Discriminative Weight Training for Unit-Selection Based Speech Synthesis

Seung Seop Park, Chong Kyu Kim and Nam Soo Kim

School of Electrical Engineering and INMC
Seoul National University
sspark@hi.snu.ac.kr, ckkim@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

Concatenative speech synthesis by selecting units from large database has become popular due to its high quality in synthesized speech. The units are selected by minimizing the combination of target and join costs for a given sentence. In this paper, we propose a new approach to train the weight parameters associated with the cost functions used for unit selection in concatenative speech synthesis. We first view the unit selection as a classification problem, and apply the discriminative training technique which is found an efficient way to parameter estimation in speech recognition. Instead of defining an objective function which accounts for the subjective speech quality, we take the classification error as the objective function to be optimized. The classification error is approximated by a smooth function and the relevant parameters are updated by means of the gradient descent technique.

1. Introduction

Recently, the waveform concatenation technique based on unit selection has become the predominant approach in speech synthesis [1]-[4]. This technique involves generating speech by concatenating a series of waveform segments which are selected from a large database. For each fundamental synthesis unit such as phone, diphone or syllable, there are a lot of waveform instances with different prosodic and spectral characteristics. But we should select one among them to produce the final speech, and the unit sequence that is expected to produce the most natural overall sound quality is selected.

Given an input text, a sequence of targets which specify the unit string with appropriate prosodic features (pitch, duration and power) is predicted based on the various contextual information resulted from the text analysis. Practically, in order to select the optimal unit among the possible candidates in the database, two kinds of cost functions are computed. One is the target cost, which measures the distance between the predicted target and each candidate database unit. The other is the concatenation cost that estimates the quality of joining two consecutive units. In [1], all of the units in a synthesis database are considered as a state transition network where state occupancy is controlled by the target cost and each state transition is associated with the corresponding concatenation cost. As the size of a speech database increases, the full search for the units in the database at synthesis time is considered to be inefficient because it is time consuming. For that reason in [2] [3], the data of each unit type (i.e. phone) is clustered in a tree structure based on the data's phonetic and prosodic context labels. This produces a number of clusters comprised of contextually and acoustically similar units. At run time, the best cluster of candidate units can be found using the decision tree.

Each target and each candidate unit in the database is characterized by a multi-dimensional feature vector and the cost functions are described in terms of the specified feature vectors. Generally, the target and concatenation costs are defined as a weighted sum of a number of sub-costs which determine the distances between the separate components of two feature vectors. The weights associated with respective sub-costs are usually trained based on a set of natural utterances held out from the synthesis database. One of the simplest approaches to estimate the weights is the grid search method in which a limited search is performed over a finite number of points in the weight space. However, the computational complexity of the grid search method grows exponentially with the number of weights and the number of grid points for each weight. Another approach to optimal weight estimation is the regression-based training technique in which the weights are determined such that an objective distance function can be well approximated by the weighted sum of sub-costs [1]. Even though the regression-based training technique is computationally more efficient and has many advantages compared with the grid search technique, it is difficult to define an objective distance function that is closely related to the subjective speech quality.

In this paper, we propose a novel approach to weight training. Our approach is based on the discriminative training technique which is frequently applied to parameter estimation in speech recognition [5]. In the proposed approach, the task of unit selection is treated as a unit recognition problem, and the weights are continuously updated so as to improve the recognition performance. A loss function which smoothly approximates the empirical classification error is defined and the weights are updated based on the gradient descent approach. The proposed approach is applied to estimate the target cost function in our experiments.

2. Discriminative Weight Training

Let $\mathbf{x} = [x_1, x_2, \cdots, x_d]$ and $\mathbf{y} = [y_1, y_2, \cdots, y_d]$ represent the target and candidate feature vectors, respectively. The target cost, $\mathcal{D}(\mathbf{x}, \mathbf{y})$ is defined as a weighted sum of the distances between the corresponding elements such that

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} w_i \mathcal{D}_i(x_i, y_i) \quad (1)$$

where $\mathcal{D}_i(\cdot, \cdot)$ is a sub-cost which indicates the distance measure assigned to the ith element. Given the definition of the target cost as shown in (1), our goal is to estimate the set of weights, $\{w_i\}$. Since the scaling of each weight by a common factor does not affect unit selection, we assume that the weights

are confined to be

$$\sum_{i=1}^{d} w_i = 1\,, \quad w_i \geq 0 \ \text{ for } \ i = 1, 2, \cdots, d\,. \tag{2}$$

Estimation of the weights is performed under the discriminative training framework and we apply the generalized probabilistic descent (GPD) technique [5]. Let us assume that $\mathbf{x}$ represents a target feature vector and there are M candidate feature vectors $\left\{\mathbf{y}^1, \mathbf{y}^2, \cdots, \mathbf{y}^M\right\}$. Without loss of generality, we also assume that $\mathbf{y}^{j^*}$ is the optimal feature vector among the M candidates. Based on these assumptions, it is desirable to estimate the weights such that

$$\mathcal{D}\left(\mathbf{x}, \mathbf{y}^*\right) = \min_{1 \leq k \leq M} \mathcal{D}\left(\mathbf{x}, \mathbf{y}^k\right) \tag{3}$$

where $\mathbf{y}^* = \mathbf{y}^{j^*}$.

Here, $\mathbf{x}$ is a target vector predicted based on the text context of token $\mathbf{y}^*$. Now, the problem of unit selection can be viewed as the classification task where $\mathbf{y}^*$ is treated as the correct classification. If the target prediction is perfect so that $\mathbf{x} = \mathbf{y}^*$, $\mathcal{D}\left(\mathbf{x}, \mathbf{y}^*\right) = 0$ for any choice of the weights and it is meaningless to estimate them. Since, however, the target prediction is usually inaccurate i.e., $\mathbf{x} \neq \mathbf{y}^*$, we should find an appropriate set of weights $\{w_i\}$ to select the optimal unit among the candidates.

The GPD approach approximates the empirical classification error by a smooth objective function defined by

$$L = \frac{1}{1 + e^{-\beta\xi}} \ , \ \beta > 0 \tag{4}$$

where

$$\xi = \mathcal{D}\left(\mathbf{x}, \mathbf{y}^*\right) + \log\left[\frac{1}{M-1} \sum_{k;k \neq j^*} \exp\left(-\eta \mathcal{D}\left(\mathbf{x}, \mathbf{y}^k\right)\right)\right]^{\frac{1}{\eta}} \tag{5}$$

with η being a positive parameter. Once the parameters, β and η are specified, the weights are trained according to the following criterion:

$$\{\hat{w}_i\} = \underset{\{w_i\}}{\operatorname{argmin}} \ L\,. \tag{6}$$

The steepest descent method is considered the easiest way to optimize the weights according to the above criterion. However, direct application of the steepest descent algorithm is found difficult due to the constraints on the weights as given by (2).

One way to overcome this difficulty is to apply the conventional steepest descent technique after parameter transformation. We adopt the following parameter transformation scheme [5] :

$$\begin{aligned} w_i &\rightarrow \tilde{w}_i \ , \ i = 1, 2, \cdots, d \\ w_i &= \frac{e^{\tilde{w}_i}}{\sum_{k=1}^{d} e^{\tilde{w}_k}} \end{aligned} \tag{7}$$

where $\{\tilde{w}_i\}$ represents the set of transformed weights. This parameter transformation scheme makes the original constrained optimization problem converted into an unconstrained one, and the conventional steepest descent technique can be easily applied to the transformed weights.

Let $\left\{\tilde{w}_i^{(n)}\right\}$ denote the set of estimates for the transformed weights at time n. Then, based on the steepest descent algorithm it is updated as follows:

$$\tilde{w}_i^{(n+1)} = \tilde{w}_i^{(n)} - \epsilon \left.\frac{\partial L}{\partial \tilde{w}_i}\right|_{\tilde{w}_i = \tilde{w}_i^{(n)}} \tag{8}$$

where $\epsilon\ (> 0)$ is a step size. With (4), (5) and (7), an elaborate computational work will lead us to

$$\frac{\partial L}{\partial \tilde{w}_i} = \beta L\,(1\text{–}L)\left[\frac{\partial \mathcal{D}\left(\mathbf{x}, \mathbf{y}^*\right)}{\partial \tilde{w}_i} - \sum_{k;k \neq j^*} p(k) \frac{\partial \mathcal{D}\left(\mathbf{x}, \mathbf{y}^k\right)}{\partial \tilde{w}_i}\right] \tag{9}$$

in which

$$p(k) = \frac{\exp\left(-\eta \mathcal{D}\left(\mathbf{x}, \mathbf{y}^k\right)\right)}{\sum_{i;i \neq j^*} \exp\left(-\eta \mathcal{D}\left(\mathbf{x}, \mathbf{y}^i\right)\right)}\,. \tag{10}$$

For example, if the traditional Euclidean metric is used to determine the distance for each element, for two feature vectors $\mathbf{x} = [x_1, x_2, \cdots, x_d]$ and $\mathbf{y} = [y_1, y_2, \cdots, y_d]$ we have

$$\begin{aligned} \frac{\partial \mathcal{D}\left(\mathbf{x}, \mathbf{y}\right)}{\partial \tilde{w}_i} &= \frac{\partial}{\partial \tilde{w}_i}\left[\sum_{j=1}^{d} w_j (x_j - y_j)^2\right] \\ &= w_i\left[(x_i - y_i)^2 - \mathcal{D}\left(\mathbf{x}, \mathbf{y}\right)\right]\,. \end{aligned} \tag{11}$$

Till now, we have been considering the case in which a single target feature vector $\mathbf{x}$ and the corresponding single set of candidate feature vectors $\left\{\mathbf{y}^1, \mathbf{y}^2, \cdots, \mathbf{y}^M\right\}$ are given. But, there are usually a lot of examples of the same unit in the training database. Moreover, several units need to share the same target cost weights for robust parameter estimation. For these reasons, it is necessary to extend the proposed approach such that it can handle multiple examples of target feature vectors. Let us assume that there are T target feature vectors $\left\{\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^T\right\}$ which have the same weights in common and each $\mathbf{x}^t$ is associated with $M(t)$ candidate feature vectors $\left\{\mathbf{y}^{t,1}, \mathbf{y}^{t,2}, \cdots, \mathbf{y}^{t,M(t)}\right\}$. Then, for each $\mathbf{x}^t$ we can define the smoothed objective function L_t,

$$L_t = \frac{1}{1 + e^{-\beta\xi_t}} \tag{12}$$

in which

$$\begin{aligned} \xi_t \ &= \ \mathcal{D}\left(\mathbf{x}^t, \mathbf{y}^{t,*}\right) \\ &+ \ \log\left[\frac{1}{M(t) - 1} \sum_{k;k \neq j^*(t)} \left(-\eta \mathcal{D}\left(\mathbf{x}^t, \mathbf{y}^{t,k}\right)\right)\right]^{\frac{1}{\eta}} \end{aligned} \tag{13}$$

with $\mathbf{y}^{t,*} = \mathbf{y}^{t,j^*(t)}$ being the optimal candidate feature vector among $\left\{\mathbf{y}^{t,1}, \mathbf{y}^{t,2}, \cdots, \mathbf{y}^{t,M(t)}\right\}$. Given L_t for $t = 1, 2, \cdots, T$, the objective function for the multiple target units is defined by

$$L_{av} = \frac{1}{T} \sum_{t=1}^{T} L_t \tag{14}$$

and the weights are updated according to the following steps:
i) Initialization: $n = 0$

$$\begin{aligned} w_i^{(0)} &= \frac{1}{d} \\ \tilde{w}_i^{(0)} &= \log w_i^{(0)} \ , \ i = 1, 2, \cdots, d \end{aligned}$$

ii) Updating Weights: $n = 1, 2, \cdots$

$$\begin{aligned}
\tilde{w}_i^{(n)} &= \tilde{w}_i^{(n-1)} - \epsilon \left. \frac{\partial L_{av}}{\partial \tilde{w}_i} \right|_{\tilde{w}_i = \tilde{w}_i^{(n-1)}} \\
&= \tilde{w}_i^{(n-1)} - \frac{\epsilon}{T} \sum_{t=1}^{T} \left. \frac{\partial L_t}{\partial \tilde{w}_i} \right|_{\tilde{w}_i = \tilde{w}_i^{(n-1)}} \\
w_i^{(n)} &= \frac{e^{\tilde{w}_i^{(n)}}}{\sum_{k=1}^{d} e^{\tilde{w}_k^{(n)}}}
\end{aligned}$$

3. Speech Synthesis Experiment and Results

3.1. Experimental Setup

In order to evaluate the performance of the proposed discriminative training approach, we conducted unit selection experiments where 3400 sentences spoken by a female speaker were used both for prosody prediction and for weight training, and another 1458 sentences provided by the same speaker were applied for performance evaluation. The training data was composed of 94997 phones and 40713 phones were discovered in the test sentences. Each sentence was sampled at 16 kHz and analyzed with a frame size of 10 ms.

The overall procedure for unit selection is as follows: First, a given text was converted to a phone sequence through the text analysis stage, and then the sequence of target units was determined based on context clustering. As the basic units of speech synthesis we used a set of clustered context-dependent phones, and the synthesized waveforms were generated by concatenating the selected phone segments in the training database. Context clustering was performed based on the approach proposed in [2] where all the waveform segments corresponding to the same phone identity were clustered into several categories by means of a binary decision tree. Splitting of each node into two child nodes was conducted based on a set of broad context questions involving adjacent phone context, syllable position in the word, word and syllable positions in the phrase, number of syllables and words in the phrase, and part of speech (POS). At each node, a single Gaussian with a diagonal covariance matrix was constructed to characterize the distribution of the 12 cepstral coefficients and their first and second derivatives, and the question which resulted in the largest likelihood gain was selected. The total number of the unit clusters that is, the terminal nodes in the constructed context clustering trees was 3730, and each cluster contained on average 25.5 phone segments.

For each leaf node of the context clustering tree, we trained a hidden Markov model (HMM) in which the number of states varied from 2 to 3 depending on the average length of the clustered phone segments. The trained HMM's had the structure of a left-to-right type without skipping and 32 mixture Gaussians with diagonal covariance matrices were computed for each state. After HMM training, all the phone segments in the training database were aligned along the corresponding HMM state sequences by means of the Viterbi algorithm. The feature vector for unit selection consisted of 1) average pitch within each state, 2) average log energy within each state, and 3) duration of the phone segment. Since the pitch is useful for representing the periodicity of a signal, the state specific pitch values were applied only to the voiced phones.

For an accurate target prediction, each component of the feature vector was predicted based on a separate decision tree constructed in a manner similar to context clustering mentioned earlier. Each leaf node of the decision trees provided not only the predicted value of the relevant feature vector component but also its variance which was later used for distance normalization. Given a target unit, the phone segments in the training database which fell on the same leaf node of the context clustering tree were used to specify the corresponding candidate units. We extracted a feature vector from each candidate unit to compute the target costs used for optimal unit selection.

3.2. Test Results

In our experiment, both the target and concatenation costs were defined as a weighted sum of the distances between the corresponding components of two feature vectors. The proposed discriminative training algorithm was applied only to estimate the target cost weights. The concatenation costs were defined as given in [1] , and they were held fixed during the experiment. In order to investigate the effect of weight training more clearly, we emphasized the target cost such that it could be the dominant factor in cost computation. All the target units with the same phone identity shared the same target cost weights, and a separate set of weights was estimated for each context-independent phone. The distance of each feature vector component was specified as the Euclidean distance normalized by its variance, which was provided by the decision tree used for target feature prediction.

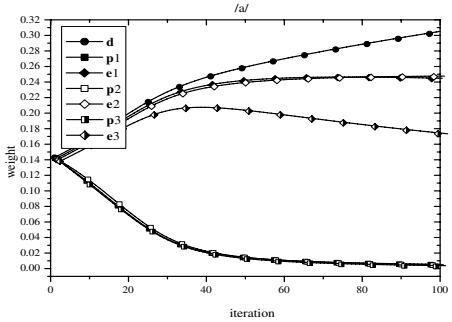

Figure 1: An example of weights training pattern.

All of the sentences in the training database were applied for training the target cost weights. The parameters used for defining the objective function, L were selected such that $\beta = 1$ and $\eta = 5$, and the step size for parameter update was set to $\epsilon = 10$ which resulted in a smooth convergence of the objective function to the minimum point. An example of the weight training pattern is shown in Fig. 1 where $\mathbf{p}i$, $\mathbf{e}i$, and $\mathbf{d}$ denote the average pitch at the ith state, average log energy at the ith state, and phone duration, respectively. It is noted that the estimated weights have different shape depending on the phone characteristic.

Both the objective and subjective quality tests were conducted for performance evaluation over the test database. For an objective measure, we computed the cepstral distance between the original phone segment and the unit segment selected with the estimated target cost weights. It was reported in [6] that the Euclidean cepstral distance produced the best result in predicting perceptual distortion among the tested objective distance metrics. Cepstral distance was computed with the use of the dynamic time warping (DTW) technique and then normalized by the frame length of the given phone segment. This nor-

malized cepstral distance was accumulated over all the phones in the test data and averaged to yield the performance measure. For the purpose of comparison, we also tried the conventional regression-based training approach to estimate the target cost weights. The procedure for regression-based weight training is summarized as follows [1] :

1. For each segment of the same phone identity in the training database perform steps *(a) - (d)*.
   (a) Treat the phone segment as a target unit.
   (b) Calculate the normalized cepstral distance between the target unit and all the candidate units.
   (c) Identify the set of n-best matches to this target unit (n = 20).
   (d) Determine the target sub-costs for the target unit and the n-best matches.
2. Collect the normalized cepstral distances and target sub-costs across all the target units and all the n-best matches.
3. Determine the weights using the linear regression technique.

method	cepstral distance
discriminative approach	22.06
regression-based approach	22.55

Table 1: *Average cepstral distance between the original and selected unit segments computed in the test database*

It is noted that in the regression-based approach there is no need to restrict the weight space to a subset of R^d as specified in (2). The result for average cepstral distance computed in the test database is shown in Table 1. From the result we can see that the discriminative training technique produced even smaller average cepstral distance compared to the regression-based approach which was designed to directly approximate the cepstral distance in the training data.

prefer A	no difference	prefer B
55.0 %	33.3 %	11.7 %

Table 2: *Result of preference listening tests (A : discriminative method, B : regression method)*

Subjective quality of the synthesized speech was measured by a preference listening test. 20 sentences which were randomly selected from the test database were used for this listening test. For a given sentence, we generated two speech waveforms where one was synthesized using the target cost weights trained based on the discriminative approach and the other was obtained with the weights estimated by the regression-based approach. 10 listeners participated in the listening test and for each sentence they compared the perceived naturalness between the two synthesized waveforms. The result for the preference listening test is given in Table 2 which confirms that the proposed approach is superior to the conventional regression-based method in making natural sounds.

4. Discussion

In training the weight parameters, prediction accuracy of the target values is important. However, given the target prediction module, the effect of the poorly predicted target values should be minimized in our approach. We believe that this is handled implicitly. By comparing the target value with the actual unit's value in the weight training phase, the relatively well predicted feature can be more weighted and has more effect on selecting the final unit.

Even though we trained the weight parameters for each phone in this work, each parameter can be specified for each cluster. If units are clustered such that certain feature values are very similar (for example, pitch), then that feature will play a less important role in selecting the final unit in that cluster, resulting in a small weight.

Although the proposed technique has been applied only to estimate the target cost weights in our experiment, we consider that it can be easily extended to train all the cost functions jointly under a well-defined conditions.

5. Conclusions

We have proposed a discriminative training approach to estimate the cost functions used for optimal unit selection in concatenative speech synthesis. Instead of defining a speech quality oriented objective function, we have introduced a smoothed classification error function which has been frequently adopted in speech recognition. The listening test showed that the proposed algorithm results in better voice quality in overall.

6. References

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", IEEE Int. Conf. Acoust., Speech and Signal Proc., pp. 373-376, 1996.

[2] R. E. Donovan, Trainable Speech Synthesis. Ph.D. Thesis, Cambridge University, 1996.

[3] Black, A., and Taylor, P. "Automatically clustering similar units for unit selection in speech synthesis", Eurospeech97, vol.2, pp. 601-604.

[4] M. Beutnagel, A. Conkie, J. Schroeder, Y. Stylianou, and A. Sydral, "The AT&T Next-Gen TTS System", 137th meeting of the Acoustical Society of America, pp. 18-24, 1999.

[5] B. -H. Juang, W. Chou, and C. -H. Lee, "Minimum classification error rate methods for speech recognition", IEEE Trans. Speech Audio Processing, vol. 5, no. 3, pp. 257-265, May 1997.

[6] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," IEEE Int. Conf. Acoust., Speech and Signal Proc., pp. 837-840, 2001.