SCRIPT DESIGN BASED ON DECISION TREE WITH CONTEXT VECTOR AND ACOUSTIC DISTANCE FOR MANDARIN TTS

Dandan Cui[1, 2]*, Dezhi Huang*[2]*, Yuan Dong*[2]*, Lianhong Cai*[1]*, and Haila Wang*[2]

[1]Key Laboratory of Pervasive Computing (Tsinghua University), Ministry of Education, China
[2]France Telecom R&D Beijing Co., Ltd.

ABSTRACT

The performance of a TTS system relies on the acoustic features of output speech, while we have only the linguistic information in text context for the corpus design phase. To reconcile that conflict, this paper proposes a design method based on decision tree. First, trees are trained using an existing speech corpus: the splitting questions are selected from context vectors and the distance metric is based on acoustic features. Then units from a new text corpus which will be the source of target script are inserted into the trees. The sentence selection strategy is a combination of coverage of major tree nodes and frequent context vector clusters. Experiment is carried out by collecting a Beijing Olympics related corpus taking advantage of an existing domain-unspecified speech corpus and supplementary domain-specified text. Informal listening test on TTS outputs confirms that the proposed method achieves a better optimization of speech. It includes not only phonetic balance as conventional methods do, but prosodic balance as well.

***Index Terms*—**Speech synthesis, clustering methods

1. INTRODUCTION

Script design of speech corpus is one of the key issues in building high quality concatenative TTS system. However, intrinsic contradiction lies as mentioned above. The problem is: what really matter in speech corpus is the acoustic features, while what we can control in the text script is merely the contexts. So, the essential problem for script design to resolve is to obtain optimal acoustic products via optimizing the text context.

Script design was initially considered as a set-cover problem with the classical method “Greedy” [1]. As the corpus and feature set expand, the greedy algorithm becomes barely competent. Then Bozkurt et al. extended the ability of classical greedy by weighting the features and events, but the problem of weight-setting was left unsettled [2]. Li et al. used a forest structure to store statistical information of context vectors (henceforth simply CV), but the empirical splitting criteria and uniform treatment for all syllables were still unsatisfactory [3]. Kawai et al. included predicted F0 and duration in metric of coverage, i.e. prosody was introduced for the first time [4]. Yet, suffered by the limited ability of prosody predicting, the method was not paid much attention. In fact, it regarded the prosody and context as parallel in the metric which seems unreasonable or might seek far and neglect what lies close at hand.

As we see, current methods can’t resolve the barrier satisfactorily. Especially, some crucial facts are not included yet: context influences different syllables in different ways [5]; interaction exists between the influences of different context features [6]. The best way out is to drive the context optimization by acoustic behavior directly. We present such a method: it takes advantage of an existing speech corpus to train decision trees, builds a new forest of CV clusters by inserting syllables of a new text corpus which will be the source of target script, and combines the covering of major nodes and frequent CV clusters in selection strategy.

This method is based on the premise that context influence on the acoustic is relatively stable between different text contents and domains for the same speaking style, and meets the fact that, nowadays, in most cases, there are existing corpora, which are just not rich or balanced enough, or can not match a new domain, i.e. usually, we don’t have to build a corpus from zero. They are proved by the result of our experiment.

The rest of this paper is organized as follows. Section 2 describes the approach, from tree training to sentence selection. Then section 3 carries out an experiment in which speech corpora are collected and compared by TTS products. Conclusion is given in section 4.

2. DESCRIPTION OF THE APPROACH

The process flow of our approach is outlined in Fig. 1. The inputs include an existing speech corpus (henceforth old corpus) and a new text corpus with the target domain or richness (henceforth text corpus).

The acoustic driving is achieved by building decision-trees from old corpus. Then, a CV forest is built by inserting new text into the trees, with adjustment if necessary. Finally, Sentence selector travels the forest iteratively to cover both major tree nodes and frequent CVs clusters.

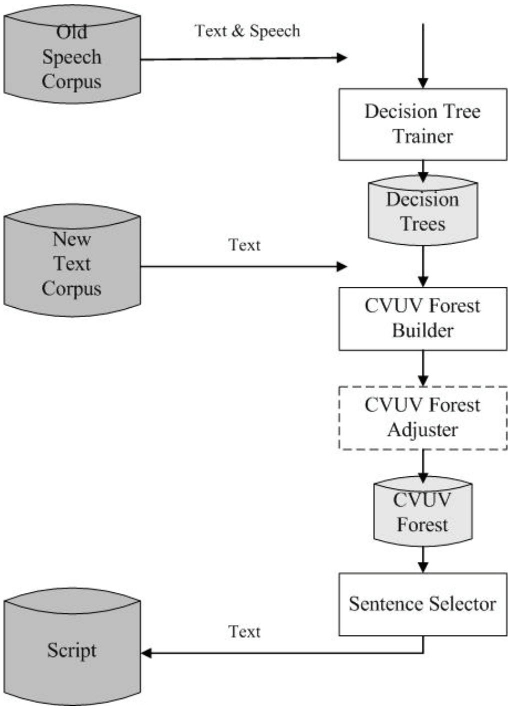

Fig. 1. Process flow of the proposed script design approach

In addition, preprocess of text may be needed, mostly using a TTS front end. The basic unit can be customized. Here we choose syllable as basic unit for Mandarin TTS corpus.

2.1. Decision tree building

Decision tree which can cluster the units hierarchically has proved helpful in TTS unit pre-selection and corpus reduction [7]. Here, CART (Classification and Regression Trees) is used. It is a widely used model of binary decision trees. It splits the initial set of speech samples of each syllable top-down; therefore obtain the clusters containing samples with similar acoustic features and the hierarchical relationship between context features and acoustic similarity of speech samples synchronously, i.e. difference of context influencing ways between syllables and interaction between influences of different context features are included.

During the splitting, every branch node is a yes/no question of a context feature, e.g. Left Final Class = 1. The principle for question selection is to maximize the increase of purity which is based on the acoustic distance.

Considering the mainstream technology of nowaday commercial TTS systems, the acoustic features employed here are prosodic ones including F0, duration and energy.

For our Mandarin TTS Corpus, context features we used are similar with the CV used in literature [3]. It is a 4-dimension CV definition:

- Left Tone Context (6 values, including 5 tones “1”~“5” and “0” for phrase boundary)
- Right Tone Context (6 values as well)
- Left Final Class (the category of the previous syllable Final, 13 categories and a phrase boundary)
- Right Initial Class (the category of the following syllable Initial, 20 categories and a phrase boundary)

The initial/final classes are also used in adjustment of sparse trees. Detailed definition is stated in literature [8].

2.2. CV forest building and adjustment

After the decision trees are trained, each syllable sample in the new text corpus which will be the source of sentences in the target speech corpus is inserted into a leaf of its corresponding tree according to its CV. Statistical information is also counted on each node.

After the inserting, adjustment is often necessary because the old corpus is imperfect, and vocabulary always changes for a new text set and a new domain.

Dealing with the sparse trees: For the old corpus may be not as rich or balance, often, there are sparse syllables with no branch. They are dealt according to their tones, initial/final classes and tree structures of similar syllables.

Calculate the average number of samples per leaf n'.

If tree t_i has n samples while with no branch, the target level of the tree is calculated with equation (1):

$$l' = \lfloor \log_2 (n/n') \rfloor \qquad (1)$$

If $l' > 0$, the tree is split up to l' levels using the following approach. Otherwise, it's a rare syllable; there is no need to split it.

I. Calculate the importance of each context x in T_r. T_r is the set of trees with the same Final Class and tone as t_i. Equation (2) calculates the importance of context x in an individual tree $t_i \in T_r$ and (3) calculates the overall one in T_r. Same tone only if none, same Final Class if still none. Here $leaf_k\ from\ branch_j$ means leaf k is split from branch j directly or indirectly. Level of root is “1”.

$$IF_i(x) = \frac{\sum_{(branch_j\ with\ x)} \sum_{(leaf_k\ from\ branch_j)} \frac{1}{level\ of\ leaf_k}}{count\,(leaves\ in\ t_i)} \qquad (2)$$

$$IF(x) = \frac{1}{\sum_i count\,(leaves\ in\ t_i)} \sum_i (IF_i(x) \times count\,(leaves\ in\ t_i)) \qquad (3)$$

II. Calculate the importance of each context x in T_l. T_l is the set of trees with the same Initial Class and tone as t_i. Initial Class only if none, tone if still none. The priority is according to statistics of all syllables using equation (2) and (3). Details will be available in literature [9], another paper from the authors.

III. Use the right contexts in T_r and left contexts in T_l to split t_i: Sort the contexts by descending importance. If there's t_i in $T_r \cup T_l$ which has $branch_{j2}\ with\ x_2$ split directly or indirectly from $branch_{j1}\ with\ x_1$ in “yes” condition, insert x_2 as “yes” child of x_1, else “no”.

IV. Terminate if have split l' levels or all contexts.

2.3. Sentence selection

After all the above steps, the CV forest is ready. Thus sentence selection can start. As we have stated, the sentence selection strategy is a combination of coverage of major tree nodes in the forest and frequent CV clusters in the new text corpus. That is achieved in the following steps.

I. Travel the forest once to cover all the syllables by selecting the leaf which has the largest number of samples.

II. Travel the forest to cover the major nodes of trees by selecting its largest leaf for each node. It is an iterative process: covering level by level in top-down sequence.

III. Terminate if the target size of script is reached.

This method is somewhat time-consuming, yet tolerable for an offline application on current computers. Experiment confirms that speech corpus designed with this method has improved TTS performance. Details are in section 3.

3. EXPERIMENT

To evaluate whether the proposed method has fulfilled its aim of achieving optimization of acoustic products via optimizing the text context, an experiment is conducted. The task is to design a new speech corpus for Beijing Olympics related applications. The source data is an existing domain-unspecified speech corpus and supplementary Olympic-related text. For comparison, corpora are designed with both the proposed method and a rule-based tree method, and collected. Then synthesis experiment is carried out. Details are presented in the following subsections.

3.1. Source data

The old speech corpus is the TH-CoSS corpus developed in Tsinghua University. It's a Mandarin speech corpus for TTS system building, testing and speech analysis. It has been used by several organizations inside and outside China. It is narrated in a formal reading style and is domain unspecified. Here, among its several sub-corpora, we use the TTS building sub-corpus from a female speaker. It contains 5433 declarative sentences and is designed via Greedy algorithm.

The new text corpus is composed of 2817 sentences of text collected from internet. The content includes all kinds of information that is all related to Beijing Olympics: sports, weather, sight-seeing, traffic, etc.

Sentence selection is based on the aggregate of the above two test sets.

Table 1. The (major) context features for the basic levels of trees in the two tree-based methods.

	Set 2	Set 3
Level 2	Left Tone	Right Tone
Level 3	Right Tone	Left Tone

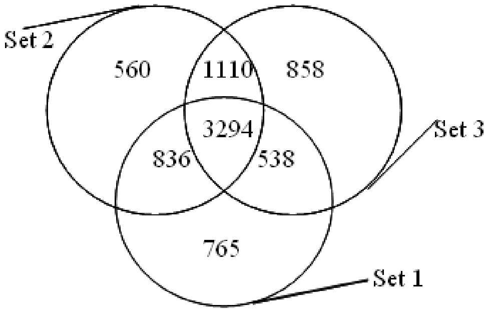

Fig. 2. Overlap between the 3 generated script sets. The circle "Set 1" denotes the script set of old corpus TH-CoSS. The circle "Set 2" denotes the script set designed by the rule-based tree method. The circle "Set 3" denotes the script set designed by our decision tree based method.

3.2. Script design and corpus collection

For comparison, script are designed by both the proposed method and the rule-based tree method in literature [3], and collected. The target size is set as 5433, i.e. the size of the old corpus. 2 forests are built via both methods with a preference of Olympic-related text. For the decision tree based approach, 1747 trees are built, with 517 sparse ones, among which 38 are split by up to 4 levels. For the 3 sets illustrated in Fig. 2, as most of the sentences in script set 2 and 3 are shorter than those in set 1, we increase the size of set 2 and 3 to 5800 to balance the number of syllable samples. As Fig. 2 shows, the script designed by decision-tree-based method has a difference of 1396 sentences (22%) with the rule-based.

Context features for the basic 2 levels in the tree-based methods is compared in Table 1: the statistics of trees from a real speech corpus is against the rule. Both prove that the real context influence on acoustic is far from the rule.

Then, the 2528 sentences which are in set 2 or set 3, but not in set 1 are narrated by the same female announcer and in a same formal reading style. Thus, 3 speech corpora with comparable size are ready for synthesis experiment.

3.3. Synthesis experiment

The 3 corpora are then integrated into a concatenative Mandarin TTS system so called BaiLing, developed in France Telecom R&D Beijing, separately. 50 sentences are inputted in to the system and synthesized: 25 are Olympic-related, collected from internet; 25 are domain-unspecified, selected from newspaper. All test sentences have not appeared in the source data in 3.1.

Therefore we have 150 synthesized speech utterances, 50 from each speech corpus. They are played in 50 groups according to text content and in random sequence within the group. 12 college students are asked to listen and give a score of preference as "1", "2", and "3" ("1" for most preferred and "3" for least).

Table 2. The result of listening experiment, presented in average score of preference. "Corpus 1" is the script set of old corpus TH-CoSS. "Corpus 2" is the script set designed by the rule-based tree method. "Corpus 3" is the script set designed by the proposed method. "Test text 1" means the 25 Olympic-related sentences. "Test text 2" means the 25 domain-unspecified sentences.

	Corpus 1	Corpus 2	Corpus 3
Test text 1	2.25	1.87	1.61
Test text 2	2.34	1.89	1.37
Overall	2.29	1.88	1.49

Average scores are presented in Table 2. Both tree-based corpora are more preferred than the old speech corpus: the hierarchical structure of context influence proves superior. Among the 3 corpora, the decision-tree-based one is most preferred, especially for the Olympic-related sentences: the proposed approach has achieved a better optimization of target speech and it has included the new vocabulary more nicely. In addition, the premise that contextual influence on the acoustic is relatively stable between different contents and domains for the same speaking style is also confirmed.

4. CONCLUSION

This paper proposes a script design method of TTS speech corpus that tries to drive the context optimization by acoustic attributes directly using a decision-tree.

- Trees are pre-trained based on an existing speech corpus: the splitting questions are context features and the distance metric is based on prosodic features.
- Then information from a new text corpus is inserted into the trees to build a CV forest, from which the target script will be generated. Adjustment of the forest will also be conducted if necessary.
- The sentence selection strategy is a combination of coverage of major tree nodes and frequent context vector clusters.

In the experiment of collecting a Beijing Olympics related corpus, an existing domain-unspecified speech corpus TH-CoSS are taken advantage of, and supplementary domain-specified new text as well. the script designed by decision-tree-based method shows about 22% difference with the rule-based tree method, and the (major) context features for basic levels of trees in the two methods are quite the contrary. Statistical results indicate that the real context influence on acoustic is far from the rule. As predicted, corpus design guidance from real speech leads to improvements.

The result of listening test on TTS outputs shows: the proposed method exceeds the rule-based tree method from literature [3] and the old speech corpus TH-CoSS designed via classical Greedy algorithm. The decision-tree based approach includes not only phonetic balance as conventional methods do, but also prosodic balance.

As we see, the whole process is acoustic-driven. The barrier between the final modality of speech for a TTS corpus and the unavailability of speech in script design phase is thus resolved. The premise that contextual influence on the acoustic is relatively stable between different contents and domains for the same speaking style also proves to work.

The method is especially effective in such applications as there is an existing corpus, which is not rich or balanced enough for a new task, or can not match a new domain.

ACKNOWLEDGEMENTS

The work in this paper is supported by National Science Foundation of China (No. 60433030, No. 60418012) and the Special Funds for Major State Basic Research Program of China (973 Program) (No. 2006CB303101).

REFERENCES

[1] Chvatal, V. "A Greedy Heuristic for the Set-covering Problem," in *Mathematics of Operations Research, vol.4, no.3*, pp.233-235, 1979.

[2] Baris Bozkurt, Ozlem Ozturk, and Thierry Dutoit, "Text Design for TTS Speech Corpus Building Using a Modified Greedy Selection," in *Proc. Eurospeech 2003*, pp. 277-280., 2003.

[3] Haiping Li, Fangxin Chen, and Li Qin Shen, "Generating Script Using Statistical Information of the Context Variation Unit Vector," in *Proc. ICSLP 2002*, pp.117-120., 2002.

[4] Hisashi Kawai, Seiichi Yamamoto, Norio Higuchi, and Tohru Shimizu, "A Design Method of Speech Corpus for Text-to-speech Synthesis Taking Account of Prosody," in *Proc. ICSLP 2000, vol.3*, pp.420-425, 2000.

[5] Zongji Wu, Maocan Lin, *Introduction to Experimental Phonetics*, Higher Education Express, Beijing, China, 1988.

[6] Zheng Yuling, Cao Jianfen, Bao Huaiqiao ,"Coarticulation and prosodic hierarchy," In *Proc. TAL 2006*, pp.145-150, 2006.

[7] Blouin, C., Bagshaw, P.C., Rosec, O., "A Method of Unit Pre-selection for Speech Synthesis Based on Acoustic Clustering and Decision Trees," in *Proc. ICASSP 2003, vol.1* , pp.692-695, 2003.

[8] Zhang J., Lu S., and Qi S., "A Cluster Analysis of the Perceptual Features of Chinese Speech Sounds," *J. Chinese Lingustic, vol. 10*, pp.189-206, 1982.

[9] Dandan Cui, Lianhong Cai, "Corpus Analysis Based on Decision Tree," *Computer Engineering, vol. 32*, pp.3-5, 2006.