PERCEPTUAL CLUSTERING BASED UNIT SELECTION OPTIMIZATION FOR CONCATENATIVE TEXT-TO-SPEECH SYNTHESIS

Tao Jiang[1,2], Zhiyong Wu[1,2], Jia Jia[2], Lianhong Cai[1,2]

[1] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2] Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
joungtao@gmail.com, zywu@sz.tsinghua.edu.cn, {jjia, clh-dcs}@tsinghua.edu.cn

ABSTRACT

In concatenative based speech synthesis, the purpose of unit selection is to select proper speech units from speech corpus by measuring how well the selected units match the given features. Perceptual test indicates that some features are always preferred to make perceptual distinction between units. Such features should be judged prior to others in unit selection. In this work, we attempt to identify the priorities for different features and try to optimize the unit selection with perceptual clustering. Out approach first clusters the speech units with hierarchical clustering based on a perceptual distance measurement between different speech units. A method to identify the questions (concerning the features) is then proposed to build the decision tree from the clustering result. The features used in the decision tree are the preferred ones, and the other features are used in the target cost function. Linear discriminant analysis (LDA) is then adopted to train the weights for the target cost function from the clustering result to make weights more reasonable and perceptual related.. Experimental results indicate that the optimized unit selection can generate synthetic speech with higher naturalness than the previous approach.

Index Terms— Perceptual clustering, decision tree, linear discriminant analyze, cost function, unit selection

1. INTRODUCTION

Large scale corpus based concatenative speech synthesis is still playing a very important role in state-of-the-art text-to-speech (TTS) synthesis systems in the application scenarios where high intelligibility and naturalness are desired. In this method, speech synthesis is achieved by concatenating speech units that are selected from a large speech inventory [1][2][3]. Given the target phonetic and prosodic features that are derived from input by text analysis and prosodic prediction modules of the TTS system, the purpose of unit selection is to select the proper speech units from the speech inventory by measuring how well the selected unit sequence matches the given features.

There are two kinds of methods for unit selection: cost function based method and classification and regression tree (CART) based method. The former method selects units from the speech inventory by minimizing the value of a cost function which is defined as the combination of *target* and *join* costs [1][2]. The *target* cost is defined as a weighted sum of sub-costs that measure the distances of the phonetic and prosodic features between the target units and their candidates; and the *join* cost is described as a weighted sum of the distortions at boundary where two neighboring units are concatenated. One important issue of the cost function based method is how to determine the weights so that these weights can represent the importance of different features in selecting the speech units. In [1], the *Regression* method has been adopted for determining the weights. However, in a unit perceptual distinction experiment to find the correlation between the features and unit perception, it is found that some features are always *preferred* to make a perceptual distinction between units. These preferred features may include articulation manner of previous unit's final and next unit's initial, etc. While a formula of weighted sum in the cost function cannot reflect such priorities of the features.

To solve the problem, CART based method is proposed by building a set of decision trees for unit selection [3]. With the decision tree, the speech units are classified into different clusters (i.e. tree leaves) by the context questions concerning the phonetic and prosodic features of units. The decision tree is build by the greedy algorithm which will first select questions (and hence related features) that are most significant in distinguishing the speech units. With this method, some of the features will be considered prior to others in selecting units. However, as in [3], several speech units are still included in a cluster (or tree leaf) and carry acoustic variations that are perceptually distinguishable.

To address the problem, the combination use of CART and cost functions is proposed in [4], where CART is used to select unit clusters and cost function (including *target* and *join* cost) is further used to find the optimal units from the clusters with the Viterbi algorithm. However, the features used in the target cost contain all the features that have been already used in the decision tree. Furthermore, the features

used in the decision tree usually get higher weights (with regression) in the target cost function. This will lead to the situation that higher weights are assigned to the features in the target cost function, but the sub-costs of these features are all the same because the candidate units all come from the same unit cluster (with the same features) in the decision tree. The use of these features in the target cost function has no effect in improving the performance of unit selection.

There are also some researches attempting to optimize unit selection using perceptual measures. The correlation between join cost and MOS is discussed to perceptually optimize the cost function [5]. The perceptual relevance of the distance measures between speech units is studied for concatenative based speech synthesis [6].

In this paper, we propose a perceptual clustering based unit selection optimization method for concatenative TTS synthesis. We attempt to identify the priorities for different features and try to optimize the *target* cost function with perceptual clustering. Our approach clusters the units using hierarchical clustering [7] method based on a perceptual distance measurement between different speech units. A method to identify the questions (concerning the phonetic and prosodic features) is then proposed to build the decision tree from the clustering result. The features used in the decision tree are the preferred ones, and the other features are used in the *target* cost function. Instead of regression training described in [1], linear discriminant analysis (LDA) [8] is used to train the weights for *target* cost function from the clustering result to make the weights more reasonable and perceptual related.

The paper is organized as follows. Section 2 describes the proposed framework for optimizing unit selection with perceptual clustering. Details include how to perform perceptual clustering with hierarchical clustering algorithm, the method for generating decision tree from perceptual clustering result, and the weight training with LDA for the target cost function. Experiments and results are presented in section 3. Finally, Section 4 concludes the paper.

2. THE PROPOSED METHOD

2.1. Framework

Figure 1 shows the framework of our proposed method for optimizing unit selection with perceptual clustering.

In the training stage, the speech units of the same tonal pinyin (i.e. pinyin with tone) from the speech inventory are clustered based on the perceptual distance measure between speech units using the hierarchical clustering algorithm. The decision tree building and the training of weights for target cost function are both based on the perceptual clustering result. The decision tree building algorithm then identifies a minimum set of features out of all phonetic and prosodic features, and constructs a map connection between these features and the unit clusters. The other features (not used in decision tree) are then used for calculating the target cost. The weights of different features in the target cost function are determined by LDA according to the hierarchical clustering result. As for the weights in the join cost function, they are trained using the regression method similar to the one in [1].

In synthesis, decision tree is first used to find the best cluster of candidate units, and the final unit selection results will be offered by the Viterbi algorithm that minimizes the cost function considering both *target* and *join* costs.

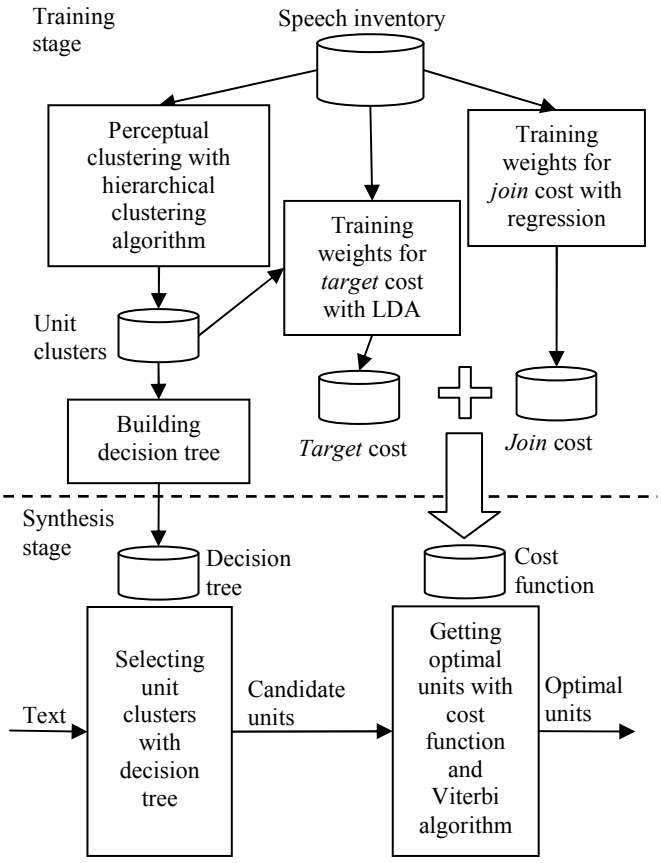

Fig.1. Framework of the proposed method for unit selection with perceptual clustering.

2.2. Perceptual clustering

The most important issue of perceptual clustering is how to choose the acoustic measure. Mean opinion score (MOS), the measure used in [5], is an optimal choice which surely represents the perception. However, it is prohibitively expensive to get the MOS scores by human perception. In attempt to measure the distance and quality of speech using objective methods, a variety of techniques have been reported in both speech synthesis and recognition areas. In the discussion of [6], mel-cepstral distance [9] and Itakura distance [10] both have reasonable correlation with perceptual evaluation. On account of the widely using of cepstral distance in speech synthesis, we use mel-cepstral distance $MCDist(U,V)$ [9] as an estimate of the perceptual distance between two speech units U and V:

$$MCDist(U,V)=\frac{1}{N}\sum_{k=1}^{N}\sqrt{\sum_{i=1}^{16}\left[MC_U(i,k)-MC_V(i,k)\right]^2} \quad (1)$$

where N is the number of the frames in the speech unit, $MC_U(i,k)$ and $MC_V(i,k)$ are the i-th mel-cepstral coefficients of the k-th frame of the speech unit U and V respectively.

Given a set of N units to be clustered, and an $N\times N$ *MCDist* matrix, the hierarchical clustering algorithm [7] is used to cluster the speech units. To guarantee that the units in one cluster are all perceptually similar, the complete-linkage clustering is used which defines the cluster distance $CDist(X,Y)$ between clusters X and Y as the maximum distance from any unit U of cluster X to any unit V of Y:

$$CDist(X,Y)=\max\{MCDist(U,V)\mid U\in X, V\in Y\} \quad (2)$$

The hierarchical clustering merges the closest pair of clusters iteratively, and produces a hierarchical tree indicating the merging relation between clusters. A simple hierarchical tree is shown in Figure 2.

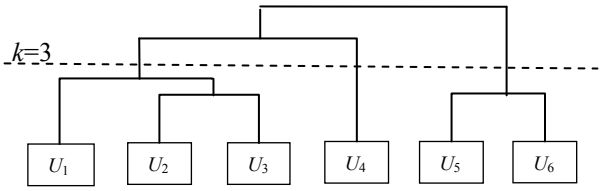

Fig.2. Hierarchical tree of 6 units and 3 clusters.

The final step of clustering is to decide the number of clusters. The longest k-1 links of the hierarchical tree can be cut to get k clusters, and the number of k can be determined by maximizing the Dunn's index for cluster validation [11]. The Dunn's validation index $DVI(C)$ of cluster set C is defined as the ratio of the minimum external distance to the maximum internal distance of clusters:

$$DVI(C)=\frac{\min\{CDist(X,Y)\mid X\in C, Y\in C\}}{\max\{\max\{MCDist(U,V)\mid U,V\in X\}\mid X\in C\}} \quad (3)$$

With the hierarchical tree, it is convenient to calculate the $DVI(C)$. Assume that the length of the k-th longest link is $Len(k)$, then the minimum external distance of k clusters equals to $Len(k\text{-}1)$, and the maximum internal distance equals to $Len(k)$. Therefore the Dunn's index of k clusters is:

$$DVI(k)=\frac{Len(k-1)}{Len(k)} \quad (4)$$

The optimal number of clusters k can be obtained by maximizing $DVI(k)$ of every possible k.

By this way, the unit clusters are generated and are highly related with the perceptual measurement. The unit clusters is the basis of the two following algorithms.

2.3. Decision tree building

The decision tree is built to use a minimum set of phonetic and prosodic features for choosing a unit cluster. In this work, the decision tree is built from the hierarchical tree of the perceptual clustering result, and it actually establishes a relation between the acoustic perception and phonetic and prosodic features of the unit.

Building the decision tree might be computationally expensive, but the agglomerative hierarchical tree makes it easier. The top k levels of the hierarchical tree can be used as the structure of the decision tree, and the problem is how to choose questions for each decision node of the tree. Our solution is based on two hypotheses: 1) in a well explained unit cluster with given phonetic and prosodic features, the two children of a non-root node must have same values of some features so that they are perceptually similar and agglomerated with minimum cluster distance; and 2) there must be at least one feature to make a distinction between two children of a node because they have perceptual differences. If these two conditions are not satisfied, it might relate with the outlier units which should be removed from the speech inventory; or the features are insufficient to represent context completely and new features are required.

Assume that N features are X_i (i=1,2,…,N), the feature vector of a unit can be expressed as $(x_1,x_2,\ldots,x_i,\ldots,x_N)$, where x_i is a value of feature X_i. To choose appropriate questions for each node, the decision vectors of the nodes can be defined as $(d_1,d_2,\ldots,d_i,\ldots,d_N)$, where d_i (i=1,2,…,N) can be a *value* of feature X_i, a "-" to eliminate feature X_i, or a *set* of possible values of feature X_i expressed as $[v_1,v_2,\ldots]$. To take a simple example, the decision vectors of the hierarchical tree with 4 features are shown in Figure 3.

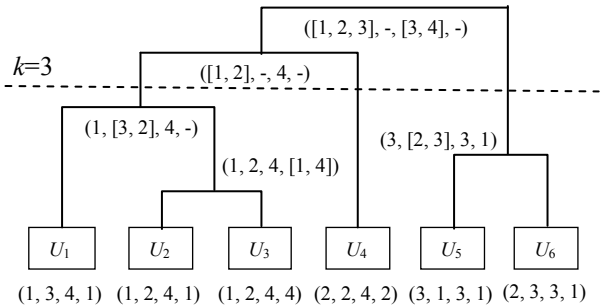

Fig.3. Decision vectors of nodes with 4 features. The elements of the decision vectors might be a 'value' type, a 'set' type (with […]) or a '-'.

To build the decision tree, the decision vectors of each nodes of the hierarchical tree are first computed; the feature of the decision question is then chosen for each node from the possible value sets in the decision vector. A standard greedy algorithm is used to select the decision questions that are related to the minimum set of features. The steps for selecting decision questions are as follows:

1) For each node of the hierarchical tree from the leaves to the root, compute the decision vector as steps a-b.

(a) Initialize the decision vectors for all leaf nodes of the tree so that the decision vector equals to the feature vector of the node;

(b) For a non-leaf node C, if the decision vectors for its two children C_l and C_r are $(d_{l1},\ldots,d_{li},\ldots,d_{lN})$ and $(d_{r1},\ldots,$

$d_{ri},\ldots,d_{rN}$) respectively, the decision vector ($d_l,\ldots,d_i,\ldots,d_N$) for node C is computed by comparing the decision vectors of the two children as:

$$d_i = \begin{cases} d_{li} & \text{if } d_{li} = d_{ri} \\ d_{li} \cup d_{ri} & \text{if } d_{li} \neq d_{ri} \text{ or } d_{li} \cap d_{ri} = \varnothing \\ - & \text{otherwise} \end{cases} \quad (5)$$

2) For each node of the tree from root to the nodes of level k-1, choose the decision question as steps c-e.

(c) Choose a *set* in the decision vector which has the maximum number of values, and assume the position of this set is i;

(d) For the feature X_i of this node, if the corresponding element of the two children's decision vector is d_{li} and d_{ri}, the decision question Q is computed as:

$$Q = \begin{cases} x_i = d_{li} & \text{if type}(d_{li}) = value \\ x_i = d_{ri} & \text{if type}(d_{ri}) = value \\ - & \text{otherwise} \end{cases} \quad (6)$$

where the condition "type(d_{li})=*value*" means that the type of the element d_{li} is *value* (not a *set* or "-");

(e) If Q is assigned as "-" after step d, it means that the feature X_i cannot be used as a decision feature for this node. Repeat steps c-e to try another feature.

Recall the example in Figure 3, the corresponding decision tree obtained by this algorithm is in Figure 4:

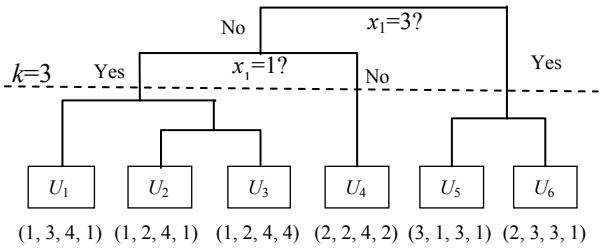

Fig.4. Decision tree generated from the hierarchical tree using the decision vectors in Fig.3.

2.4. Weight training for cost functions

The *join* cost measures the acoustic distortions at the boundary where two neighboring units are concatenated. The weight in join cost function is trained using the regression method as described in [1].

As for the *target* cost, we use LDA [8] to obtain the weights. The most important issue of training the weights is to find reliable perceptual index predicted by the *target* cost. In this work, unit clustering result is an appropriate choice for such perceptual index, because the clustering reasonably correlates with acoustic perception.

Assume that a set of M features is used in the *target* cost, which is a subset of the N features. The feature vector of a unit used in *target* cost can be expressed as ($x_1,x_2,\ldots,x_M$). The cluster No. i of a unit serves as the perceptual index to be predicted by the *target* cost. LDA is used to train the weights of features by performing dimension reduction while preserving as much of the class discriminatory information as possible. Because the objective of LDA is to perform dimension reduction while maximizing the separability of the units, the distance predicted by this *target* cost should be small in a cluster and large between clusters, and that is more reasonable correlated with acoustic perception that distinguishes the units into clusters.

The weights for different tonal pinyin are generated separately as the influence of phonetic and prosodic context features might be different with different tonal pinyin.

3. EXPERIMENTS AND RESULTS

The experiment is based on our homegrown Crystal [4] TTS system with a Chinese speech inventory of 25,000 sentences with text scripts extracted from the year 2000 People's Daily newspaper. There are 1,694 tonal pinyins in Mandarin in total. The units for each tonal pinyin are clustered using the hierarchical clustering algorithm, and a decision tree is build based on the unit clustering result for each tonal pinyin. Finally, 1,694 decision trees are built based on the unit clustering result.

The average number of the clusters (i.e. leaves) for all the decision trees is 7.68. And 89.3% of the decision trees have less than 16 clusters (leaves). The average number of features used in the decision trees is 3.57. This means most of the decision trees can use only 4 features to distinguish the unit clusters related to them.

The candidate features for building the decision tree include 30 features concerning 3 types of phonetic or prosodic information: the boundary types of the current syllable, the tonal pinyin and related features of the previous and next syllable, and the index, length and position information at different prosodic levels including prosodic word, prosodic phrase and utterance.

The features used in all the decision trees are summarized in Table 1, where UFC shows statistics of the frequency count of each feature used in all decision trees. There are only 14 features used in all the decision trees, and 8 of them are used much more frequently than the others. These 8 features are considered as preferred features, and the other 22 features are used in the *target* cost function.

Finally, a perceptual pair comparison (PC) test was conducted to compare the naturalness of the synthesized speech with this new unit selection method against the old one. A set of 20 test sentences with varying length from 10 to 30 were selected randomly from the year 1998 People's Daily newspaper and the news report on the internet of July 2011. The sentences are also checked to ensure that only Chinese characters and Arabic numerals are contained to avoid text normalization errors. 40 files of the synthesized speeches were generated from these 20 sentences by the two versions of the systems. And these 40 files were presented to subjects for listening test, and the order of the speech files in each pair is randomized. 10 subjects were invited in this

PC test. The subject was asked to listen to the 20 pairs of speech files and select which file is better.

Figure 5 shows the result of the perceptual PC test. Out of the 200 possible tests (20sentences×10subjects), 148 tests (74%) prefer the synthetic result by the new method. The perceptual test shows that the newly proposed method for unit selection has a better performance in generating speech with higher naturalness.

Table.1. Features used in all the decision trees.
(UFC: frequency count used in decision trees)

Feature	UFC	Note
p.py	3733	Pinyin of previous syllable
n.py	3429	Pinyin of next syllable
sylTuttP	1150	Relative position of syllable in utterance, meaning percentage of the position of current syllable over the whole length of utterance
sylTpphP	708	Relative position of syllable in prosodic phrase
pboud	521	Previous boundary type of current syllable
nbound	484	Next boundary type of current syllable
sylTpphL	271	Syllables count in prosodic phrase
sylTuttL	116	Syllables count in utterance
sylTuttF	18	Syllable index from head of utterance
sylTpphF	17	Syllable index from head of prosodic phrase
sylTpwdL	14	Syllable count in prosodic word
pwdTuttL	13	Prosodic word count in utterance
pwdTuttP	12	Relative position of prosodic word in utterance
sylTpwdF	10	Syllable index from head of prosodic word

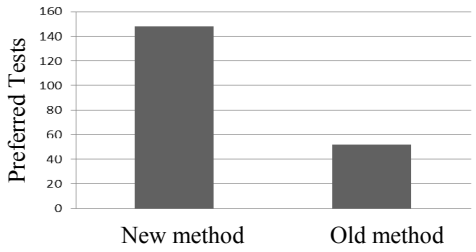

Fig.5. Preference of the two methods in PC test.

4. CONCLUSIONS AND FUTURE WORK

This paper describes a new view of unit selection by considering features with different priorities according to the perceptual clustering result and proposes a perceptual clustering based unit selection optimization method for concatenative based TTS synthesis.

Our approach uses mel-based cepstral distance that is correlated to perceptual and hierarchical clustering method to cluster speech units. Based on the unit clusters, the features of each unit cluster, concerning the phonetic and prosodic context, are screened to produce an optimal decision tree that uses least number of features to distinguish and select unit clusters. The features used in the decision tree are the preferred ones, and the other features are used for calculating the *target* cost. The weights of different features in *target* cost are determined by LDA using the non-preferred features of a unit and the unit cluster information it belongs to. During synthesis, decision tree is first used to find the best clusters with the preferred features, and the final unit selection results will be derived by the Viterbi search to minimize the target and join costs considering the non-preferred features. The perceptual pair comparison test indicates that the proposed method (i.e. the perceptual clustering optimized unit selection) can generate speech with higher naturalness than previous method.

We note an additional effect of the proposed clustering-based method. The bad units that do not belong to any of the clusters can be removed, which might be helpful for cleaning the speech inventory and improving the quality of the synthetic result, and we will test it in future work.

5. ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (60928005, 60805008, 60931160443 and 61003094), the Ph.D. Programs Foundation of Ministry of Education of China (200800031015) and the Science and Technology R&D Funding of the Shenzhen Municipal.

6. REFERENCES

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using large speech database", In: *Proc. of ICASSP*, pp.373 -376 1996.

[2] J. Vepa and S. King, "Join cost for unit selection speech synthesis", *Text to speech synthesis*, S. Narayana, A. Alwan, Eds. Prentice Hall, 2004.

[3] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," In: *Proc. of Eurospeech*, pp. 601-604, 1997.

[4] Z. Wu, G. Cao, H. Meng and L. Cai, "A unified framework for multilingual text-to-speech synthesis with SSML specification as interface", *Tsinghua Science and Technology*, vol.14, no.5, pp.623-630, 2009.

[5] H. Peng, Y. Zhao and M. Chu, "Perpetually optimizing the cost function for unit selection in a TTS system with one single run of MOS evaluation", In: *Proc. of ICSLP*, 2002.

[6] J. Wouters and M. Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis", In: *Proc. of ICSLP*, vol. 6, pp.2747-2750, 1998.

[7] S.C. Johnson, "Hierarchical clustering schemes", *Psychometrika*, vol. 32, pp.241, 1967.

[8] R.A. Fisher, "The use of multiple measures in taxonomic problems", *Ann. Eugenics*, vol. 7, pp.179-188, 1936.

[9] R. F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment", In: *Proc. IEEE Pacific Rim Conf. Communications, Computers and Signal Processing*, 1993.

[10] F. Itakura and T. Umezaki, "Distance measure for speech recognition based on the smoothed group delay spectrum", In: *Proc. of ICASSP*, pp.1257-1260, 1987.

[11] J.C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters", *Jour. Cybemetics*, vol.3, no.3, pp.32-57, 1973.