Constructing Scalable TTS System based on Corpus Approach*

ZHANG Wei[1]
1, Department of Computer Science
Ocean University of China
Qingdao, Shandong, 266100
e-mail: weizhang@ouc.edu.cn

LING Zheng-hua[2], DAI Li-rong[2]
2, Department of Electronic Engineering and Information Science
University of Science & Technology of China
Hefei, Anhui, 230027

Abstract—**Pruning redundant synthesis instances or tailoring TTS voice font is an important issue of Corpus-based TTS. But pruning redundant synthesis instances, usually results in loss of non-uniform. In order to solve this problem, this paper proposes the concept of virtual non-uniform. According to this concept and the synthesis frequency of each instance, the algorithm named StaRp-VPA is constructed as to make up the loss of non-uniform. In experiments, the naturalness scored by MOS remains almost unchanged when less than 50% instances are pruned off, and the MOS does not severely degrade when reduction rate is above 50%.**

Keywords—**text-to-speech system; speech synthesis; scalable speech synthesis system; scalable text-to-speech system**

I. Introduction

Corpus-Based approach, or Selection-based approach, is a successful technology and has been applied in most state-of-art Text-to-Speech systems [1]~[4]. This approach can generate highly natural speech due to its utilizing not only digital signal process techniques but also data-driven techniques from knowledge discovery and data mining.

Generally, the basic unit chosen for synthesis is syllable in Mandarin or Cantonese. When being synthesized, proper syllables are selected from a very large speech database by Viterbi [5] algorithm. In database, all recorded speeches are index by trees, named non-uniform units. A non-uniform unit includes one syllable or several succeeding syllables. Acoustic instances (variants or voice fonts) belongs to same non-uniform are indexed to a tree according to their prosody, phonetic and part of acoustic contexts. The tree, which is named non-uniform tree, is constructed by clustering (generally using CART [6] approach for clustering) instances based on questions concerning prosody, phonetic and part of acoustic context. Figure 1 gives a example of non-uniform trees.

With Corpus-based TTS systems, speech synthesis becomes a problem of collecting, annotating, indexing and retrieving from a very large speech database [1]~[4]. In order to synthesize nature-sound speech, several or even tens of hours speech waveforms are required from diverse text input. Thus, storing, loading and searching such a huge corpus become inevitable issues in many applications. Because of these reasons, Corpus-based TTS usually requires high performance hard wares to synthesis natural-sound speech. If there is the approach of keeping the naturalness Corpus-based TTS but properly shrinking the speech database, the Corpus-based TTS will be more flexible and scalable to all kinds of hard wares. This problem is called pruning redundant synthesis instances, or tailoring TTS voice font.

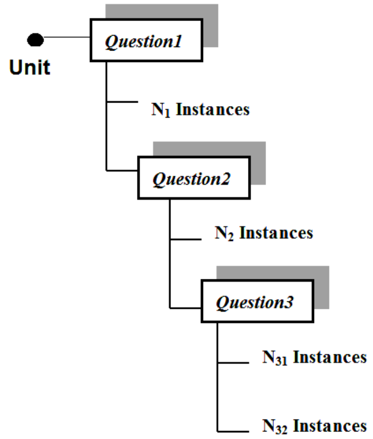


Figure 1: non-uniform tree

There is redundancy in speech database in deed. For example, some instances are almost never used for synthesis. While some instances can be replaced each other. Several approaches for reducing redundant synthesis instances have been proposed. The approach described in [7] clusters similar units (diphones) with a decision tree that asks questions concerning prosodic and phonetic context. Units that are furthest from the cluster center are pruned. It claimed that pruning up to 50% of units produced no serious degradation in speech quality. The method proposed in [8] is based on a unified HMM framework. Only instances (single or multiple) with the highest HMM scores are kept to represent a cluster of similar ones. Kim et al. presented a weighted vector quantization (WVQ) method that prunes the least important instances [9], [10]. 50% reduction rate is reached without significant distortions. In paper [11], Rutten et al. proposed a database reduction technique based on the statistical behavior of unit selection. They claimed that pruning the database down to 50% of its original size without a significant drop in the

* Supported by of National Science Foundation of China (No. 60602017).

output speech quality. Zhao et al. [12] proposed prosodic outlier criterion, the importance criterion and the combination of the two, and pruning voice fonts with those criterions. According to their paper, the naturalness remained almost unchanged when 50% of instances were pruned off with the combined criterions. We have done researches on clustering synthesis-instances-pruning approach of embedding system [13].

Non-uniform is very important concepts in corpus-based TTS. Non-uniforms of different granular increase the matching between the texts being synthesized and the corpus of speech database. In sense of speech naturalness, quality of non-uniforms determines the TTS systems performance. So almost all state-of-arts Corpus-based TTS systems utilize non-uniform technique. But tailoring TTS voice font, or pruning redundant synthesis instances, usually results in loss of non-uniforms. All pruning methods above haven't mentioned this opened problem.

In order to solve this problem, this paper proposes the concept of virtual non-uniform. According to this concept and the synthesis frequency of each instance, the algorithm named StaRp-VPA is constructed on KBCE[1] TTS systems as to make up the loss of non-uniform. In experiments, the naturalness scored by MOS remains almost unchanged when less than 50% instances are pruned off, and the MOS does not severely degrade when reduction rate is above 50%.

This paper is organized as follows: the key problems of synthesis instances pruning, the concepts of virtual non-uniform and are investigated in Section II. The framework of StaRp-VPA algorithm base on virtual non-uniform is described in Section III. Experiments and object/subject evaluations are discussed in Section IV. Final discussion is presented in Section V.

II. Problem investigation and virtual non-uniform

The redundancy of speech database can only exist in units (instances index-trees) or acoustic instances. Generally speaking, in mandarin or Cantonese, units are those words or succeeding syllables that frequently appears in literature [3]. Pruning a unit means Loss of some prosody and phonetic environment. Thus there is little chance of redundant units. In fact, Redundancy usually comes from redundant acoustic instances. Some redundant instances hardly selected by TTS system, so these instances should be pruned. Some redundant instances have little difference between each other, so it ought to reserve one and pruned others. Therefore, the purpose of this paper is to research how to removing redundant instances from speech database automatically and flexibly.

Concerning with pruning redundant instances there are two problems: (1) how many instances of a unit are redundant; (2) in a unit, which instances are the redundant. In another word, it is to say that how to determine the importance of instances which belongs to the same unit.

For problem one, because most state-of-art TTS adopt the framework of [3] or something likely, it can be considered that more instances in a unit means more redundancy. Thus, we propose an approach named vibration-rate pruning: Keep total reserving rate as configuration, tune the instances reserving rate of each unit according to their instances amount (More instances means smaller reserving rate), and at last properly offset those reserving rates which are too small.

There are two quantities can measure the importance of a instances: Firstly, the frequency of a instance selected by the TTS (Frequently selected instance must be reserved); Secondly, the capability of a instance replacing other instances (the instances that can replace more others instances means more important, and the replacing should include different granular non-uniforms). The importance measure is a function of these two quantities (a proper function is their product), named Instance-Importance-Score function. So for problem two, we arrange the instances of same unit according to their IIS, the instances have least IIS value are redundant instances.

Based on the discussion above, there are four key points need for more consideration.

1. Reserve rate compute of vibration-rate pruning. Suppose we want pruning the speech database to $\beta(0<\beta\leq 1)$ of origin. From the analysis of problem one, different unit i need different reserve rate g_i (pruning rate $t_i=1-g_i$), and the total reserve rate is equal to β. The g_i is computed as following:

Suppose the instances belong to unit i is pi of the total instances,

$$\sum_{i=1}^{I} p_i g_i = \beta$$

Let $p_i g_i=\beta/I$, so

$$g_i=\beta/I/p_i \tag{2. 1}$$

Equation2.1 shows g_i is in inverse proportion to pi. This means more instances smaller reserve rate, which consistent with the discussion of problem one. To those units whose β satisfy $\beta/I/p_i >1$, their $g_i=1$, The remnant is (2.2):

$$\sigma=\sum_{i=1}^{I}(\frac{\beta}{I}-p_i g_i)=\sum_{g_i=1}(\frac{\beta}{I}-p_i)+\sum_{g_i<1}(\frac{\beta}{I}-p_i g_i)$$

Expectation Probability Efr_i describes the probability of unit i appears in a text. The value of Efr_i can be well estimated by the statistical compute from large corpus (In this paper, we used all kinds of texts 300 MB to estimate each Efr_i). Current Frequency Ratio Sfr_i describes, during current iteration, the ratio of unit i 's frequency to the frequency of all units. The value of Sfr_i can be accurately computed by counting instances number of current speech database (Sfr_i is updated by each iteration of reserve rate compute).

[1] KBCE is the prototype TTS system of www.iflytek.com 's commercial TTS system named interphonic. KBCE is a Corpus-Based Continuous Chinese-English Text-to-Speech Engine, and has been awarded best performance in the routine china national “863” evaluation. Because KBCE used an ingenious non-uniform technique to synthesize high natural speech, it is of great importance that pruning approach shrinks the speech database of KBCE without severely degrading the naturalness.

$$x_i = Efr_i/Sfr_i$$

describes the gap between Sfr_i and Efr_i, xi is used for reserve rate offsetting (2. 3) to all $g_i<1$:

$$G_i = g_i + \sigma \frac{x_i}{\sum_{g_i<1} x_i} = \frac{\beta}{Ip_i} + \sigma \frac{x_i}{\sum_{g_i<1} x_i}$$

If $G_i \le 1$, $g_i=G_i$; $G_i>1$, $g_i=1$. Processes of offsetting iterates and terminate when reserve rates are less than or equal to 1. Reserve rate offsetting is to prevent over-pruning those instances that belongs to a unit including many instances. So offsetting keeps the prosody and phonetic integrality of original speech database.

2.Virtual non-uniform and Instance Importance Score. There are two parameters of IIS: usage and replacing score. The usage of instance L is defined as the loss of dynamic coverage after deleting the unit's leaf that instance L belongs to. Suppose the coverage (see paper [3] for computation detail) before deleting is A_{0L}, and A_L after deleting, thus usage of instance L is:

$$\alpha_L = (A_{0L}-A_L)/A_{0L}$$

Usage of instance is to weigh the importance from corpus. If the prosody and phonetic environment is same, the usages of different instances are the same. Otherwise, usages are usually different.

Pruning redundant synthesis instances, usually results in loss of non-uniform. In order to solve this problem, we introduce the concept of virtual non-uniform in the following content.

Let's remove a given instance from TTS system, then let TTS system selects a replacement of this instances using a measure or an algorithm, such as Viterbi, acoustic distance, trainable approach and so on (the instance itself is named replacement R_0). This replacement is named the 1st replacement R_1 of R_0. R_1 is not a real non-uniform; it's the best replacement of non-uniform R_0. In order to select the best one,

Selection only happens in instances with fidelity to original prosodic and phonetic environment of R_0. We name this process Speech Completion.

In similar way, remove R_0 and R_1, let TTS selects the 2nd replacement R_2 of R_0. Generally, remove the R_0, …, R_{N-1}, let TTS selects the Nth replacement R_N. Replacement R_i ($0 < i < N$) is named the Virtual Non-Uniform of instance R_0.

The measure or algorithm (in this paper, we use Viterbi) which TTS uses to select the replacement gives each R_K a cost Q_K, the cost just describes the difference between virtual non-uniform and real non-uniform， and satisfies monotonicity: $Q_0=0$, $Q_{K-1} \le Q_K$.

The score of each replacement:

$$M_K = \exp(-\frac{Q_K^2}{\sigma})$$

σ named width, is used to control the response range of Q_K. for original non-uniformR_0, $M_0=1$, because Q_K is monotonic, M_k satisfies monotonocity: $M_K \le M_{K-1} < 1 = M_0$。

Pruning synthesis instances is ultimately pruning redundant syllable instances, thus we should add the score of each replacement to the syllable instances, which compose those replacement. For example, a replacement $V_1V_2V_3$ (suppose the score of $V_1V_2V_3$ is M_F) is composed by V_1, V_2 and V_3. We should add to each of V_1, V_2 and V_3:

$$\alpha_F M_F$$

α_F is the usage of original real non-uniform $V_1V_2V_3$.

The Instance-Importance-Score (IIS for short) of a syllable instance m is:

$$S_m = \sum_{j \ge L_m}^{J} \eta_j mark_j \text{ , } mark_j = \sum_{n=0}^{N} F_n \text{ , } F_n = \alpha_n M_n$$

Here we consider syllable number of a instance as the length of this instance. M_n is the score of instance n; instance n is a length-j replacement (real or virtual non-uniform) that is composed by instance m. α_n is the usage of instance n, and F_n is the sum of weighted replacement score. If we considered α_n as a weight, $mark_j$ is the weighted sum of all length-j replacements that are composed by instance m. Here L_m is syllable number of instance m. The IIS of instance m is weighted sum of all different length replacements that are composed by instances m. here the weight is η_j (usually η_j is 1), which to keep balance of different instance lengths.

From the above, we can see that IIS is a measure of replacement ability (score of replacement) of each instance. In IIS, we take into account that how frequently each instance is selected by TTS. The frequency is demonstrated by usage of instance. If we consider the length (syllable number) of instance as granularity, in IIS, we also take into account for all kinds of different granular replacements.

3. Non-uniform adjustment. When a syllable instance is pruned, all those instances, which are composed by that syllable instance, are also pruned passively. Thus each instance should records the information of N replacement. If one-syllable instance, which compose the replacement R_K of R_0, is pruned, the replacement R_K is pruned passively. Then replacement R_L, with least L in the reserved replacements, is the final virtual non-uniform replacing R_0. This adjustment reduces loss of non-uniform: those R_0s with High IIS value are reserved, in this situation, $R_L=R_0$, there is no loss of non-uniform; If R_0 is pruned, R_0 is replaced by virtual non-uniform R_L. Because R_L is worse than R_0, there is a little loss of non-uniform; Only in the situation that all replacements are pruned, the non-uniform are thorough lost. So the value of N is very important: If N is too small, all replacements of a majority of instances are probably pruned. If N is too large, the compute will be time and storage consuming. In this paper, N=5.

4. Associated Scoring Elimination. There is a problem when we score each instance: If two instances can be replaced each other, maybe they will be scored twice. For example, if V_1 and V_2 can be replaced each other, V_2 is scored when scoring all replacements of V_1 and V_1 is scored when scoring all

replacements of V_2. Although only one of V_1 V_2 is need to reserve, V_1 and V_2 are probably reserve together because of repetitively scoring. This problem is named associated scoring. We propose a process as following to eliminate associate scoring; the process is named Associated-Scoring Elimination and ASE for short.

With each instance V there are two structures: V.REL composed by the replacement that can replace V, V.RIL composed by the replacement that V can replace. So IIS of V can express by:

$$S_V = \sum_{R \in V.RIL} score_R$$

We arrange the instances from high to low according to their IIS value. Suppose the arrangement is V_1, V_2 … V_H, there are only k (k<H) of them can be reserved according to reserve rate. Firstly, reserve instances V_1 because of its highest IIS value; $\forall$ R_x□V_1.REL , R_x□V_1.REL⇔V_1□R_x.RIL, thus remove V_1 from R_x.RIL. Secondly, adjust IIS value of R_x (R_x=V_2, …V_H):

$$S_{R_x} = S_{R_X} - score_{V_1}$$

At last, rearrange the left instances V_2, …V_H, reserve the instance with highest IIS value, and so on.

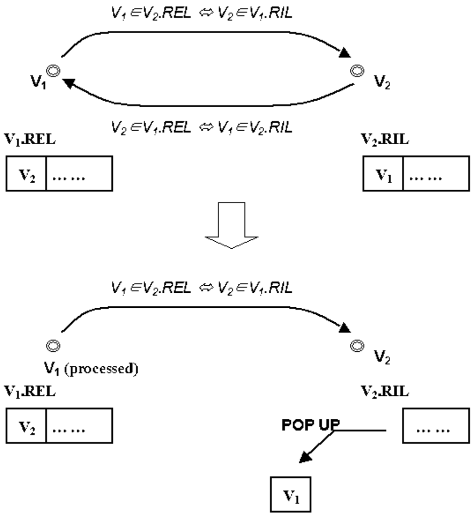

$V_1 \in V_2$.REL ⇔ $V_2 \in V_1$.RIL
V_1
V_2
$V_2 \in V_1$.REL ⇔ $V_1 \in V_2$.RIL
V_1.REL
V_2.RIL
V_2 ……
V_1 ……
$V_1 \in V_2$.REL ⇔ $V_2 \in V_1$.RIL
V_1 (processed)
V_2
V_1.REL
V_2.RIL
V_2 ……
POP UP
……
V_1

Figure 2: Associated-Scoring Elimination

If V_2 and V_1 can be replaced each other, there are two conclusion: (a) V_1□V_2.REL⇔V_2□V_1..RIL and (b) V_2□V_1.REL⇔V_1□V_2..RIL. But ASE remove V_1 from V_2..RIL, this make conclusion (b) not true. Thus V_1 and V_2 are not scored repetitively, associate scoring is eliminated. These are shown in figure 2. By using ASE, Pruning synthesis instances is deduced to a problem of graph theory: Construct an edge-weighted directed graph. Then begin with the vertex of maximum outgoing degree, search the vertices with not only maximum degree of outgoing but also minimum degree of incoming [14].

III. STARP-VPA ALGORITHM

This section describes **Sta**tistics & **Rep**lacing based **V**ariant **P**runing **A**lgorithm (StaRp-VPA for short). There are three main steps of StaRp-VPA:

Step1: Computing the instances reserve rate of each unit

Input: overall reserve rate of speech database

Output: instance reserve rate of each unit U, namely Reserve_rate(U)

The compute of step1 is just according from equation (2.1) to equation (2.3).

Step2: IIS scoring of instances

Input: every Reserve_rate(U), and other frequency information of instances

Output: information on reserved instances, replaced instances and pruned instances

The process of step2 is described as following:

```
For L=max length To 1, execute (1) and (2):
(1) ∀ V, Length(V)=L, execute (1.1) and (1.2):
  (1.1) V.REL={$R_0$, $R_1$ … $R_N$}, scores=$F_0$, $F_1$ … $F_N$。
  (1.2) ∀ W∈ V.REL, W.RIL={V} ∪ W.RIL。
(2) ∀ U, execute from (2.1) to (2.2):
  (2.1) Reserve_Num=Number (U)×Reserve_rate(V)
  (2.2) ∀ V∈ U, execute (2.2.1):
    (2.2.1) For i=1 to Reserve_Num,
              Execute ASE: Associated-Scoring Elimination
    (2.2.2) Tailor all Variants left by ASE
```

Note: L is syllable number of instance; its maximum is max length. V=Variant represents instances; U=Unit represents units. Number(U) is number of instances that unit U includes.

Step3: Adjusting speech database

Input: information on reserved instances, replaced instances and pruned instances and original speech database

Output:: speech database after pruning

(1) To reserved instance, nothing needs to do;

(2) To replaced instance, replace the instance with its virtual non-uniform replacement;

(3) To pruned instance, Delete all information connect with it from speech database

IV. OBJECT AND SUBJECT EVALUATION

Using StaRp-VPA, we get several speech databases of KBCE with different reserve-rates. In this section, some practical results of StaRp-VPA are evaluated in the following.

A. Objective evaluation

Because the purpose of StaRp-VPA is to make up the loss of non-uniform, it is natural to evaluate the results pruned by StaRp-VPA with the distribution of non-uniforms (including syllables) after prune. The objective measurements here are: proportion between number of reserved non-uniforms and number of original non-uniforms (rONU), proportion between number of virtual non-uniforms and number of original non-uniforms (rVNU), proportion between number of pruned non-uniforms and number of original non-uniforms (rTNU), and the λ_O, λ_{OV} (that will be illuminated late in this section). All the distributions of non-uniforms under different reserve rates are showed in Figure 3.

β (%)	*rONU* (%)	*rVNU* (%)	*rTNU* (%)	λ_O	λ_{OV}
73	62.53	36.43	1.24	0.86	1.36
61.9	47.80	48.85	3.35	0.77	1.56
50	33.98	56.44	9.58	0.68	1.81
30	15.93	48.22	35.85	0.53	2.14
10	4.36	19.15	76.49	0.43	2.35

Figure 3: the distributions of non-uniforms

Here β is reserve rate; so pruning rate of speech database is 1-β,

$$\lambda_O = \frac{rONU}{\beta}, \quad \lambda_{OV} = \frac{rONU + rVNU}{\beta}$$

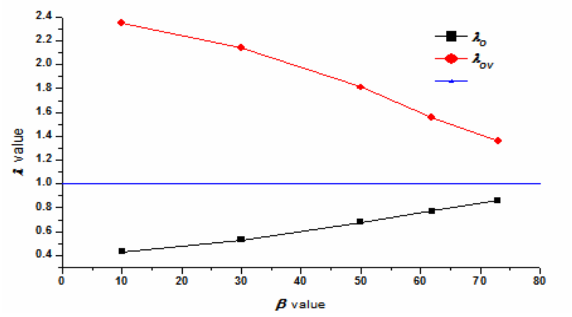
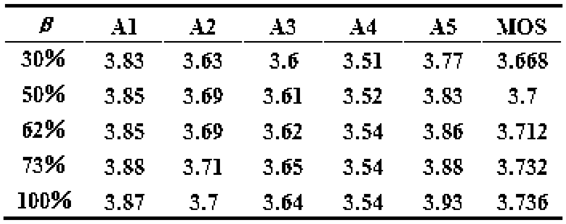


Figure 4: λ_O and λ_{OV} with different β

λ_O, λ_{OV} describe the effect of changing reserve rate on the loss of non-uniforms. Generally speaking, there is certainly loss of non-uniform when pruning speech database. In fact, the theoretically optimality β=*rONU* is impossible. The reason is explained as following: Connecting with a syllable instance V, there are L_V instances of different lengths composed by V. When V is pruned, those L_V original instances are still pruned. Different V means different L_V. If a instance V with large L_V is pruned, *rONU*<<β。 Figure 3 and 4 show that λ_{OV} descends slowly when β descends. This demonstrates StaRp-VPA tends to reserve those syllable instances with large L_V. The loss of non-uniform is in some degree made up with virtual non-uniform, as λ_{OV} shows in figure 3 and 4 (usually, λ_{OV} >1)。

B. *Subjective evaluation*

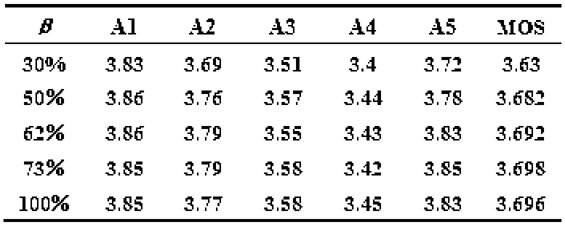
We use texts of two kinds performing listening test to evaluate the effect of StaRp-VPA on synthesis quality. The subjective measurement is MOS (Mean Opinion Score).



β	A1	A2	A3	A4	A5	MOS
30%	3.83	3.63	3.6	3.51	3.77	3.668
50%	3.85	3.69	3.61	3.52	3.83	3.7
62%	3.85	3.69	3.62	3.54	3.86	3.712
73%	3.88	3.71	3.65	3.54	3.88	3.732
100%	3.87	3.7	3.64	3.54	3.93	3.736

Figure 5: MOS of front 100 sentences in text 1

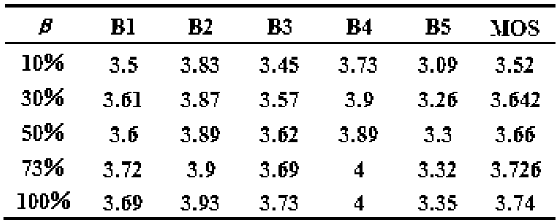
Text 1 includes 150 sentences, which is automatically gathered from a large corpus by using the approach of paper [3]. Front 100 sentenses of text 1 are of high coverage, while Rear 50 sentenses are of low coverage. Five formal listeners perform the listening test on speech database of KBCE with different reserve rate. The MOS of front 100 and rear 50 sentences are showed in figure 5 and 6. Figure 5 and 6 demonstrate that MOS degrades quite slowly when pruning rate rise. Even though the pruned speech database is 30% of origin, MOS just degrades within 0.07. Specially, the MOS of 73% is a little higher than that of origin for rear 50 sentences.



β	A1	A2	A3	A4	A5	MOS
30%	3.83	3.69	3.51	3.4	3.72	3.63
50%	3.86	3.76	3.57	3.44	3.78	3.682
62%	3.86	3.79	3.55	3.43	3.83	3.692
73%	3.85	3.79	3.58	3.42	3.85	3.698
100%	3.85	3.77	3.58	3.45	3.83	3.696

Figure 6: MOS of rear 50 sentences in text 1

Text 2 includes 100 sentences, which is automatically gathered from the Internet still by using the approach of paper [3]. Another five formal listeners perform the listening test on speech database of KBCE with different reserve rate. The MOS of text 2 are showed in figure 7. Figure 7 demonstrates that MOS of text 2 still degrades slowly but a little more quickly than text 1. We even pruned the speech database to 10% of origin with the MOS degrading only 0.22.



β	B1	B2	B3	B4	B5	MOS
10%	3.5	3.83	3.45	3.73	3.09	3.52
30%	3.61	3.87	3.57	3.9	3.26	3.642
50%	3.6	3.89	3.62	3.89	3.3	3.66
73%	3.72	3.9	3.69	4	3.32	3.726
100%	3.69	3.93	3.73	4	3.35	3.74

Figure 7: MOS of text 2

In Figure 8, the curve of MOS descends slowly when reserve rate descends. When reserve rate is above 50%, MOS is almost unchanged. When reserve rate is under 50%, the MOS doesn’t severely degrade.

V. DISCUSSION AND CONCLUSION

In two listening tests mentioned above, the MOS doesn’t degrade severely. The reason maybe: 1. Because of StaRp-VPA’s mechanism, the instances reserved are the instances which usually able to replace others and are frequently selected by TTS. 2. Utilizing virtual non-uniforms, in some sense, make up the loss of replaced non-uniforms. 3. Vibration-rate pruning reserve most prosody and phonetic environments, only unimportant instances are pruned.

In this paper, we proposed the concept of virtual non-uniform in order to make up the loss of non-uniform. And based on virtual non-uniform and the usage of instances, we constructed the algorithm StaRp-VPA and used StaRp-VPA pruning speech database of KBCE to different size. In listening tests, the naturalness scored by MOS remains almost unchanged when less than 50% instances are pruned off, and the MOS does not severely degrade when reduction rate is above 50%.

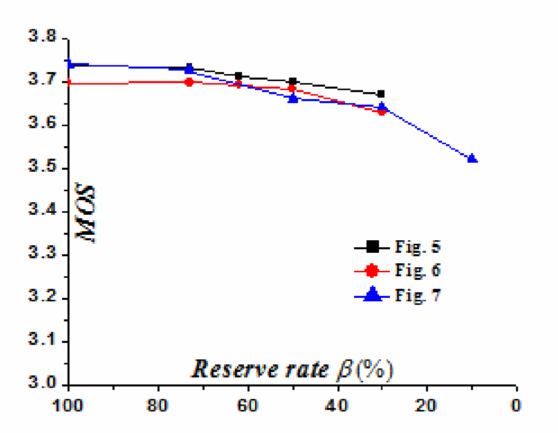


Figure 8: the change of MOS with different reserve rate

VI. Acknowledgment

The paper is supported by the project *Research on Scalable Mandarin TTS System* (No. 60602017) of National Science Foundation of China.

References

[1] Hunt, A., Black, A., Unit selection in a concatenative speech synthesis system using a large speech database, Proceedings of ICASSP1996, vol. 1, 373-376, 1996

[2] Sagisaka, Y, Kaiki, N., Iwahashi, N. and Mimura, K., ATR-v-TALK speech synthesis system, Proceedings of ICSLP1992, vol.1, 483-486, 1992

[3] Liu, Q. F., Speech synthesis study based on perception quantification, doctor thesis, university of science and technology of china, 2003

[4] Chu, M., Peng, H., Yang H., and Chang, E., Selection non-uniform units from a very large corpus for concatenative speech synthesizer, Proceedings of ICASSP2001, 2001

[5] Rabiner, L. R., A tutorial on hidden markov models and selected application in speech recognition, IEEE Proceedings, 77 No. 2, 257-285, 1989

[6] Breiman, L., Friedman J., Olsen, R. and Stone., C., Classification and regression trees, wadsworth & Brooks, Pacific grove, CA, 1984

[7] Black, A. W., Taylor, P. A., Automatically clustering similar units for units selection in speech synthesis, Proceedings of Eurospeech1997, vol.2, 601-604, 1997

[8] Hon, H., Acero, A., Huang, X., Liu, J. and Plumpe, M., Automatic generation of synthesis units for trainable text-to-speech systems, Proceedings of ICASSP1998, vol. 1, 293-296, 1998

[9] Kim, S. H., Lee, Y. L. and Hirose, K., Pruning of redundant synthesis instances based on weight vector quantization, Proceedings of Eurospeech2001, 2231-2234, 2001

[10] Kim, S. H., Lee, Y. L. and Hirose, K., Unit generation based on phrase break strength and pruing for corpus-based text-to-speech, ETRI Journal, vol. 23, No. 4, 168-176, Dec., 2001

[11] Rutten, P., Aylett, M., Fackrell, J. and Taylor, P., A statistically motivated database pruning technique for unit selection synthesis, Proceedings of ICSLP2002, 125-128, Denver, 2002

[12] Zhao, Y., Chu, M., Peng, H. and Chang Eric, Custom-tailoring TTS voice font-keeping the naturalness when reducing database size, Proceedings of Eurospeech2003, 2957-2960, 2003

[13] Ling, Z. H., Hu, Y., Shuang, Z. W. and Wang, R. H., Compression of speech database by feature separation and pattern clustering using STRAIGHT, Proceeding of ICSLP2004, 766-769, 2004

[14] Bondy, J A, Murty, U. S. R. Graph theory with application. American Elsevier, New York, 1976