Custom-Tailoring TTS Voice Font
— Keeping the Naturalness When Reducing Database Size

Yong Zhao, Min Chu, Hu Peng and Eric Chang

Microsoft Research Asia, Beijing, 100080
{yzhao;minchu;hupeng;echang}@microsoft.com

Abstract

This paper presents a framework for custom-tailoring voice font in data-driven TTS systems. Three criteria for unit pruning, the prosodic outlier criterion, the importance criterion and the combination of the two, are proposed. The performance of voice fonts in different sizes which are pruned with the three criteria is evaluated by simulating speech synthesis over large amount of texts and estimating the naturalness with an objective measure at the same time. The result shows that the combined criterion performs the best among the three. The pre-estimated curve for naturalness vs. database size might be used as a reference for custom-tailoring voice font. The naturalness remains almost unchanged when 50% of instances are pruned off with the combined criterion.

1. Introduction

Most state-of-the-art text-to-speech (TTS) systems adopt corpus-driven approaches due to their capability to generate highly natural speech. In these systems the most suitable speech segments are selected from a very large unit inventory first, and then are concatenated to generate the output speech. With this type of systems, speech synthesis becomes a problem of collecting, annotating, indexing and retrieving from a large speech database [1] [3] [6]. The naturalness of synthesized speech to some extent depends on the size and the coverage of the unit inventory, where many instances with phonetic and prosodic variations are kept for target units. Generally, several hours of speech waveforms are required for synthesizing natural-sound speech from diverse text input. Thus, storing, loading and searching such a huge corpus become inevitable issues in many applications.

Several approaches for reducing the size of voice font have been proposed. The approach described in [2] clusters similar units 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 [4] 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 [5]. 50% reduction rate is reached without significant distortions. All above methods tend to keep the most likely units, i.e. units that are close to the centroid, normally measured in segmental features of a cluster.

This paper proposes a new framework for custom-tailoring voice font in our two-module TTS system [6], which does not have a prosody model to predict numerical target values for prosodic features, and consequently does not apply pitch or time scaling on the selected units. Both prosodic outliers and non-frequently used instances are pruned with the objective of minimizing decrease in voice quality.

The paper is organized as follows: problems in the two-module TTS framework are investigated in Section 2. The new custom-tailoring framework is described in Section 3. The final discussion is presented in Section 4.

2. Problem investigation

In our Mulan Mandarin TTS system [6], a very large speech corpus is used. The corpus covers approximately 64% of all possible unit variations that are distinguished by their syllable-dependent descriptive contextual variation vector [6]. Similar to what had been done in [1], automatic classification and regression tree (CART) is used to index all instances of each unit in the corpus. However, instead of using a very complicated acoustic vector as the impurity measure, we choose a much simpler vector that contains only prosody related features, i.e. the average f_0, the dynamic range of f_0 and the duration of an instance. The question set covers only prosody related features such as the position in phrase, position in word, left tone and right tone. The splitting criterion for CART is to maximize the reduction in the weighted sum of the MSEs (Mean Squared Error) of the three prosodic features. The MSE of each feature is defined as the mean of the squared distances from the feature values of all instances to the mean value of their host leaves. After the trees are grown, all instances on the same leaf node are assumed to possess similar prosodic features and the main difference among them is inside segmental features. Two phonetic constraints, the left and right phonetic contexts, and a smoothness cost, are used in unit selection to assure the continuity of the concatenation between units. One of the most important advantages of this framework is that the synthetic utterances could achieve richer intonation by inheriting the prosodic habit of the original voice talent.

However, strange prosodic salience arises occasionally. A main cause for the instability is identified as misusage of prosodic outliers in the leaf nodes of CARTs. Although several prosodic contexts of instances are queried when growing a CART, there are still other prosody related factors, such as stress or emphasis, which are not considered. Thus, some instances on a leaf node may have prosodic features far away from the center of the cluster and will cause unnatural prosody if they are put in an improper context. Furthermore, some instances have strange pitch or duration because of mistakes in unit boundary alignment or break-indices labeling. Figure 1 shows the scatter diagram of the instances on one leaf node of the tree for the syllable “shi4”. The horizontal axis is average f_0 and the vertical axis is the duration. Obviously, there are some outliers out of the circle around the center of the node. These outliers possess higher potential than others to generate an unnatural prosodic salience. Thus, they should be pruned off. A two-dimension graph is drawn for ease of understanding. The situation is similar when three or more features are considered.

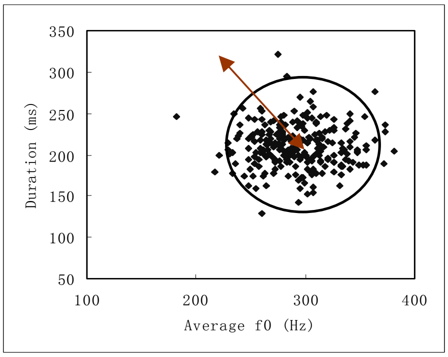

Figure 1: Scatter diagram of instances on one leaf node of the tree for the syllable “shi4”.

Another issue to be addressed is the redundancy in the unit inventory. Since the speech corpus is designed to cover as many prosodic variations as possible, it inevitably contains a good deal of “repeated” instances, i.e. both the prosodic context and phonetic context of these instances are the same, or alike. Meanwhile, the unit selection algorithm is incapable of differentiating them and picks one at random. In this case, only one of these repeated instances is frequently chosen for synthesizing speech, and others are unused yet occupy storage space.

The idea of identifying redundant instances might be generalized as that the less frequently used instances are less important than the frequently used ones. The importance of an instance can be measured by its contribution to synthetic speech, defined as the usage frequency of the instance divided by the accumulative usage frequency of all instances after synthesizing a large amount of texts. Figure 2 shows the distribution of the contributions of instances in the original unit inventory. Instances are grouped by their contributions on the horizontal axis. The accumulative contribution of a set denotes the sum of instance contributions in this set. It is noticed that the contributions of instance are uneven. On one hand, about 18.2% of instances in the unit inventory have never been used (i.e. their contribution is 0). Instances with contribution less than 10^{-6} account for 52.6% of all instances and their accumulative contribution is only 2.2%. On the other hand, instances with contribution more than 10^{-5} account for 11.6% of all instances and their accumulative contribution reaches up to 69.0%. Therefore, instances with the least contribution should be pruned first.

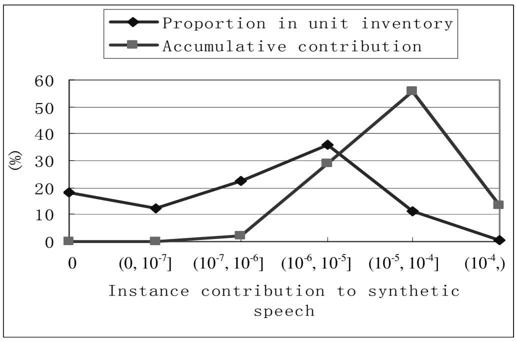

Figure 2: Distribution of the contributions of instances in the original unit inventory (calculated by synthesizing speech from a 200MB text corpus).

3. Custom-tailoring voice font

The naturalness of synthesized speech will not be decreased if only prosodic outliers and idle units are pruned off. However, if we go further, i.e. prune off more units, it is desired to have a clear view on potential decrease in voice quality when reducing the size. This paper proposes a framework for custom-tailoring voice font so that a balance point between quality and database size can be decided according to the requirement of practical applications.

3.1 An objective measure for naturalness

When pruning units, it is always desirable to know whether it will decrease the quality of synthesized speech and how much the decrease will be. So, an objective measure for speech quality is required.

In our previous study, a formal subjective evaluation was done to investigate the relationship between the naturalness of synthetic speech and some objective measures [7]. ACC (Average Concatenative Cost) shows pretty high correlation with MOS (Mean Opinion Score), as shown in Figure 3. The high correlation reveals that the ACC predicts, to a great extent, the perceptual behavior of human beings. Thus, the linear regression equation at the right top corner of Figure 3 can be used to estimate MOS from ACC. More details on the definition of ACC and the subjective evaluation can be found in [7].

Once some outliers or less important synthesis instances are pruned off from the original unit inventory, the remaining ones form a new voice font, the performance of which can be re-estimated by synthesizing a large amount of text scripts with it and calculating the overall ACC at the same time. The overall MOS of speech synthesized with the new voice font can then be estimated by the equation in Figure 3. The best balance point is identified in two ways. In one way, when a size limitation has been decided by the application, units are pruned to achieve the minimum increase in ACC or the minimum decrease in the estimated MOS. In the other way, when a quality limitation (i.e. target ACC) is requested by the application, units are pruned to achieve the minimum size of voice font. Three pruning criteria are proposed.

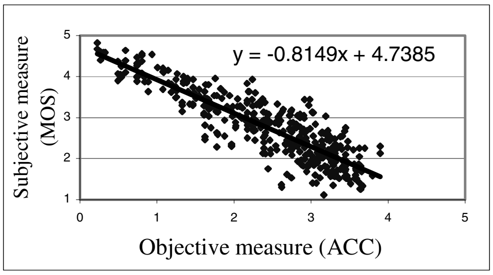

Figure 3: Correlation between ACC and MOS.

3.2 The prosodic outlier criterion

As illustrated in Figure 1, instances locating outside the circle are treated as prosodic outliers. Different choices of the circle radius result in voice fonts with different sizes. When the radius is set to be very large, all instances are included in the circle, as a result, no instances are pruned off. When the radius is shrunk step by step, more and more instances move outside and are pruned off. If the radius becomes small enough, only a few instances closest to the cluster center are preserved. These prosodic neutral instances are believed to be better prosodic representatives of the cluster than the others. The drawback of merely using prosodic neutral instances is that the concatenated speech sounds flattened and monotonous in prosody. Furthermore, when more and more instances from different phonetic contexts are pruned, the chance of producing unsmooth concatenation at the unit boundary will increase. Consequently, overall naturalness will be destroyed. Therefore, the radius of the circle should be controlled carefully.

To make it easily understood, the relative proportion of pruned instances (i.e., the number of instances to be pruned off over the number of all instances), which can be derived from a given radius, is used as a control parameter for illustration.

3.3 The instance importance criterion

As illustrated in Section 2, since the importance of an instance is measured by its contribution to synthetic speech, instances whose contributions are smaller than a preset threshold are pruned. The larger the threshold is, the smaller the voice font will be. However, when a large threshold is used, the naturalness of synthesized speech will drop. Again, for easy illustration, the relative proportion of pruned instances that is derived from a given threshold is used as a control parameter.

3.4 The combined criterion

The two criteria discussed above are complementary. Pruning units with the outlier criterion tends to preserve prosodic neutral units and the scheme adopting the importance criterion tends to keep the frequently used ones. They are integrated in order to obtain better performance, i.e., prune the prosodic outliers first and then prune the less important units. Various settings of the radius in outlier criterion and the threshold in importance criterion result in possible voice fonts with different sizes. Still, the relative proportion of the pruned instances is used as a control parameter. The performance of each possible voice font is measured with ACC, from which MOS is estimated.

3.5 Balance between naturalness and database size

Various settings for the radius and the threshold have been investigated by synthesizing speech from a 200MByte text corpus with the correspondingly generated voice fonts. The ACC over the corpus for a voice font is calculated at the same time and then MOS for the voice font is estimated. Most texts in the corpus are selected from newspapers, and others from novels, essays and weather reports. Four curves are plotted in Figure 4 to illustrate the relationship between the proportion of pruned instances and the naturalness for the remaining voice font. Curve D is obtained by applying the outlier criterion only. The estimated MOS drops rapidly with the increase of the pruned instance proportion. This is in accordance with the concerns mentioned in Section 3.2, i.e. the outlier criterion is not suitable for pruning large number of instances. Curve A is obtained by applying the importance criterion only. It demonstrates a rather flat portion when less than 50% instances are pruned. This is because that the pruned instances contribute only 2.2% to the synthetic speech.

The estimated MOS of any voice font generated with the combined criterion is anticipated to locate between curve A and D. This does not mean the naturalness is not improved. Since the unnatural prosody salience caused by outliers is difficult to be measured by ACC, the improvement in naturalness brought in by pruning some prosodic outliers might not be reflected in the estimated MOS. Curve B and C are generated by two different settings in radiuses in outlier criterion and a moving threshold in importance criterion, respectively. Curve C adopts a radius which makes the curve lower than A. From this value up, the radius is extended gradually to a value which results in curve B which almost overlaps with curve A. It is reasonable to believe that the naturalness of voice fonts on curve B is higher than that of the corresponding same-sized voice fonts on curve A owing to eliminating prosodic outliers.

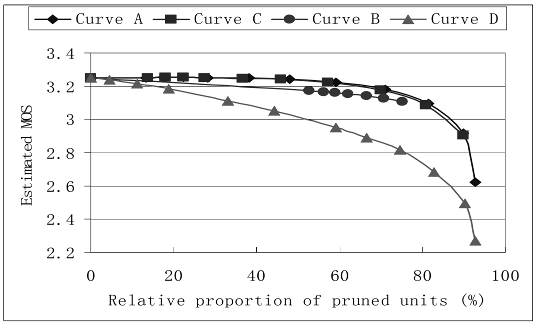

Figure 4: Relative proportion of pruned instances vs. estimated MOS of new voice font.

To verify the assumption, a listening test is done to compare utterances synthesized with two voice fonts on the two curves. One is the 22.3% off voice font on curve A and the other is the 22.3% off voice font on curve B. 30 sentences are selected as testing set so that utterances synthesized with the two voice fonts have some differences. The 30 pairs of synthetic utterances are played randomly to 18 subjects who are asked to select the more natural one from each pair. The preferring rate for the voice font on curve A and B is 44% to 56%, i.e. utterances generated with the voice font on curve B sound a little better than those synthesized with the corresponding one on curve A.

Therefore, curve B is chosen as a reference curve for custom-tailoring voice font. According to the curve, users may obtain a rough estimation on the potential quality decrease when certain proportion of instances are pruned off, or the size of voice font if a certain level of naturalness is requested to be maintained.

4 Discussions

This paper presents a framework for custom-tailoring voice font in a corpus-driven TTS system, in which selected units are concatenated directly without any pitch or time scaling. Unlike conventional unit pruning schemes which often tend to keep segments closest to the center of a cluster measured in segmental distance, three pruning criteria are proposed in this paper, including the prosodic outlier criterion, in which prosodic distance from the center of a cluster is measured, the importance criterion, in which the importance of instances are measured by their usage frequencies, and the combined criterion. It shows that, with elaborately adjusted parameters, the last criterion generates the best results. A reference curve revealing the relationship between naturalness and database size is obtained by simulating speech synthesis over a large text corpus. From the curve, it is observed that the naturalness of synthetic speech will be almost unchanged when 50% of instances are pruned off with the combined criterion, as is consistent with those reported in [2] and [5]. Furthermore, when about 80% of instances are pruned off, the estimated MOS is still above 3.0, i.e. the naturalness is acceptable. Beyond that point, the naturalness will drop rapidly. The curve is helpful for users to find a balance point between quality and database size according to requirements of practical applications.

5 References

[1] Hunt, A. and Black, A., “Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database”, *Proceedings of ICASSP1996*, vol. 1, pp. 373-376, 1996.

[2] Black, A. W. and Taylor, P. A., “Automatically Clustering Similar Units for Units Selection in Speech Synthesis”, *Proceedings of Eurospeech1997*, vol. 2, pp. 601-604, 1997.

[3] Huang, X., Acero, A., Adcock, J., “WHISTLER: A Trainable Text-to-Speech System”, *Proceedings of ICSLP1996*, Philadelphia, vol. 4, pp. 2387-2390, 1996.

[4] 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, pp. 293-296, 1998.

[5] Kim, S. H., Lee, Y. J., and Hirose, K., “Pruning of Redundant Synthesis Instances Based on Weighted Vector Quantization”, *Proceedings of Eurospeech2001*, pp. 2231-2234, 2001.

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

[7] Chu, M. and Peng, H., “An objective measure for estimating MOS of synthesized speech”, *Proceedings of Eurospeech2001*, 2001.

[8] Chu, M, Peng, H., Zhao, Y., Niu, Z and Chang, E., “Microsoft Mulan — a bilingual TTS systems”, *Proceedings of ICASSP2003*, Hong Kong, 2003.