Applying Pitch Connection Control in Mandarin Speech Synthesis

Yi Zhou+, Yi-qing Zu++, Zhen-li Yu++, Dong-jian Yue++, Gui-lin Chen++

+Shanghai Jiaotong University, P.R. China
++Motorola China Research Center, Shanghai, P.R. China

Abstract

In this paper, a novel tone-based pitch connection control in unit selection is described to improve naturalness of output speech for Mandarin text-to-speech (TTS) baseline system. This study mainly focuses on pitch connections of concatenative syllables. To improve the concatenation quality, we apply offset pitch of preceding syllable and onset pitch of following syllable in unit selection. According to the statistical result on corpus, three types of pitch connection constraints are proposed. Based on the property of pitch connection constraint, corresponding tone-based cost functions play important role in unit selection for continuity improving at concatenation point. By applying the defined cost functions in unit selection, more suitable units are selected and more natural-sounding synthesized speech is achieved.

1. Introduction

With the fast growing of computational power and memory capacity, more and more systems use concatenative text-to-speech (TTS) synthesis method[1-2]. By applying concatenative synthesis, intelligibility of synthesized speech is not among the big challenge, while high naturalness is now becoming the main concern. To achieve a natural-sounding speech synthesis, an important trend for most TTS research groups is to collect a large and prosody-rich database for speech synthesis. Then, how to select suitable units from a rich corpus becomes an interesting issue.

This paper proposes a novel approach to improve the speech unit selection. It gives constraints on unit selection module. Differing from [1,3], which focused on modifying selected units for reducing the discontinuities, in this paper, we focus on how to select more suitable units.

Motivation of our approach comes from observation of unnatural-sounding output. For the unnatural-sounding cases of synthesized speech, syllables’ mean pitch, duration and energy all “seem” suitable for concatenation. After analyzing the selected syllables, we found that the selected syllables have obvious unsuitable pitch jumps around concatenation points, though corresponding syllable’s mean pitch is suitable for concatenation. Original purpose of this paper is to remove or decrease this kind of pitch jumps. Compared with spectrum modification method, finding out more suitable units from rich corpus may be a better choice for achieving high quality of synthesized speech. For the natural speech read in intermediate speed, most of syllables have “clear” waveform boundaries or spacings. For syllable is tone carrier in Chinese, conventional pitch control method can hardly be used for reducing unsuitable pitch jumps in our Mandarin TTS system.

Literature [4] presented an approach to select target pitch contours for parameter-based speech synthesis. It classified all tone-based pitch contours into several representative contours (codebook). Based on tone concatenation rules, pitch contour representatives are selected for synthesis. In this approach, if the representatives are good enough, pitch jump between two successive syllables can be controlled through selecting proper pitch contour representatives. While for a real concatenative TTS system, pitch contour can hardly keep the same between target and selected units. If no extra pitch control processing is included, pitch jump problem cannot be solved automatically. Both [5] and [6] proposed pitch contour prediction based on statistical method. Focusing of these two literatures is the prediction and selection of pitch contour, in which no extra pitch connection properties between neighboring concatenated syllables were involved. [5] and [6] just provided a solution of pitch prediction, and how to solve neighbor syllables’ unsuitable pitch jump was not included.

In this paper, a tone-based pitch connection control method for speech unit selection is proposed. Neighbor syllables’ pitch contour connection property is used for statistics. Based on the statistical result, pitch connection constraints are proposed. By applying the proposed constraints in unit selection, a better pitch contour control for speech synthesis is achieved.

2. Pitch connection statistics

For Mandarin, natural speech can be classified into three parts, which are voiced part, unvoiced part and silence part. It is well known that pitch exists only in the voiced part. In this paper, for the simplicity and without ambiguity, pitch contour in the voiced part will be described as syllable’s pitch contour. For instance, onset pitch indicates onset pitch of syllable’s voiced part.

Statistics on neighbor syllables’ pitch contour connection is based on pitch “jump” between successive syllables, which is called pitch difference in this paper and denoted as $\Delta P_c$ (see in Fig. 1). Definition of pitch difference $\Delta P_c$ is

$$\Delta P_c = f_{z,\text{onset}} - f_{l,\text{offset}}$$

where $f_{z,\text{onset}}$ represents onset pitch of syllable 2 and $f_{l,\text{offset}}$ is offset pitch value of syllable 1 respectively. Syllable 1 and syllable 2 are neighbor syllables in an utterance, and syllable 1 is followed by syllable 2. Obviously, some of neighbor syllables’ pitch contour may be connected with no pitch difference between $f_{z,\text{onset}}$ and $f_{l,\text{offset}}$ for there is no unvoiced part in syllable 2. For this case, $\Delta P_c = 0$.

Syllables are select as basic speech units for our Mandarin concatenation speech synthesis. All syllables are associated with four lexical tones (denoted as tone 1 ~ 4) and one neutral tone (denoted as tone 5). Therefore, statistics on syllable’s pitch connection is tone-based, which can be briefly described as following. Suppose we are investigating the pitch connection property of syllable tone j followed by syllable tone k, which is simplified to tone pair {tone j , tone k}, all {tone j , tone k} pairs’ pitch connection results are sorted into same tone pair group. For a 5-tone (tone 1~5) Mandarin TTS system, there are 25 different pitch connection tone pairs. For the simplicity, we use the token (j-k) to denote tone pair {tone j , tone k}, where j , k = 1, 2, 3, 4, 5. In order to keep match
between corpus and synthesized speech, only the corpus used for synthesis is considered in statistics. Namely, speech in corpus is considered to be the most natural sounding. As we know, in an utterance, pitch at break point may be more flexible and may have a big pitch dynamic range. Pitch property at break point is quite different from non-break point. Accordingly, our statistics just focuses on the non-break points of utterance.

Pitch connection statistics is based on tone-based $\Delta P_c$ distribution. Discrete value is adopted to represent the distribution. Methodology of getting $\Delta P_c$ discrete coverage distribution is described as follows.

A sufficiently wide range of pitch difference (in Hz), for instance, $[0, 240]$ Hz (this range covers all pitch differences in statistics) is selected. This range is further partitioned into bins with a width of 5 Hz, such as $\{0\}$, $\{0, 5\}$, $\{5, 10\}$, $\{10, 15\}$,..., $\{235, 240\}$. $\{0\}$ is a special bin which has only one discrete value 0. The tone-based pitch difference is calculated throughout corpus. Each absolute value of $P_{c\Delta}$, denoted as $P_{c\Delta}$, is used for generating statistics. If a $P_{c\Delta}$ belongs to a bin $(f_a, f_b)$ (suppose the bin $(5, 10)$ is considered, $f_a$ refers 5, and $f_b$ refers 10), an appearance count will be added to this bin. Once all $P_{c\Delta}$ have been counted, distribution (count) of $|\Delta P_c|$ along with different small frequency bins is achieved. Based on the tone-based $|\Delta P_c|$ distribution, discrete coverage distribution $I(x)$ is calculated based on summations as shown in equation (2).

$$I(x)_{j,k} = \frac{\sum_{j=0}^{x} \text{Counts of } |\Delta P_c|_{j,k} \text{ in } b_{j,k}}{\sum_{j=0}^{x} \text{Counts of } |\Delta P_c|_{j,k} \text{ in } b_{j,k}} \times 100\%$$

where, $b_{j,k}$ denotes frequency bin $(f_a, f_b)$. Concerning Eq.(2), discrete coverage distribution $I(x)$ changes along with different $x$. For the simplicity, parameter $x$ in function $I(x)$ is called as pitch connection threshold which means that within the threshold $x$, coverage of pitch connection in frequency bin $[0,x]$ is $I(x)$. Fig.2, Fig.3 and Fig.4 give the coverage distribution of pitch connection for tone pair (1-1), (1-2) and (1-5) respectively.

Commonly, histogram (as shown in Fig.5) of pitch difference may be used in rule generating. While, our original try showed that coverage distribution $I(x)$ is better.

3. Pitch connection control

By drawing all coverage distributions in form of curve, such
as Fig. 2 ~ 4 for instance, we find that some of curves are quite similar in shape, while some are quite different. Some of pitch connection pairs’ coverage increases ‘fast’ with the increasing of pitch connection threshold x (for example, tone pair (1-1), as shown in Fig.2). While for some other pairs, the coverage increases even ‘slow’ (such as tone pair (1-2) shown in Fig.3), and there are still several pairs whose coverage changes with a ‘middle’ speed (as shown in Fig.4, tone pair (1-5), for instance).

To cluster these 25 coverage curves, a hybrid method is employed. Here, hybrid means that both automatic stage and manual stage are included. Automatic stage applies K-mean clustering method. For applying the same vector size, x (as shown in Equation (2)) is normalized to [0,1] and each coverage curve is equally divided into 50 points. Based on K-mean clustering result, manual-checking stage will be performed on the clustered items. Brief manual clustering benchmarks are as follows.

1. At least a certain number (3 in this paper) of candidates (curves) should be survived for each cluster. Distance between two ‘legal’ clusters should be more than a threshold, which is determined empirically.

2. For most clusters (not all clusters), distance between each candidate (curve) in same cluster is the smaller the better.

In our system, three classes of pitch connection constraints are assigned to 25 tone pairs. For the simplicity, each class is denoted with a name. The names are Less Than Constraint (LTC), More Than Constraint (MTC) and None Constraint (NC) respectively.

In the LTC class, most of pitch connection difference (\(\Delta P_c\), see in Fig.1) is under a certain threshold x, meanwhile, the threshold should be a relative small value. We call this threshold as ‘benchmark threshold’ for the corresponding tone connection pair. Determination of benchmark threshold is based on corresponding tone pair’s coverage distribution empirically. Tone pair (1-1), as shown in Fig.2, is a member of this class, and 40 Hz is selected as the benchmark threshold for this tone pair.

In the MTC class, most of pitch connection difference \(\Delta P_c\) is bigger than a threshold, meanwhile the threshold should be a relative small value. Tone pair (1-2), as shown in Fig.3, belongs to this class.

All pitch connection pairs that follow neither LTC nor MRC are classified as the third class, which is named as None Constraint (NC) class. Here none constrain does not mean ‘full free’ of pitch connection. In fact, exact meaning of this class is that the rule of pitch connection for these tone pairs hasn’t the property of LTC or MTC class, and it may be controlled through other type of pitch connection constraint.

Tone pair (1-5), as shown in Fig.4, is an example of this class.

As a cost function based TTS system, all three classes of constraints are employed. Each pitch connection constraint is implemented through a defined cost function. For LTC and MTC class, cost functions are defined by equation (3) and (4), respectively. Different from LTC and MTC, cost function of NC class is defined by equation (5), which follows the concept that the best candidate is the one that has the same (or the most similar) pitch connection difference comparing with the predicted target pitch.

\[
C_{NC} = \begin{cases} 
0.5 \cdot (\Delta P_c / \text{Threshold})^4, & |\Delta P_c / \text{Threshold}| < 1 \\
|\Delta P_c / \text{Threshold}|^4, & |\Delta P_c / \text{Threshold}| \geq 1 \\
\exp((2-2\cdot|\Delta P_c / \text{Threshold}|)/2), & 1 \leq |\Delta P_c / \text{Threshold}| < 2 \\
0.5 \cdot |\Delta P_c / \text{Threshold} - 2|^2, & 2 \leq |\Delta P_c / \text{Threshold}|
\end{cases}
\]

\[
C_{NC} = \begin{cases} 
0.5 \cdot (\Delta P_c / \text{Threshold})^4, & |\Delta P_c / \text{Threshold}| < 1 \\
|\Delta P_c / \text{Threshold}|^4, & |\Delta P_c / \text{Threshold}| \geq 1 \\
\exp((2-2\cdot|\Delta P_c / \text{Threshold}|)/2), & 1 \leq |\Delta P_c / \text{Threshold}| < 2 \\
0.5 \cdot |\Delta P_c / \text{Threshold} - 2|^2, & 2 \leq |\Delta P_c / \text{Threshold}|
\end{cases}
\]

\[
C_{NC} = \begin{cases} 
0.5 \cdot (\Delta P_c / \text{Threshold})^4, & |\Delta P_c / \text{Threshold}| < 1 \\
|\Delta P_c / \text{Threshold}|^4, & |\Delta P_c / \text{Threshold}| \geq 1 \\
\exp((2-2\cdot|\Delta P_c / \text{Threshold}|)/2), & 1 \leq |\Delta P_c / \text{Threshold}| < 2 \\
0.5 \cdot |\Delta P_c / \text{Threshold} - 2|^2, & 2 \leq |\Delta P_c / \text{Threshold}|
\end{cases}
\]

In equation (3) and (4), \(\Delta P_c\) follows Eq.(1) and parameter \text{Threshold} is the benchmark threshold, which may have different value for different tone pair. In equation (5), \(\Delta P_c = \Delta P_{c,j} - \Delta P_{c,T}\), in which \(\Delta P_{c,j}\) is pitch connection difference of predicted target, \(\Delta P_{c,T}\) is pitch connection difference of candidate.

Reasonability of cost function defined in equation (3), (4), and (5), can also be supported by the concept presented in [7]. As shown in [7], there are a lot of pitch and rhythm uncertainties among natural utterances. Absolute pitch value of a syllable can always be variable for the same speaker, same text and the same emotion. Syllable’s pitch value is always presented within a certain dynamic range. Based on this, [7] proposes a concept that a syllable’s pitch value is considered to be illegal when it is out of a given range. From this point of view, concept used in defining pitch connection cost in this paper has some similarities to the [7]. In [7], available dynamic pitch range is used to pre-filtering candidates. The filtering procedure can be considered as a hard mask filtering. Compared with this, cost functions defined in our TTS system can be considered as a soft mask filtering.

Furthermore, rationality of tone-based pitch connection control can also be verified from the idea of [8, 9]. [8] shows relationships between two successive Mandarin syllables on pitch contour among one utterance. Offset pitch contour of previous syllable will be changed through some rules to meet onset pitch contour of successive syllable. In other words, there are some relations between two successive syllables’ pitch contour, and pitch contour will be affected by its neighbor syllables. [9] discusses the maximum speed of pitch change. The conclusion of this literature is that the maximum speed of pitch change is often approached in speech. In other words, pitch transition property between two successive syllables is stable for same speaker. It may be true that the relation between two successive syllables’ pitch contour follows some underlying rules.

4. Synthesis comparison

Pitch connection constraint error rate is adopted to evaluate performance of employing pitch connection constraints. Here, both LTC and MTC are considered in evaluation. If one pitch connection pair (tone-based) matches corresponding LTC or MTC (benchmark is the corresponding benchmark threshold), then we count it as right; otherwise, we count it wrong. Equation (6) gives the definition of error rate

\[
\text{Error rate} = \frac{\text{Count\_wrong}}{(\text{Count\_wrong} + \text{Count\_right})} \times 100\%
\]

where, for an evaluation set, Count\_wrong is the total count of pitch connection pairs, which pitch connection property does not match the corresponding pitch connection constraints. Count\_right is the total number of pitch connection pairs, which pitch connection matches the corresponding constraints.

4.1 Matching performance of pitch connection

In this section, matching performance between corpus real pitch connection status and proposed pitch connection constraints is evaluated. Details of this part are as follows.

a) condition 1:

All break connections are not included in evaluation, and words (with two or more Chinese syllables) are separated into isolated syllables. In other words, all un-break connection
pairs that should be controlled by LTC or MTC are counted. Based on this condition, Count_wrong = 8600, Count_right = 38028. Applying the definition Eq.(6), Error_rate = 18.4%.

b) condition 2:

Based on condition 1, and if all intra-word pitch connections pairs are not included in evaluation, then matching performance becomes Count_wrong = 4520, Count_right = 16872. So, following Eq.(6), Error_rate = 21.1%. According to the condition 1 and condition 2’s result, intra-word pitch connection has the performance: Count_wrong = 8600 − 4520 = 4080, Count_right = 38028 − 16872 = 21156. Error_rate = 16.2%. Comparing between matching performance of condition 1, condition 2 and intra-word condition, intra-word has the best pitch connection constraint’s matching performance. There is 21.1% − 16.2% = 4.9% absolute error rate difference between intra-word and inter-word (condition 2). This phenomenon can be briefly explained as following. Ordinary, syllables among one word may have a better smooth/natural transition from one to another, while inter-word transition may have a bigger dynamic transition. Break transition or sentence transition can be acted as an extreme example. From this point of view, pitch constraints proposed in this paper may have a better controlling effect on intra-word’s pitch transition.

4.2. Synthesis experiment

1800 sentences are randomly selected from People’s Daily 1999 to do real synthesis evaluation experiment. Corresponding sentence length distribution is shown in Fig.6, where “Count” represents the number of sentence with corresponding length among the 1800 sentences.

![Figure 6: Length distribution of tested sentences.](attachment:figure6.png)

To emphasize contribution of the proposed pitch connection constraints, intra-word connections and break connections are all excluded from this evaluation.

Based on the 1800 synthesized sentences, error rate of pitch connection is 24.3% for un-employing pitch connection constraints. After employing pitch connection constraints, the error rate decreases to 19.0%. Absolute error rate is dropped by 5.3%. This phenomenon can be briefly explained as follows. Unit selection algorithm in our TTS system is based on cost functions. Each cost component does not have the dominated position and only plays its own contributions to the unit selection. All selected units are under the rule of global optimization. Purpose of employing pitch connection constraints is to decrease unnatural pitch jumps between concatenated syllables. Based on the concept of global optimization and the un-dominated pitch connection cost definition, it is impossible to eliminate all unnatural pitch jumps practically for a real (limited footprint) TTS corpus.

As the proposed pitch connection constraints are derived from un-break connection pairs, corpus pitch connection matching performance of condition 1 (18.4%) can be acted as the benchmark for synthesis evaluation. Based on this, relative error rate difference for employing proposed constraints is (19.0 − 18.4)/18.4 × 100% = 3.3%, while for un-employing proposed pitch connection constraints, it is (24.3 − 18.4)/18.4 × 100% = 32.1%. From this point of view, by applying the pitch connection constraints, synthesized speech has a better similarity with the corpus. Obviously, the most natural synthesized speech is the corpus itself, so it is easy to understand that the more similarity between synthesized speech and corpus, the more naturalness can be possibly achieved. Informal perceptual test on synthesized speech shows that after employing pitch connection constraints, unnatural pitch jumps between syllables is obviously decreased.

5. Summary

The goal of this study is to propose a novel pitch control method for Mandarin TTS system. By applying the proposed method, unnatural pitch connection jumps between concatenated syllables can be reduced significantly. Based on statistical method, three different pitch connection constraints are proposed. According to the constraints, corresponding cost functions are defined. After adding the defined cost functions to the unit selection algorithm, our TTS system achieved better performance. Synthesis experiment and informal perceptual test validate the proposed method.

6. Acknowledgements

The authors would like to thank the TTS team and all the member of Motorola China Research Center (MCRC), for their valued suggestion and discussion on this paper.

7. References