# Chapter 4. Initial Phonetic Segmentation via HMM and DTW

In general, since phones are usually regarded as the smallest units of speech, the goal of phonetic segmentation is to label the boundary of each phone. However, the basic synthesis units are usually syllables in most of Mandarin TTS/SVS systems, i.e., the goal of the phonetic segmentation turns out to label the boundary of each syllable. Fig. 4.1 demonstrates a typical example of the manual phonetic segmentation results of a Mandarin sentence, “請把這籃兔子送走” (“ging2-ba3-zhe4-lan2-tu4-z5-song4-zou3”). In one of our prior work [52], we generated the synthesis units for an on-the-fly Mandarin SVS system by using automatic phonetic segmentation based on the forced alignment of Viterbi search.

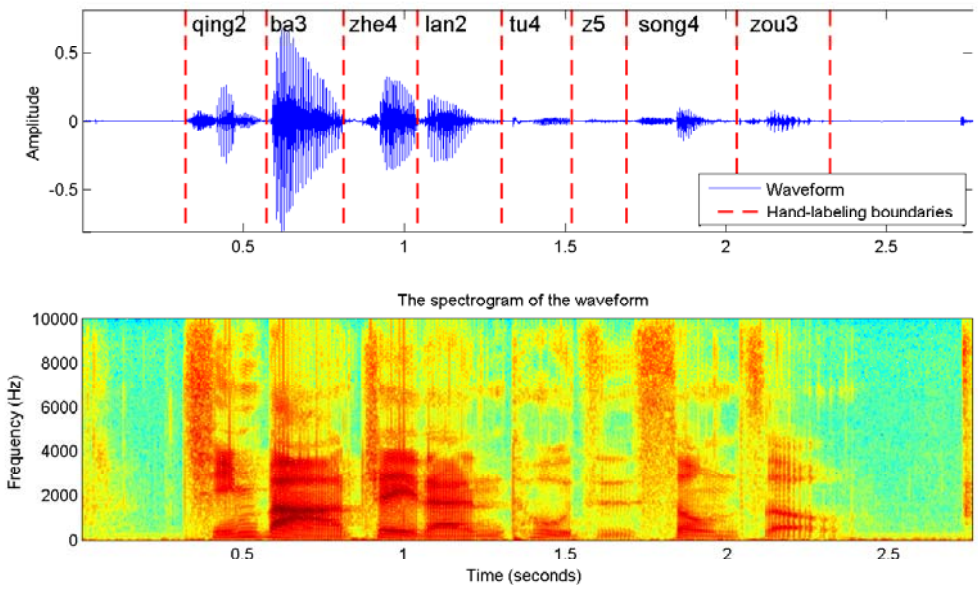


Fig. 4.1. The manual phonetic segmentation results of a Mandarin sentence, “請把這籃兔子送走” (“ging2-ba3-zhe4-lan2-tu4-z5-song4-zou3”).

As noted in Chapter 1, two kinds of initial estimates can be obtained by HMM and DTW. Once the initial estimates are available, we apply the proposed hybrid approach to

perform subsequent boundary refinement. In this chapter, we introduce the speech/singing voice corpora as well as the construction of HMM-based and DTW-based recognizers. Furthermore, we also address the problems of the two kinds of recognizers and make a performance comparison between them.

### *4.1. Speech/Singing Voice Corpora*

In this study, we have two speech corpora and a singing voice corpus which are TTS-455, TCC-300 [53] and SVS-1384, respectively. The detailed descriptions of these corpora are explained as follows.

**TTS-455 (speech data):**

- It is a speech corpus composed of 455 sentences spoken by one professional male recordist.
- Its sampling rate and encoding bit rate are 20 000 Hz and 16 bits.
- It covers about 6000 syllables and overall duration is about 30 minutes (66 MB).
- There are a total of 408 base syllables and 1196 tonal syllables.

**TCC-300 (speech data):**

- The speech corpus composed of 8913 sentences was collected by by National Taiwan University, National Cheng Kung University, and National Chiao Tung University.
- The speech data of each university are provided by 100 speakers (50 males and 50 females). Totally TCC-300 contains speech data from 300 speakers.
- Its sampling rate and encoding bit rate are 16 000 Hz and 16 bits.

- It covers about 334 000 syllables and overall duration is about 27 hours.
- There are a total of 411 base syllables and 1394 tonal syllables.

**SVS-1384 (singing voice data):**

- It is a singing voice corpus composed of 1384 sentences sung by a professional female recordist.
- Its sampling rate and encoding bit rate are 16 000 Hz and 16 bits.
- It covers about 9561 syllables and overall duration is about 70 minutes (132MB).
- There are a total of 385 base syllables.

### *4.2. HMM-based Alignment with MFCCs*

In general, HMM-based recognizers can be categorized in various ways. For example, some use context-dependent HMM, while others use context-independent HMM [54]. Also, there are various types of HMM training methods, including speaker-dependent (SD), speaker-independent (SI), and speaker-adapted (SA) models. To make a performance comparison among these models, we thus employed different types of model training to construct an HMM-based recognizer for forced alignment, as described below:

1) 1st model: SI-based model constructed by using the TCC-300 corpus.
2) 2nd model: SD-based model constructed by using the TTS-455 corpus, with uniform segmentation.
3) 3rd model: SD-based model constructed by using the TTS-455 corpus, with initial segmentation performed by the model trained using the TCC-300 corpus.

4) 4th model: SI-based model constructed by using the TCC-300 corpus first and then adapted by using the TTS-455 corpus. In other words, this model is SA-based.
5) 5th model: SI-based model constructed by using the TCC-300 corpus first and then adapted by using the SVS-1384 corpus. In other words, this model is SA-based.

Each of these acoustic models was constructed based on context-dependent tri-phones. The acoustic feature used in each model is 39-dimensional MFCCs (13 static, 13 delta and 13 delta delta). In addition, since TTS-455 and TCC-300 have different sampling rates, down sampling procedure should be performed on TTS-455 corpus (higher sampling rate) before training 1st model and 4th model for achieving the higher recognition rate. In our implementation, all these models are constructed by using HTK software [55]. The difference between the 2nd and 3rd models lies in the initial segmentation for training. The 2nd model uses uniform segmentation, while the 3rd model uses the segmentation derived by the recognizer trained using the TCC-300 corpus. Both of them can be viewed as SD-based models derived from the TTS-455 corpus.

For SA-based models, two common methodologies including the maximum a posteriori (MAP) [56] estimation and maximum likelihood linear regression (MLLR) [57] method can be adopted. In this study, we constructed SA-based models with the MLLR method. MLLR is a model adaptation technique that estimates a set of linear transformation for the mean and variance parameters of a Gaussian mixture HMM system. It is noted that only the mean vector of each Gaussian density function is adapted in this study.

In order to evaluate the performance of these different types of recognizers, we performed forced alignment over the TTS-455 and SVS-1384 corpora whose all boundaries

were labeled by human in advance. The assessment of performance is to estimate the segmentation accuracy within different tolerant interval, i.e., the percentage of the cases whose error range is less than N milliseconds. Here error range means the absolute difference between a true boundary (human labeled) and an automatic-segmentation boundary. For example, we have 50% cases whose error range < 10 ms after the forced alignment via a HMM-based recognizer. Table 4.1 shows the segmentation results of forced alignment by using different training models. It should be noted that the evaluated data for all training models is TTS-455 except for 5$^{th}$ model. SVS-1384 is used to evaluate the performance of 5$^{th}$ model-training recognizer. It can be seen from Table 4.1 that the 4$^{th}$ model outperforms other three models (1$^{st}$, 2$^{nd}$ and 3$^{rd}$) in most of cases; we thus chose the 4$^{th}$ model as the HMM-based recognizer to perform the phonetic segmentation on speech data.

According to the experimental results, it appears to be useful to adopt the speaker-adaptation model for forced alignment. Consequently, we chose the 5$^{th}$ model as the HMM-based recognizer to perform the phonetic segmentation on singing voice data. However, the segmentation accuracy of the 5$^{th}$ model is evidently lowest as compared with other models. It is probably caused by that the initial SD-based model is trained through using speech data (TCC-300) instead of singing voice data. That is, there is exactly the instinct difference between speech and singing voices. For example, the pitch range variation of a singing voice is much wider than that of speech; the average singing rate is generally slower but has a greater variance; the sounding effects (portamento, vibrato, etc) that frequently appear in singing voices seldom occur in speech. Fortunately, since the music score information is available for each sentence in our singing voice corpus (SVS-1384), we thus employ dynamic time warping (DTW) with pitch contours to perform the phonetic

segmentation on singing voices. We shall give the detailed introduction of DTW-based recognizer in the coming section.

**TABLE 4.1 SEGMENTATION ACCURACY RESULTS W.R.T. DIFFERENT TRAINING MODELS**

| Segmentation Error Tolerance / Model – Evaluated Data | **≦10 ms** | **≦20 ms** | **≦30 ms** | **≦50 ms** |
|---|---|---|---|---|
| 1st model – TTS-455 | **49.47%** | 70.58% | 84.24% | 95.10% |
| 2nd model – TTS-455 | 45.02% | 69.83% | 81.96% | 92.36% |
| 3rd model – TTS-455 | 43.49% | 65.55% | 79.60% | 91.51% |
| 4th model – TTS-455 | 46.09% | **72.07%** | **87.40%** | **95.80%** |
| 5th model – SVS-1384 | 22.53% | 51.62% | 68.31% | 80.33% |

### *4.3. DTW-based Alignment with Pitch Contours*

In addition to traditional segmentation method via HMM, DTW approach could be qualified to the task as well. In past studies, DTW has been adopted successfully in melody recognition [30][31]. Furthermore, we had developed an automatic singing voice rectifier system [32] successfully and obtained acceptable performance. It is noted that the feature used in the DTW-based alignment is pitch information instead of MFCCs. This is because that each sentence in SVS-1384 has its corresponding music score information, it is thus intuitive to utilize pitch information as the feature to carry out DTW-based time alignment. However, on the other hand, it also implies that the pitch contours need to be estimated as correctly as possible such that the DTW-based alignment is able to achieve the optimum performance. Since correct pitch tracking is critical to both DTW and singing voice synthesis, we used a robust pitch tracking method [46] in our implementation.

Let us take an example to briefly introduce how DTW works. For a singing voice sentence, supposed that the input pitch (semitone) vector is represented by $t(i) = 1,2,...,M$ ,

and the referenced pitch vector is represented by $r(j) = 1,2,...,N$. The two vectors are not necessarily of the same length. Then we can construct a $M \times N$ DTW table using the following recurrence (4.1).

$$D(i,\ j) = \left|t(i) - r(j)\right| + \min \begin{Bmatrix} D(i-2,\ j-1) \\ D(i-1,\ j-1) \\ D(i-1,\ j-2) \end{Bmatrix} \qquad (4.1)$$

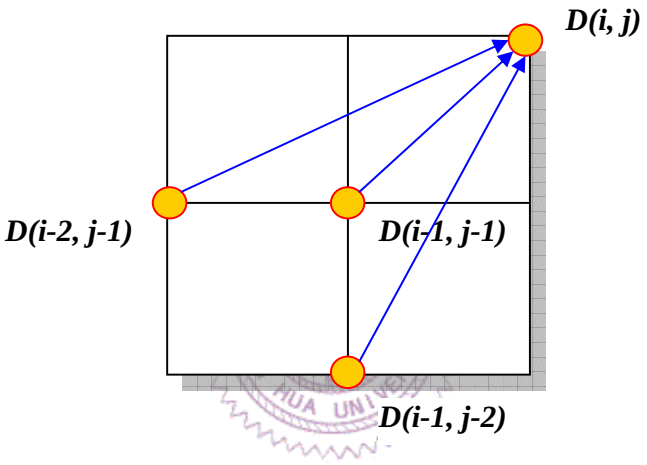


Fig. 4.2. The local constraint of DTW alignment.

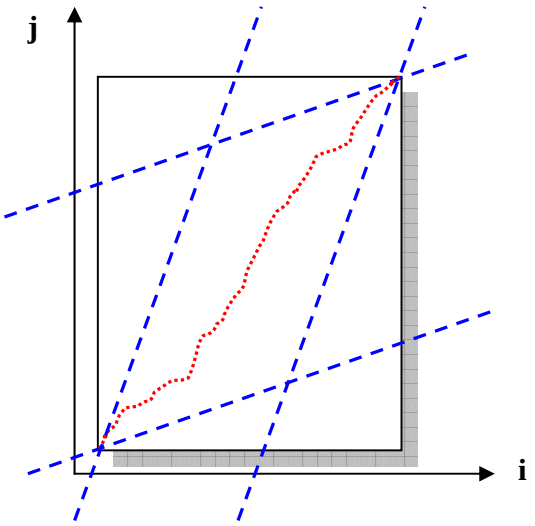

Fig. 4.3. The global constraint of DTW alignment.

Fig. 4.2 shows the local constraint of DTW and Fig. 4.3 demonstrates the global constraint of DTW in our study. After the DTW table is constructed, we can find the optimal ending point as the minimum element of the last column. The corresponding best alignment path can be obtained through back tracking. However, the key transposition must be considered since the recording artist is likely to drift slightly from the melody of a song. Accordingly, we shifted the identified pitch vector to the same mean of the reference pitch vector at the first. Subsequently, we employed an iterative procedure to shift the identified pitch incrementally by 0.1 semitone within the range [-3, 3], in order to find the best amount of shift that can generate the minimum DTW distance between the input pitch and its corresponding reference pitch. Fig. 4.4 shows a typical example of DTW-based alignment. Fig. 4.5 demonstrates an example of the key transposition. Table 4.2 shows the performance of the DTW-based alignment. Here the evaluated data is SVS-1384.

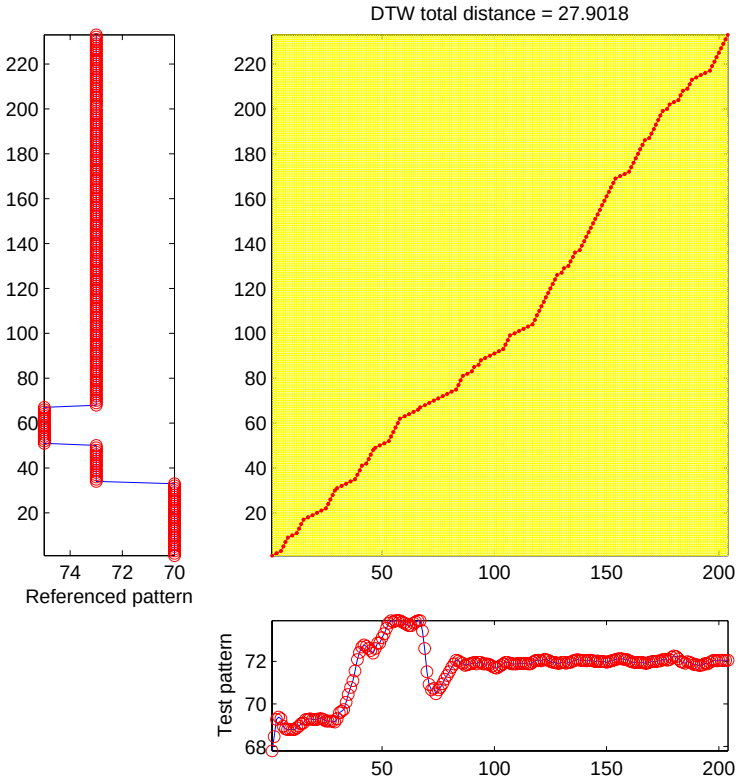


Fig. 4.4. An example of DTW-based alignment.

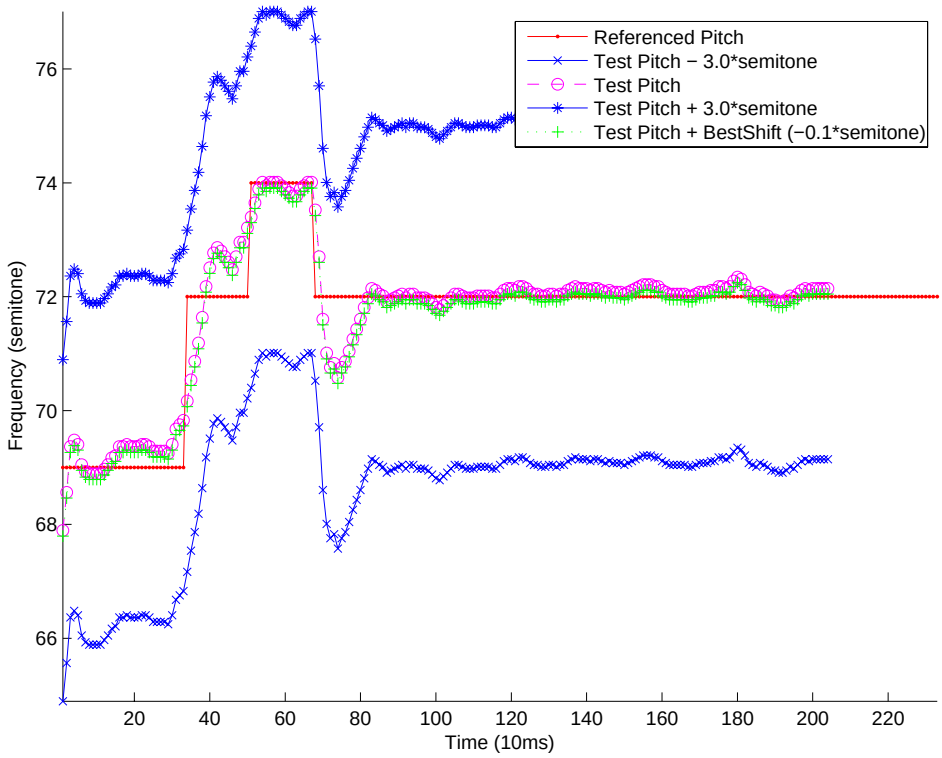


Fig. 4.5. The key transposition for getting the best amount of pitch shift.

**TABLE 4.2 THE SEGMENTATION RESULTS OF DTW-BASED ALIGNMENT.**

| Segmentation Error Tolerance | **≦10 ms** | **≦20 ms** | **≦30 ms** | **≦40 ms** | **≦50 ms** |
|---|---|---|---|---|---|
| DTW-based alignment | 49.20% | 65.12% | 70.73% | 73.71% | 76.22% |

Through the comparison between Table 4.1 and Table 4.2, DTW-based alignment produces better results in the cases which have smaller segmental errors (≦30 ms) as compared with HMM-based alignment, whereas it does not outperform in the other cases that have larger segmental errors (>30 ms). Moreover, no matter which method we choose, the overall performance difference between the two methods is not apparent. Nevertheless, if we are able to combine the segmental results obtained by the two methods perfectly, it will improve the performance of phonetic segmentation obviously. Table 4.3 shows the

performance after the perfect combination of the different segmental results obtained from the HMM-based and DTW-based recognizers. Here the perfect combination means that the result with smaller segmentation error is always able to be picked out between the given two segmental results obtained by HMM and DTW. (The evaluated data is SVS-1384.)

**TABLE 4.3 THE SEGMENTATION RESULTS OF THREE DIFFERENT MANNERS.**

| Segmentation Error Tolerance | **≦10 ms** | **≦20 ms** | **≦30 ms** | **≦40 ms** | **≦50 ms** |
|---|---|---|---|---|---|
| HMM-based alignment with MFCCs | 22.53% | 51.62% | 68.31% | 76.11% | 80.33% |
| DTW-based alignment with pitch contours | 49.20% | 65.12% | 70.73% | 73.71% | 76.22% |
| Perfect combination of DTW and HMM | **59.52%** | **78.38%** | **85.62%** | **89.08%** | **91.46%** |

It can be seen from Table 4.3 that the overall performance after the perfect combination is obviously much better than that of both the previous two methods (HMM and DTW). In other words, if we could integrate the two segmental results of HMM and DTW effectively, the phonetic segmentation would be more reliable for corpus-based SVS.