Pronunciation Variation Generation for Spontaneous Speech Synthesis Using State-Based Voice Transformation

Chung-Han Lee, Chung-Hsien Wu and Jun-Cheng Guo

Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan
Email: {chlee, chwu} @csie.ncku.edu.tw

ABSTRACT
This study presents an approach to Hidden Markov Models (HMM)-based spontaneous speech synthesis with pronunciation variation for better spontaneity. Pronunciation variation generally occurs in spontaneous speech and plays an important role in expressing the spontaneity. In this study, a state-based transformation function is adopted to model the relation between read speech and the corresponding spontaneous speech with pronunciation variations. The transformation function is then used to generate the state-based pronunciation variations. Due to the lack of training data, the articulatory features are used to cluster the transformation functions using Classification and Regression Trees (CARTs) such that the unseen pronunciation variation with the same articulatory features can be generated from the transformation function in the same cluster. Objective and subjective tests are conducted to evaluate the performance of the proposed approach. The experimental results show that the proposed transformation function achieves a significant improvement on spontaneity in synthesized speech.

Index Terms: Pronunciation variation, speech synthesis, transformation function

1. INTRODUCTION
In recent years, two main trends for speech synthesis were proposed to produce a highly intelligible and decent quality speech. The first one, concatenative synthesis method [1], is the representative technology. The HMM-based synthesis method [2] is another popular one; the stochastic methods dominated the development of the data-driven synthesis systems. HMM-based synthesis systems have been adopted for spectral, duration and pitch generation, and possess the portability and flexibility for high quality speech synthesis. The stochastic models can be adapted effectively and quickly among different speakers or emotions; models can be converted using a small amount of training corpus [3]. However, most of these synthesis systems were applied to “read speech” synthesis. The spontaneity property in spontaneous speech is still not well modeled. For read speech, the speaking speed is stable and the pronunciation is clear, while for spontaneous speech, the speaking rate is unstable and pronunciation variations generally occur. In this study, the generation of pronunciation variations is desired to achieve better spontaneity property in speech synthesis.

Several kinds of pronunciation variations for Mandarin spontaneous speech have been defined in [4]. The Mandarin Conversational Dialogue Corpus (MCDC) [4] contains four types of variations, including 1) syllable contraction, 2) nasalization, 3) assimilation, and 4) lengthening, in which the ratio of syllable contraction is about 84% of all variations. In this approach, a linear transformation function is constructed for pronunciation variation modeling. For linear transformation function construction, the acoustic features of the input read speech and the corresponding pronunciation variation in spontaneous speech are extracted using the STRAIGHT algorithm [5]. The parallel features for each read speech and pronunciation variation phone pair are then modeled by an HMM; the transformation function for each state of this pair is derived using the trained HMM. In order to eliminate the lack of the pronunciation variations in the collected speech database, a CART is constructed to select a suitable transformation function for pronunciation variation generation in spontaneous speech synthesis. The articulatory features are considered to construct the CART for transformation function clustering from the training data. Thereafter, in the synthesis phase, this CART is used to select the transformation function based on the articulatory features of the input text for pronunciation variation generation.

2. PRONUNCIATION VARIATIONS IN SPONTANEOUS SPEECH
To show the pronunciation variation in spontaneous speech, the dynamic time warping (DTW) algorithm is adopted to align the mapping between the spectra of normal speech and spontaneous speech. Fig. 1 shows the aligned result using DTW between normal speech and spontaneous speech.

Figure 1: Aligned result between normal and spontaneous speech.

The vertical lines (between the 2nd purple line and 3rd purple line) describe the mapping from normal speech to...
spontaneous speech; this presents the deletion effect of normal speech to the spontaneous speech. On the other hand, the horizontal lines (between the 3rd red line and 2nd yellow line) show the insertion effect of normal speech to the spontaneous speech. The syllable boundaries of spontaneous speech with pronunciation variation can be identified by the aligned boundaries of normal speech using DTW.

3. TRANSFORMATION FUNCTION MODELING

The purpose of this study is to construct the transformation functions between the data samples of the read speech and spontaneous speech with pronunciation variations. Fig. 2 shows the flowchart of the proposed method.

![Flowchart of the proposed method.](image)

In the training phase, a parallel speech database is collected, and the feature vectors, consisting of spectrum and duration parameter vectors, are extracted for HMM training. The STRAIGHT algorithm is adopted to extract the spectral features. The transformation function is derived using the trained HMM for transforming from normal phone to variation phone state-by-state. The articulatory features of the input text are extracted for model selection from the F-CART and D-CART. The normal (read-style) speech is synthesized from the input text by HMM-based TTS system (HTS). Finally, spontaneous speech with pronunciation variation is synthesized from the converted mel-cepstral coefficients and scaled durations obtained from the F-CART and D-CART using the MLSA filter.

In transformation function modeling, the parallel observation sequences from normal and variation speech are used for training. The sequences \( X = \{x_1, x_2, \ldots, x_N \} \) and \( Y = \{y_1, y_2, \ldots, y_N \} \) are the parallel source (normal speech) and target (variation speech) feature (MFCC) sequences of \( N \) frames, respectively. A multi-dimensional linear regression model is adopted as the state-based transformation function for converting the normal speech \( X \) to the variation speech \( Y \):

\[
Y = AX + R
\]

where \( A \) is the linear transformation matrix and the residual matrix \( R \) is determined as the difference between the converted spectrum through transformation matrix \( A \) and the target spectrum. In the training phase, a statistical model \( \lambda \) for the normal and variation speech features, which is estimated to maximize the likelihood function of their joint distribution, is modeled as [6]:

\[
P(X, Y|\lambda) = \sum_{q_y} P(X, Y, q_y) = \sum_{q_y} \prod_{t=1}^{T} a_{q_{t-1}, q_t}(x_t, y_t),
\]

\[
b_j(x, y) = b_j(y|x)b_j(x)
\]

\[
b_j(y|x) = N(y_j; A_j x + R_j, \Sigma_j)
\]

\[
b_j(x) = N(x; \mu_j, \Sigma_j)
\]

where \( q = \{q_1, q_2, \ldots, q_T\} \) denotes the \( T \)-state sequence shared by the source and target feature streams; \( \pi \) is the initial probability; \( a \) is the transition probability from state \( q_t \) to \( q_{t+1} \); and \( b(\cdot) \) means the state observation probability density function (PDF) for state \( j \). \( N(x; \mu, \Sigma) \) represents the Gaussian distribution of \( x \) with mean vector \( \mu \) and covariance matrix \( \Sigma \). The parameters for each state can be estimated using the EM algorithm [7]. The re-estimation formulas for state \( j \) can be derived as [6]:

\[
\mu_j = \frac{1}{T} \sum_{t=1}^{T} r_j(j) x_t
\]

\[
\Sigma_j = \frac{1}{T} \sum_{t=1}^{T} r_j(j)(x_t - \mu_j)(x_t - \mu_j)^T
\]

\[
A_j = \left( \sum_{t=1}^{T} r_j(j)(y_t - A_j x_t)x_t^T \right)^{-1}
\]

\[
R_j = \frac{\sum_{t=1}^{T} r_j(j)y_t A_j x_t}{\sum_{t=1}^{T} r_j(j)}
\]

\[
\Sigma_j = \frac{\sum_{t=1}^{T} r_j(j)(y_t - A_j x_t - R_j y_t)(y_t - A_j x_t - R_j y_t)^T}{\sum_{t=1}^{T} r_j(j)}
\]

where \( \gamma_j(j) \) denotes the state occupancy probability of
frame \( t \) belonging to state \( j \). The spectrum of normal speech can be transformed to the variation speech by (1) importing (6) and (7). Duration model can be trained using the conventional HMM-based TTS system. The ratio between source duration \( L^s \) and target duration \( L^\ell \) can be used to scale the duration information. The normal duration can be scaled to the variation duration.

4. TRANSFORMATION FUNCTION CLUSTERING

Pronunciation variations with similar articulatory features, which are classified in the same cluster, are assumed to have similar transformation function. The transformation functions and duration ratio can be shared in the cluster with the same articulatory features. In order to select the transformation function and duration information, the transformation functions and duration information are classified by an F-CART and a D-CART considering the articulatory features, respectively. The CART is adopted to model the relation between the articulatory features and the transform function clusters; The F-CART is then used to retrieve an appropriate transformation function for spectrum conversion and the D-CART is used to retrieve duration ratio for duration information scaling. The articulatory features used in this study include voiced/unvoiced plosive, fricatives, affricatives, nasals, liquids, front/central vowels, etc.

In the training phase of the CART, each sample is composed of cluster index and the corresponding features, including linguistic features and articulatory features. For F-CART, the generation error is adopted as the splitting criterion and is estimated as:

\[
GenErr_i = \sum_{m=1}^{M} \| y_m - (A x_m + R) \|^2
\]

where \( y_m \) and \( x_m \) denote the \( m \)-th frame of target (variation speech) and source (normal speech) feature vectors, respectively. Generation error is determined as the difference between the transformed spectrum generated from the transformation function and the target spectrum. The GenErr represents the information gain for a split. In order to minimize the generation error, the reduced generation error (RGE) represents the potential information generated by splitting a parent node with \( GenErr_p \) samples into child nodes with \( GenErr_i \) samples and is calculated as:

\[
RGE = GenErr_p - \sum_i W_i GenErr_i
\]

where \( GenErr_p \) and \( GenErr_i \) denote the generation errors of the parent node and the \( i \)-th child node, respectively; \( W_i \) is the number of items in cluster \( i \). The split with the largest RGE in all possible splits is chosen and the generation error of split nodes in each cluster are re-estimated by (1) importing (6) and (7); The tree growing stops when there is no significant information gain for all nodes. For D-CART, the duration information \( L = (L^s, L^\ell, \ldots, L^s, L^\ell)^T \) (\( n \) is the number of states) for each node are classified according to the articulatory features. The mean square error (MSE) is adopted as the splitting criterion and is estimated as:

\[
MSE_i = \sum_{j=1}^{n} (x_j - \bar{x})^2
\]

where \( MSE_i \) denote the MSE of the \( i \)-th node; \( x_j \) denotes the duration information of the \( j \)-th sample in the \( i \)-th node; \( n \) is number of items in node \( i \), and \( \bar{x} \) is the mean value of all samples in node \( i \). The reduced mean square error (RMSE) represents the potential information generated by splitting a parent node with \( MSE_p \) samples into child nodes with \( MSE_i \) samples and is calculated as:

\[
RMSE = MSE_p - \sum_i \left( \frac{M}{M_p} \times MSE_i \right)
\]

where \( MSE_p \) and \( MSE_i \) denote the mean square error of parent node and the \( i \)-th child node, respectively; \( M_p \) and \( M \) denote the numbers of items in parent node and the \( i \)-th child node, respectively. Each leaf node of the D-CART consists of the mean and variation of the duration information for this cluster.

5. EXPERIMENTS AND RESULTS

For experiments, a phonetically balanced small-sized parallel speech database pronounced by three speakers, which contains read speech and spontaneous speech with pronunciation variations, was designed and collected to train the pronunciation variation models. The numbers of training sentences were 200, 100 and 50 for each speaker. The speakers were one female and two male native speakers and all the collected speech databases were manually transcribed and marked. For synthesis, the normal speech was generated by the conventional HMM-based TTS. The Tsinghua-Corpus of Speech Synthesis (TH-CoSS) [8] database containing 5,406 sentences and 98,749 syllables in read speech (normal) was adopted for HTS training. Experiments were designed and conducted to assess the performance for pronunciation variation modeling, transformation function selection and duration scaling using CART. For feature extraction, smoothed spectrum was extracted by the STRAIGHT algorithm.

5.1 Comparisons with Other Variation Generation Methods

Mean square error was adopted as the performance measure for objective evaluation. \( MSE \) was calculated between the converted and the target spectrum feature vectors as:

\[
MSE = \frac{1}{N} \sum_{n=0}^{N} (y_n - \hat{y}_n)^2
\]

where \( N \) is the total number of feature vectors. \( y_n \) and \( \hat{y}_n \) denote the target and the converted feature vectors.
For objective evaluation, the variation phone in the parallel speech corpus is considered as the target feature vector. Three different variation generation methods were adopted for comparison as follows: 1) MLLR adaptation from normal phone to variation phone; 2) duration scaling; 3) phone-based variation transformation function. Experiments were conducted on the comparisons of the proposed state-based voice transformation method, adaptation method, duration scaling and phone-based pronunciation variation transformation methods. Fig. 3 shows the MSE of all the collected speech data at the phone level for the four different methods. The analytical results indicate that the state-based pronunciation variation generation method can achieve lower MSE than the others.

![Figure 3: MSEs of different variation generation methods](image)

5.2 Evaluation via Formal Listening

For subject evaluation, the speech spontaneity and quality of the proposed method were assessed based on formal listening test. Fifteen listeners were asked to compare each synthesized sentence with the conventional HTS, MLLR adaptation and the proposed method. The listeners were asked to give a score from one to five, with one for “bad” and five for “excellent”. Fig. 4 presents the results from the listening test.

![Figure 4: MOS Listening test](image)

The quality of the synthesized speech using HTS are generally better than that using adaptation and state-based transformation methods. The reason is because the speech quality directly synthesized by HTS without conversion was not affected by the errors from adaptation or transformation. However, the synthesized speech using the proposed state-based transformation method obtains much better spontaneity than the read speech synthesized by the HTS. In addition, the proposed state-based method only slightly degrades the speech quality compared to that generated from HTS. The MLLR-based adaptation method cannot obtain good speech quality as well as improve the spontaneity of the synthesized speech.

6. CONCLUSIONS

This study presented a method to improve the spontaneity property in conventional HMM-based speech synthesis. The pronunciation variation can be modeled by a linear transformation function trained from the parallel normal and variation speech data. The acoustic and articulatory features are considered for transformation function clustering to construct a function selection CART (F-CART) and a duration scaling CART (D-CART). The F-CART and D-CART are then adopted to select appropriate transformation function and duration information using the articulatory features, respectively. From the evaluation results, the proposed state-based transformation method can generate the pronunciation variations to effectively improve the spontaneity in spontaneous speech synthesis compared to the conventional HMM-based HTS and the MLLR-based adaptation methods.

7. REFERENCES


