A Miniature Chinese TTS System Based On Tailored Corpus

Zhi-Wei Shuang, Yu Hu, Zhen-Hua Ling, Ren-Hua Wang

iFly Speech Laboratory
University of Science and Technology, China
szw@mail.ustc.edu.cn

ABSTRACT

Miniature Text to Speech (TTS) systems are broadly applied to embedded system and speech chip, where limited resource requires the corpus to be relatively small and the computing complexity to be low. In general, speech synthesized by conventional miniature TTS systems lacks naturalness due to the limitation of corpus size. In this paper, a method of automatic building a small corpus from a large speech database is described. A new way of distance measurement among candidate instances is also proposed. Based on the tailored corpus, a miniature Chinese TTS system is built, which can produce speech with high naturalness.

1. INTRODUCTION

Recently, corpus based approaches have been successfully applied to TTS system. With this method, the quality of output speech has been greatly improved. However, the required speech database is too large to be used in a miniature TTS system. To get an optimal tradeoff between the capacity and the performance is essential. The aim of the corpus design is to have a decent coverage of phonological context based on the idea that phonetic features will stay the same under same phonological environment [5]. However, different phonological instances may have similar phonetic features, especially when some prosodic modification is allowed. So tailoring the corpus is possible.

In this paper, we propose an automatic method to build a small corpus from a large speech database. A new way of distance measurement among candidate instances is also proposed. Based on the distance measurement, a small quantity of instances can be selected automatically to substitute the original whole speech database as our corpus with least degradation to the performance of synthesized speech. This tailored corpus is fit for the tradeoff between speech quality and memory size.

With the small constructed corpus, a miniature TTS system is built, which can produce speech with high naturalness. Here we define a miniature TTS system as one that can be chosen to replace another instance with little degradation in perception. Prosodic features based clustering

2. DATABASE REDUCTION

Our original database is built by 2-hours’ female speech pronounced in a narrative mode and recorded in 16kHz’s sampling rate, with about 20 instances for each syllable. Segmentation and labeling has been finished. And a TTS system has been built successfully based on this corpus [2].

To guarantee the performance of the tailored corpus, we need to ensure the prosodic and phonetic coverage of the original speech database. Here, to cover an instance under sense of prosodic features is defined as one which can be picked out and modified to a different instance by signal processing with little degradation, while to cover an instance under sense of phonetic features is defined as one which can be chosen to replace another instance with little degradation in perception.

2.1. Prosodic features based clustering

Since prosodic mismatch is much more perceivable than that of phonetics, the coverage on prosodic features should be guaranteed first.

Since it is advantageous to allow prosodic modifications, if major discontinuities can be avoided during synthesis [5], then prosodic distance can be defined as the degradation caused by the prosodic modification from one instance to another, instead of the difference between their prosodic attributes.

TD-PSOLA is adopted to adjust the prosody during synthesis because it causes little degradation in speech quality when the modification is small –scaled and its low computing complexity is fit for miniature TTS systems.

2.1.1. Pitch data preprocessing

To facilitate our comparison of different instances’ pitch envelope, we need to represent them with a vector with equal dimension.

For those syllables who have more than one voiced phoneme, normalization should be carried out on them.

Figure 1. Diagram shows two instances of syllable ma4, lower contours are the pitch contours, the vertical lines mark the segmenting point of m, a
As shown in Fig 1, the length of phonemes of different instances may vary in a large scale. Simply sampling the pitch contour may cause compare of pitch between different phonemes, which is obviously meaningless.

The normalization of syllables with two voiced phonemes is illustrated as follows.

Suppose there are N instances for this syllable, note the $i$ th instance’s split point of two voiced phoneme as $F_{pos(i)}$, the $i$ th instance’s duration as $Duration(i)$. Then the relative position of the split point in the instance, noted as $fpos(i)$, can be calculated by:

$$fpos(i) = F_{pos(i)}/Duration(i)$$

And the average of $fpos(i)$ can be computed as

$$favg = \frac{1}{N} \sum_{i=1}^{N} fpos(i)$$

If the length of whole instance’s pitch vector is $N_i$, the first voiced phoneme ‘s pitch will be represented by a $favg \times N_i$’s vector, while the later will be represented by a $(1 - favg) \times N_i$’s vector.

Instances with more phonemes can be dealt with by similar methods.

### 2.1.2. Evaluation of prosodic distance

What we need to measure is the degradation in speech quality when one instance is converted to another by modification of pitch and duration. A model is built to map the difference in prosody to the degradation in speech quality.

By perceptual experiments in modifying the pitch and duration of instances separately with PSOLA, we have found the degradation, noted as $d(r)$, is related to the square of scale of modification mode, noted as $R$, which can be represented as:

$$d(r) = \begin{cases} a_r r^2, & r > 0 \\ a_r r^2, & r < 0 \end{cases}$$

$a_r$ indicate the degree of degradation with the same scale’s modification. The same scale’s modification of duration will cause much less degradation than modification of pitch. $a_r$, $a_\_\_$ are mainly determined by the type of the instance’s vowel part. Experiments were done to acquire different vowels’ $a_r$, $a_\_\_$. We choose 5 syllables for one vowel, modifications with scale $r$ from –0.4 to 0.4 and a step of 0.05 were made to every syllable’s pitch and duration separately. 10 experienced researcher were invited to evaluate the grade of quality of modified speech on a score scale of 1 (bad) to 5 (good), noted as $g$, which can be fit by:

$$g = 5 - d(r)$$

By the *method of least squares*, $a_r$, $a_\_\_$ can be solved for, which will be used as the $a_r$, $a_\_\_$ of the vowel.

Then we can calculate the degradation caused by modifying the $m$ th instance’s prosody to the $n$ th instance’s prosody:

$$d_{mn} = \frac{1}{N_i} \sum_{i=1}^{N_i} d_{i}(F_{mi}/F_{ni} - 1) + d_{i}(D_{mi}/D_{ni} - 1)$$

Where $F_{ni}$, $F_{mi}$ mean the $i$ th dimension’s value of $n$ th, $m$ th instances’ pitch vector, while $D_{ni}$, $D_{mi}$ mean the $n$ th, $m$ th instances’ duration.

Thus the distance in prosodic features between the $n$ th, $m$ th instance, noted as $D_{p_{nm}}$, can be calculated by:

$$D_{p_{nm}} = \frac{1}{2}(d_{mn} + d_{nm})$$

### 2.1.3. Prosodic distance based clustering

![Figure 2](a)

Figure 2. a: the pitch contours of all 20 instances for one special syllable b: the pitch contours of class center when all instances’ pitch contours are partitioned into 5 clusters

After being classified into 5 clusters, the average difference in one dimension between instances in one cluster is below 5Hz. A limited modification can be performed by PSOLA with so little degradation to the quality that every instance can be selected to substitute the whole cluster.

There is only one instance both in cluster 1 and cluster 5, which indicates that these two instances are quite different from the other 18 instances. We may assume these two instances as bad instances and discard them. So we only need to retain arbitrary one instance for each cluster 2,3,4 while acquiring 90%’s coverage of all the instances in prosodic features.

The algorithm is described below, where $D_{p_{nm}}$ will be used as the distance between the $n$ th and $m$ th. A prosodic
coverage threshold \( T_p \) and degradation threshold \( D_p \) can be set according to practical needs.

1. \( C_n = 1 \).
2. All \( N \) instances for one syllable are partitioned to \( C_n \) clusters by k-Means method \([7]\), based on the prosodic distance between samples.
3. Calculate the average distance \( D_{\text{Dis}} \) among every two instances in one cluster.
4. If \( D_{\text{Dis}} < D \) is true for every cluster, go to 5. Else \( C_n = C_n + 1 \), go to 2.
5. Sorting all the clusters according to clusters’ number of instances, noted as \( n[k] \), the cluster no. after sorting will be recorded as \( a[k], k = 1, 2, C_n \).
6. \( \text{Sum} = 0, k = 1 \)
7. \( \text{Sum} = \text{Sum} + n[a[k]] \)
8. If \( \text{Sum} > T_p \times N \) go to 9 else \( k = k + 1 \) go to 7.
9. The current \( k \) will be recorded as \( C_n \).

Then all the instances are partitioned into \( C_n \) clusters, and \( C_n \) is determined dynamically by the distribution of the instances. To acquire coverage of \( D_p \), those clusters that embody more instances will be retained first and those clusters that embody only 1 or 2 instances will be likely dropped out as bad instances.

\( 2 \sim 4 \) clusters will be retained for most syllable when \( T_p = 0.9 \), \( D_p = 0.8 \), which will be used for later work.

2.2 Phonetic distance based instance selection

As selecting a certain instance from every retained cluster could satisfy prosodic coverage, the strategy of choosing instances is made to satisfy the coverage on phonetic features. The measure of distance between instances on phonetic features will be discussed first.

Based on the idea that similar phonological context leads to similar phonetic features \([5]\), a rule-based distance table constructed by listening tests that relies on a set of context attributes is applied to determine the similarity between two Text instances, which has been used as the cost function for segment selection during synthesis in a former corpus-based TTS system \([10]\). The position of a syllable in the word, in the phrase and in the sentence, the previous syllable’s vowel part and the next syllable’s initial will be taken into account when calculating the rule based distance between \( n \) th, \( m \) th instance, noted as \( D_s(m, n) \).

Since the first and the last frame’s spectrum is important to the smooth concatenation with the former and the next syllables, the distance of the first and the last frame’s spectrum between the units should be considered to reach a high phonetic coverage. The Kullback-Leibler distance \([8]\) is used to calculate the distance of two spectrums, which are found related closely with the human perception. \([6]\) The symmetric version is also used, computed as:

\[
d_{\text{kl-sym}}(S, S') = \int_{-\pi}^{\pi} \log(\frac{S(w)}{S'(w)})[S(w) - S'(w)] \frac{dw}{2\pi}
\]

Thus the first and the last frames’ spectral distance between the \( n \) th instance and the \( m \) th instance can be computed as:

\[
D_s(m, n) = d_{\text{kl-sym}}(S_{nf}, S_{mf}) + d_{\text{kl-sym}}(S_{nf}, S_{mf})
\]

Where \( S_{nf}, S_{mf} \) are the first frames’ spectrums while \( S_{nf}, S_{mf} \) are the last frames’ spectrums of the \( n \) th, \( m \) th instances.

The distance between the \( n \) th instance and the \( m \) th instance is then computed as:

\[
D_s(m, n) = D_s(m, n) + \alpha \times D_s(m, n)
\]

\( \alpha \) is a weight factor to adjust the importance of \( D_s(m, n) \).

The instances in one cluster are partitioned again to several sub-clusters based on the phonetic distance between instances. The method is similar with the strategy of partitioning the instances based on prosodic distance, except that the k-Medoids method \([7]\) is used. And the medoids in the sub-clusters will be retained to construct our final inventory.

A phonetic distance threshold \( T_s \) and coverage threshold \( D_s \) can also be adjusted to balance the capacity and coverage on phonetic features. Since the instances with similar prosody are likely to have similar phonetic features, it’s reasonable to partition every cluster to 2-3 sub-clusters to reach a high phonetic coverage.

Setting \( D_s = 0.8 \), two final corpus are constructed with different \( T_s \).

1. With a relatively high \( T_s \), only 1-2 samples are retained for every subset, or 3-4 samples are retained for one syllable. Compressed by G.723.1 encoder in 8K’s sampling rate, the size of the lib is about 1Mbytes, noted as \( \text{Lib}_1 \).
2. With a relatively low \( T_s \), 2 or 3 samples are retained for every subset, or 4-8 samples are retained for one syllable. Compressed by G.723.1 encoder in 8K’s sampling rate, the size of the lib is about 1.5Mbytes, noted as \( \text{Lib}_2 \).

3. A PRACTICAL MINIATURE TTS SYSTEM

![Diagram](image-url)
Based on the tailored corpus constructed above, we build a miniature TTS system whose capacity is less than 2M Bytes. Fig3 shows the system block diagram.

$D_{pt}$, $D_{ph}$ will be used as prosodic distance and phonetic distance between targets and instances in corpus in unit selector.

10-experienced listeners are invited to evaluate the grade of quality of synthesized speech based on $Lib_{a}$ and $Lib_{b}$ on a scale of 1(bad) to 5(good). The synthesized speech of the original corpus-based TTS system [10] noted as $Lib_{a}$ and the original miniature TTS system [11] based on fixed acoustic corpus noted as $Lib_{b}$ are also evaluated for compare.

The result is shown in below:

![Figure 4: the grade of different system](image)

From Figure 4, we can find the performance of the miniature TTS system based on either $Lib_{a}$ or $Lib_{b}$ is close to that of the original corpus-based system and is much better than the original miniature system based on the fixed acoustic corpus. It shows that our strategy of instance reduction is feasible.

It is also clear that the performance of the system based on $Lib_{a}$ is better than that of the system based on $Lib_{b}$, though the difference is not obvious. A more careful analysis shows their performances are quite similar for many sentences, which indicates the importance of prosody in Chinese speech synthesis. However, some declination in performance of system based on $Lib_{a}$ may occur when the best candidate shows too much phonetic discrepancy from the targets. Therefore the performance of system based on $Lib_{b}$ is more robust.

4. CONCLUSION

The proposed strategy for tailoring corpus has been approved. Based on the tailored corpus the developed miniature Chinese TTS system achieved a comparable performance with the original corpus based system. And the tailored corpus is also highly scaleable. The tradeoff between speech quality and memory size can be acquired easily just with different threshold.

Both prosodic attributes and phonetic attributes are included in our method to acquire a decent prosodic and phonetic coverage. Also our prosodic and phonetic distance measurement has more close relation with human perception than that obtained through HMM’s.

The degradation caused by prosodic modification is strongly related to the algorithm used. And degradation caused by phonetic mismatch can also be decreased by the phonetic modification with better algorithms. Some new modification algorithms recently have been tested, and the essential problem is to reduce the computing complexity to let them be fit for miniature systems.

5. ACKNOWLEDGEMENTS

This research is supported by the National Natural Science Foundation of China (69975018)

6. REFERENCES


[7] Jiawei Han, Micheline Kamber, “Data mining concepts and techniques”, Mogan Kaufmann Publishers, 2001


