Feature Integration and Dimension Reduction in Unit Selection TTS

Ling Cen, Minghui Dong, Paul Chan, Haizhou Li
Institute for Infocomm Research (I²R), A*STAR, 1 Fusionopolis Way, Singapore 138632
{lcen, mhdong, ychan, hli}@i2r.a-star.edu.sg

Abstract

Unit selection based speech synthesis is able to generate high quality speech. During unit selection process, a set of unit candidates are provided for each target unit. Its performance is degraded with a large number of candidates. In order to improve the speed of the unit selection process, it is important to determine a small set of candidates. This needs a candidate selection method that does not miss good candidate units. In this work, a dimension-reduction based method is proposed to fuse different features together and use these features as the pre-selection criteria in the compilation of the candidate set in the unit-selection based speech synthesis method. Experiments in Mandarin TTS shows that speech quality can be improved using the proposed method.

Keywords

Text-to-speech, unit selection, unit features, dimension reduction.

1 Introduction

The unit-selection based approach to speech synthesis (Hunt et al., 1996; Black et al., 1997, Clark et al., 2006; Schroder et al., 2006) has become one of the most popular methods to generate high-quality synthesized speech. A unit selection based speech synthesis system consists of a large unit database, in which there are various instances for each unit as candidate units. The instances in the database are designed to cover the variations of each unit as much as possible. During the synthesis process, a set of candidate units are associated with each target unit. A search process is designed to find an optimal sequence based on a cost function.

In order to maintain the natural quality of synthetic speech, one needs to predict the prosody for each unit. The prosody model predicts the prosodic parameters, which normally describes the pitch, duration and energy of speech units. These prosodic parameters are used as part of the criteria for unit selection.

However, using prosody alone is not enough. The prosodic parameters can only ensure its prosodic suitability in the synthesis. Another important aspect of natural speech, the spectral suitability of the speech unit, also needs to be considered. Although many current systems do not explicitly consider spectral features when selecting units, the spectrum is somewhat controlled during unit selection. The spectral suitability is normally established in

two ways: (1) the use of contextual information of the unit to make sure that the selected unit is chosen from a context similar to that of the target unit. The underlying assumption is that the units from the same context would have similar spectral properties. (2) the use of spectral information as part of the joining cost during the concatenation of the units. This is achieved by comparing the spectral mismatch on the boundaries of the units that will be concatenated. However, these methods cannot effectively ascertain the smoothness of speech. Acute spectral mismatches are still often perceived in the speech that is synthesized.

To better describe each unit, it is expected to use both prosodic and spectral features together in unit specification. The spectrum is normally represented with a vector; prosody of speech units can also be represented with a set of parameters. If we integrate spectral and prosodic vectors for different parts (e.g. segments of HMM states) of a unit, the joint vector will be very long. It is not only time-consuming to process such a long vector, but, more importantly, it also contains a lot of redundant information. This prompts us to find a way to reduce the dimension and remove the redundancy. In this work, we use the Principal Component Analysis (PCA) (Jolliffe, 2002) approach to achieve a dimension-reduced vector that is a compact form of unit features. The statistical model is used to predict the parameters. The vector is used for candidate unit selection in speech synthesis process.

In the remaining part of the paper, we will first introduce the data we used, and the feature calculation process. After that, we will introduce the PCA method that is used for dimension reduction. Next, the unit selection process is described. Finally, the experiments are presented and our conclusion is drawn.

2 Data Processing

Our work is based on a Mandarin TTS corpus. The corpus consists of 6 hours of Mandarin speech in 6000 utterances.

2.1. Forced alignment of speech and phone sequence

Mandarin is a syllable-based language, in which each Chinese character is pronounced as a mono-syllable. There are about 408 base syllables in Mandarin. Each base-syllable can be said to be composed of an Initial-Final structure similar to the Consonant-Vowel structure in other languages. Each base syllable consists of either an Initial followed by a Final or a single Final. The Initial is the initial consonant part of a syllable and the Final is the vowel part including an optional medial or a nasal ending. In Mandarin Chinese, there are 22 different Initials (including a null-initial) and 38 different Finals (Hon, et al., 1994). In our system, we further divided the finals into 1-4 phonemes, similar to the phone set used for English speech recognition. Hence, we defined 43 phones as shown in Table 1. The advantage of using the smaller unit is that we are able to handle missing syllables easily.

Table 1. ***Mandarin phone set***

18 vowels	a aa ah e ea ee een eeng eh er i iz izh o oh oo u v
25 consonants	b c ch d f g h j k l m n ng p q r s sh t vh wh x yh z zh

The speech utterances are automatically force-aligned to the pronunciations with HTK. Phone-sized speech segments are defined as our basic unit. For forced alignment, 39-dimensional MFCC features are used for training the phone models. The frame size is 25ms and the frame shift is 10ms. Three states are defined in each context independent HMM

model for each phone. The phone models are first trained using the speech corpus. Unit boundaries are then obtained by the forced alignment of speech with its phonetic sequence.

2.2. Acoustic feature calculation

A set of parameters are first defined, which describes the spectral and prosodic features of each HMM state, and the boundary (start and end) frame. The main values that we capture include the statistical values of each individual HMM state and the values of boundary frames of the unit. The initial parameters used consist of the following:

- Spectral features: MFCC mean for the 3 HMM states, MFCC for boundary frames.
- Pitch features: Mean, maximum, minimum, and range of pitch values and pitch derivative and acceleration values for 3 HMM states, and boundary frames.
- Duration features: Durations of the 3 states, duration of the unit.
- Energy features: Mean energy of frames in the 3 HMM states, and boundary frames.

Cascading all the parameters together, we have a long 308-dimensional vector. High-dimensional feature vectors not only make the unit selection process slow, but also introduce redundant information. In our method, the PCA approach is employed to reduce the dimension of the feature vector. The reduced vector is a compact form of representation of the prosodic and spectral features of the unit. Via dimension reduction using PCA, we manage to reduce this to a 40-dimensional vector.

2.3. The Prosodic Parameters

The acoustic parameters define both spectral and prosodic information. However, there are more parameters conveying spectral information than those conveying prosodic information defined in the long vector. Prosodic information is thus actually less prominent in the acoustic vector. Nevertheless, we still need a set of prosodic parameters to emphasize the prosodic properties in speech. The prosodic parameters for each unit consist of the following:

- Pitch mean of the unit
- Duration of the unit
- Energy mean of the unit
- Pitch range of the unit.

2.4. Linguistic Features

Linguistic features are derived from input text. They are used for predicting the acoustic parameters. The features we used include (the number of parameters is denoted in brackets):

- Context units: phone identities of the previous 2 and next 2 units. (4)
- Tone information: the tones of the current, previous two and next two syllables. (5)
- Phone location in syllable: number of phones in the syllable, position of the phone counting from left boundary, position of the phone counting from right boundary. (3)
- Word information: length and part-of-speech of the previous word, current word and next word, position of the syllable in word. (8).
- Prosodic phrase information: lengths of prosodic phrases of different levels, syllable locations of prosodic phrases of different levels. (12)

Altogether, we have a linguistic feature vector of 32 elements for Mandarin.

3 Model Training

3.1. Dimension reduction for acoustic features

We have initially defined a long vector of 308 dimensions to describe the unit. As the long vector contains a lot of redundant information and it is also time consuming to process quite long vector, the PCA is employed to remove redundant information contained in the feature vector and thus reduce the number of dimensions needed to describe the vector.

The PCA (Jolliffe, 2002) method is able to transform a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. For a data matrix, *X*, which is normalized to zero mean, where each row represents a different repetition of the experiment, and each column gives an observed parameter, the PCA transformation is given by:

$$Y^T = X^T W = V\Sigma, \tag{1}$$

where $V\Sigma W^T$ is the singular value decomposition (SVD) of X^T. W is a $m \times n$ matrix, where *m* is the original dimension and *n* is the reduced dimension, and $m > n$.

3.2. Parameter prediction

The process for acoustic parameter prediction calculates the parameters from the linguistic features. The prediction can be represented with the following formula:

$$y_i = F_i(X), \tag{2}$$

where y_i is the i^{th} parameter for the unit and *X* is the linguistic feature vector for the unit.

In our system, the linguistic features are the predictors and the acoustic and prosodic parameters are the responses. We build our models using the Classification and Regression Tree (CART) (Breiman, et al., 1984) approach. Each individual parameter is predicted separately with a CART tree.

4 Unit Selection

There are two steps in the unit selection process. In the first step, a candidate set is determined for each target unit. Then a search process is applied to calculate target cost and dynamically determine the joint cost in the second step.

4.1. Candidate set determination

The candidate determination process calculates the pre-selection cost. All the candidates are then sorted in the ascending order of the pre-selection cost. The first *m* candidates are kept as the candidate set for unit selection.

The pre-selection cost is the weighted sum of acoustic cost and linguistic cost. It is defined as follows

$$c_s = w_{ta}c_{ta} + w_{tl}c_{tl}, \tag{3}$$

where c_{ta} and c_{tl} are the cost of acoustic parameters and linguistic features, respectively, w_{ta} and w_{tl} represent their corresponding weights.
The cost of acoustic parameters c_{ta} is defined as

$$c_{ta} = \sum_{i=1}^{n_a} ((u_i - v_i)/s_i)^2, \tag{4}$$

where n_a is the dimension of the acoustic feature vector, u_i and v_i are the predicted parameter vectors for the target unit and the actual parameter vector for the candidate unit, respectively, and s_i is the standard deviation of the i^{th} parameter.

The cost of context linguistic features c_{tl} is defined according to the difference between the features of the target unit and those of the candidate units. Whenever the values of features in the target and candidate units are different, a cost value is given. The total cost is the sum of all costs for each individual feature. In this function, we give a higher cost value to the more important factors (e.g. the identities of previous unit and next immediate unit, the accent of the unit, the stress of the unit, etc).

4.2 Unit sequence determination

In this part, we describe how we define the cost function. The unit selection process is based on the cost function that consists of two parts: (1) a target cost to measure the difference between the target unit and the candidate unit; (2) a joint cost to measure the acoustic smoothness between the concatenated units.
Our target cost is defined as

$$c_{tp} = \sum_{i=1}^{n_p} w_i ((p_i - q_i)/t_i)^2, \tag{5}$$

where n_p is the dimension of the prosodic feature vector, p_i and q_i are predicted parameter vectors for the target unit and the actual parameter vector for candidate unit respectively, t_i is the standard deviation of the i^{th} parameter, and w_i is the weight of the i^{th} parameter.

The joint cost, c_j is defined as the squared value of the Euclidean distance between the vector of the end frame in the previous unit E_{i-1} and the vector of the start frame in the current unit S_i as

$$c_j = (E_{i-1} - S_i)(E_{i-1} - S_i)^T. \tag{6}$$

The total cost *c* is calculated with the following function.

$$c = w_t \sum_{i=0}^{n} c_t(i) + w_j \sum_{i=1}^{n} c_j(i), \tag{7}$$

where n is number of units in the sequence, $c_t(i)$ is the target cost of unit i, $c_j(i)$ is the joint cost between unit i-1 and unit i, and w_t and w_j are weights for target cost and joint cost respectively.

The best unit sequence is determined by searching for the best path among the candidate unit lattice to minimize the total cost of the selected sequence. The Viterbi algorithm is used to find the best sequence. The weights in the cost function are manually tuned.

5 Experiments

The Mandarin speech corpus consists of about 6 hours of speech in 6000 different utterances.

5.1. Dimension reduction

The principal component analysis is performed for the different parameters. The eigenvalues with the index of principal components are illustrated in Fig. 1. The eigenvalues represent the importance of each individual principal component. It can be seen from the figure that the first 40 components have high values, thus are more important. With the increase of the index, the value decreased to almost 0 when the index is larger than 100. Based on the analysis, we keep the first 40 principal components in our work.

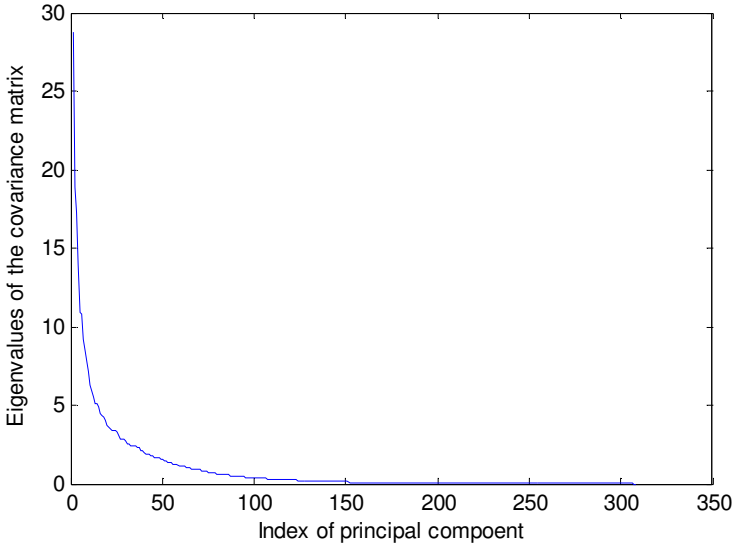

Figure 1. Eigenvalues for principal components

5.2. Validation of defined parameters

Now we test whether the defined acoustic cost in formula (3) helps to improve the speech quality. Experiment is performed by controlling the weights in the formula. We conduct two tests, in each of which top 10 candidates are kept by the candidate set determination process. The cost function and dynamic search process are kept the same in the two tests. The weights in the two tests are as shown in Table 2.

Table 2. Weight settings for 2 tests

Test	Setting
A	$w_{ta} = 0, w_{tl} = 1$
B	$w_{ta} = 1, w_{tl} = 1$

In test A, we disable the acoustic parameters, while in test B, we enable the acoustic parameters. We use the two settings to generate 20 utterances and ask 10 people to listen to the pairs of utterances and indicate their preferences. The listening test results are shown in Table 3. It is indicated from the listening test that method B is better than method A. This shows that the use of acoustic parameters in the TTS process helps to improve the speech quality.

Table 3. Listeners' preferences

Test	Result
A	31%
B	69%

6 Conclusion

In this paper, we propose to integrate multiple feature vectors to describe unit features in unit-selection based speech synthesis. Principal component analysis approach is used to reduce the number of dimensions and remove redundancy in the long vectors. The dimension-reduced parameters are used in candidate set determination in unit selection speech synthesis. Experiment with Mandarin TTS shows that the method helps to improve the speech quality of synthesized speech.

7 References

A Hunt, A W Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database", Proc. of ICASSP 1996.

A. W. Black, P. Taylor, "Automatically Clustering Similar Units for Unit Selection in Speech Synthesis," in Proc. Eurospeech 97, vol 2 pp 601-604, Thodes, Greece.

R. Clark, K. Richmond, V. Strom, S. King, “Multisyn voice for the Blizzard Challenge 2006,” Blizzard Workshop 2006.

M. Schroder, A. Hunecke, S. Krstulovic, “OpenMary – Open Source Unit Selction as the Basic for Research on Expressive Synthesis,” Blizzard Workshop 2006.

I.T. Jolliffe, “Principal Component Analysis”, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002.

H. W. Hon, et al. Towards large vocabulary Mandarin speech recognition. Proceedings of ICASSP 1994. pp:545-548.

L. Breiman, , J. H. Friedman, R. A. Olshen, and C. J. Stone, “Classification and Regression Trees”. Monterey, Calif., U.S.A.: Wadsworth, Inc., 1984.