Chapter 1. Introduction

Corpus-based speech synthesis is an important approach to text-to-speech (TTS) due to its high degree of fluency and the natural feel of the generated speech. There are numerous studies on corpus-based TTS [1]-[5] during the past few years. The essence of corpus-based TTS is to employ an effective unit selection scheme to pick up the suitable synthesis units directly from the corpus as the desired output. These units are usually non-uniform, such as a diphone, a syllable, a word, and a sentence. In addition, the prosodic modification is probably needed for corpus-based TTS due to the limited speech corpus. Recently, the corpus-based approach has also been applied to the singing voice synthesis (SVS) [6][7]. Generally speaking, the better synthetic quality can be achieved if a larger corpus is available. In other words, collecting a very large corpus for corpus-based TTS/SVS systems seems to be a tendency for current speech/singing voice synthesis research. These corpus-based systems require a large amount of efforts in several essential preprocessing tasks, such as corpus design/collection/recording, pitch estimation/marking, phonetic transcription/segmentation, etc. Fig. 1.1 demonstrates these preprocessing tasks of corpus-based TTS/SVS.

Here we briefly introduce these tasks shown in Fig. 1.1. Initially, we collect a speech/singing voice corpus based on a corpus design method. After the corpus was design, each text sentence was recorded very carefully by a professional recordist to ensure the consistency of volume or speaking rate among whole data. Once the corpus is prepared, we can apply the forced alignment of the Viterbi search using a hidden Markov model (HMM) or a word segmentation algorithm to obtain the phonetic transcription. Subsequently, we

perform the automatic phonetic segmentation to acquire the position of each phoneme.

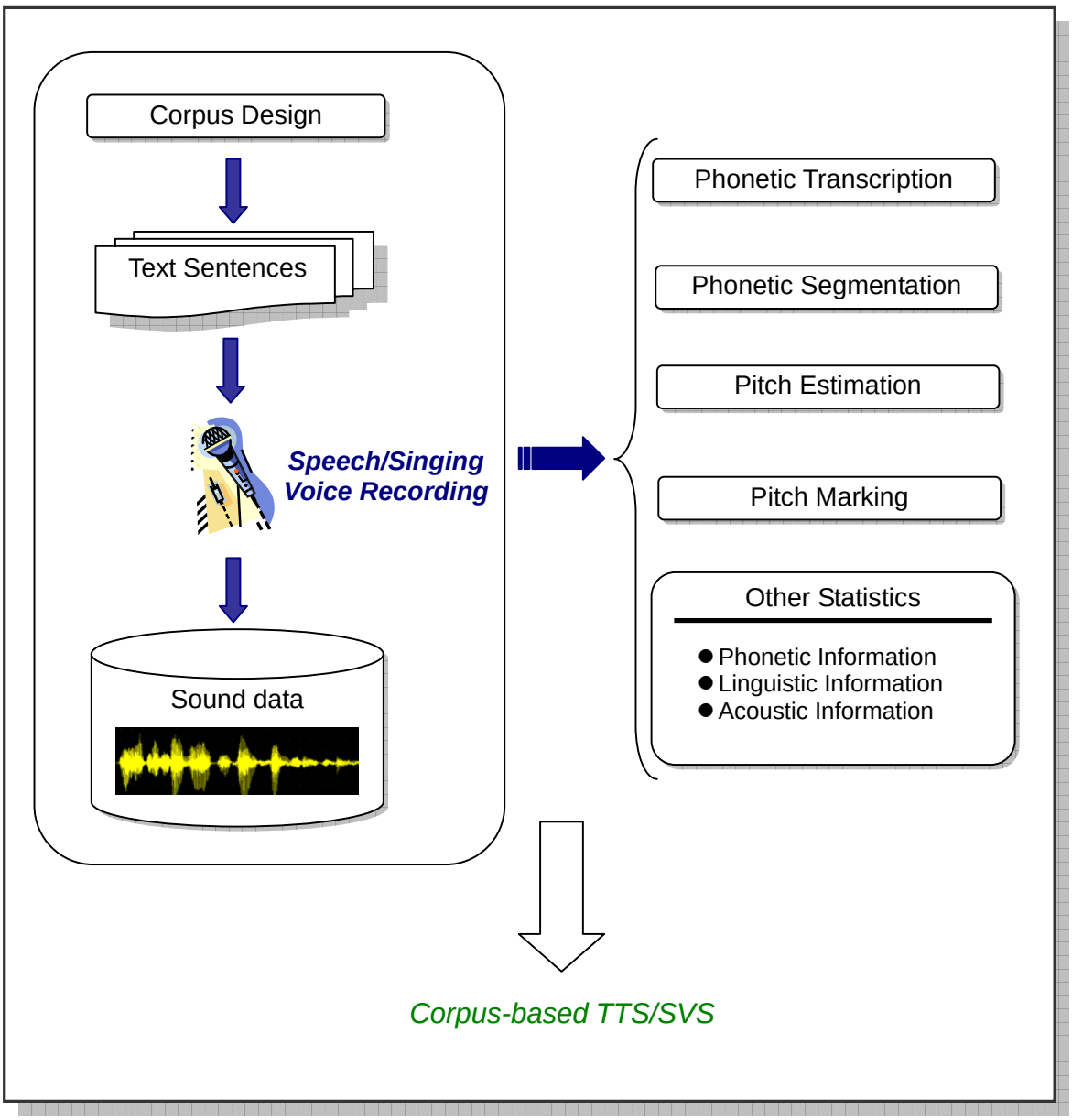

Fig. 1.1. The preprocessing tasks of corpus-based TTS/SVS

Correct pitch estimation is very significant to corpus-based TTS/SVS systems. It is necessary to utilize pitch and other acoustic information to measure the prosody difference between target units and source units when processing the unit selection procedure or

before carrying out the pitch-scale modification. Various pitch estimation methods have been proposed in the literature. These methods include the autocorrelation factor [8], pattern recognition [9], least-square fitting [10], etc. Nevertheless, the manual checking to ensure the correctness of pitch estimation results is needed no matter which methods we adopted.

Normally, in order to obtain the better speech quality with efficient computing, altering pitch contours usually relies on time-domain pitch synchronous overlap and add (TD-PSOLA) algorithm [11] instead of other methods, such as sinusoidal modeling [12], harmonic noise modeling [13], and so on. An essential part of TD-PSOLA is pitch marking, which tries to find the glottal closure instant (GCI) in order to perform synchronous analysis. If the result of pitch marking is not good enough, TD-PSOLA will produce low-quality speech. Moulines *et al.* [14] proposed a pitch marking method to detect abrupt changes at GCI. In one of our prior studies [15], we proposed an effective pitch marking algorithm based on a two-phase pitch marking concept. Although the method has satisfactory performance on labeling pitch marks, the manual checking/revisions were processed further in order to ensure the correctness of all pitch marks.

On the other hand, we also gathered statistics of corpus data which includes several types of information concerning the corpus; they are phonetic, linguistic, and acoustic information. For phonetic information, we analyze the distribution of tones, vowels or consonants. For linguistic information, we extract part-of-speech (POS), word length, sentence length, etc. For acoustic information, we calculate the duration, energy, pause, etc. These statistics are usually essential for speech/singing voice synthesis.

The phonetic segmentation plays a significant role among these fundamental tasks

mentioned above. If the boundary information of each phone/syllable is not accurate enough, the statistics of acoustic information or the results of pitch estimation/marking will be also unreliable. In other words, we will require a lot of manual efforts to revise these errors. In view of this, effective phonetic segmentation is practical to the developments of corpus-based TTS/SVS. In this dissertation, we will focus on the issues of phonetic segmentation.

Admittedly, for speech/singing voice synthesis the precise phonetic segmentation is significant because that a small segmentation error may cause an audible error in the synthetic sound. In order to obtain high-quality synthetic speech, the phonetic segmentation usually relies on large amounts of manual efforts in the past. However, the task is extremely labor intensive and time consuming especially when the corpus size is very large. Moreover, to develop TTS/SVS systems with different languages and voices quickly, using automatic phonetic segmentation techniques are usually necessary. Generally speaking, methods for automatic phonetic segmentation involve two essential steps. First we perform a rough phonetic segmentation by forced alignment of the Viterbi search using a HMM with Mel frequency cepstral coefficients (MFCCs) [16]. Then we apply a boundary refinement procedure as a post-processor to fine-tune the results obtained by the HMM.

In theory, for speech recognition the use of HMMs does not rely on manual phonetic segmentation. An initial segmentation (ex. flat segmentation) is used directly to train the HMMs since HMM training is an averaging process that tends to smooth segmentation errors. Nevertheless, some studies point out that speech recognition can benefit from more precise initial segmentation in training [17][18]. The HMM-based recognizers can be categorized in various ways. For example, some use context-dependent HMM, while others

use context-independent HMM [19]. Also, there are various types of HMM training methods, including speaker-dependent, speaker-independent and speaker-adapted models. Although the HMM-based speech recognizer using MFCCs is well known for its excellent speech recognition ability, its use of automatic phonetic segmentation does not always produce precise and satisfactory results for the development of TTS/SVS. That is the reason why a boundary refinement procedure is usually employed to refine the results obtained by the HMM.

Several studies [20]-[27] have focused on the boundary refinement in the past few years. For example, Jan P. H. van Santen *et al.* [20] adopted broad-band and narrow-band edge detection. Bonafonte *et al.* [21] took Gaussian probability density distribution as a similarity measure. Toledano *et al.* [22] tried to mimic human labeling using a set of fuzzy rules. In [23], Chou *et al.* proposed a speaker-dependent HMM model plus simple boundary correction rules for Mandarin Chinese. Wang *et al.* [24] proposed a post-refining method with fine contextual-dependent Gaussian mixture models (GMMs) and employed classification and regression tree (CART) to cluster acoustically similar GMMs. A multilayer-perceptron (MLP) was also applied to achieve an improved accuracy of segmentation [25][26][27]. Drawbacks of some of the representative approaches are summarized as follows:

1) Toledano *et al.* [22] adopted multiple features based on manually tuned subject rules for each category of phonetic transition. This approach is labor intensive and not easily scaled up because different phonetic sets have different rules to be identified.

2) Wang *et al.* [24] proposed the use of MFCCs alone to refine the boundaries of all categories of phonetic transitions. This is too assertive. For example, if a boundary is of the case “silence + fricative”, other simple features, such as energy, may well outperform MFCCs.
3) Constructing the system proposed by Chou *et al.* [23] is time consuming because of the iterative procedure used for forced alignment, the correction rules and, re-training. In addition, it becomes particularly inefficient if the speech corpus is updated incrementally and regularly, such as by adding one hour of speech data per week.
4) In particular, most methods mentioned above do not elaborate the issue of error analysis, as what categories of phonetic transitions tend to be more error-prone and how to deal with these transitions.

In theory, it should be better if a boundary refinement is constructed by using multiple acoustic features rather than single feature alone (e.g. MFCCs). Moreover, using statistics-based methods to refine boundaries are more reliable than using rule-based methods in general. In addition, refining boundaries would be more effective if we can adopt different acoustic features for different phonetic transition categories.

According to these viewpoints mentioned above, initially we propose an effective boundary refinement approach to refine the boundaries according to various phonetic transition categories. For each phonetic transition category, a set of training data is collected and split into two classes, “correct” and “wrong” according to the distance to a true, manually labeled, boundary. Subsequently, we adopt statistical pattern recognition

(sequential forward selection, k-nearest neighbor rule and leave-one-out error) to identify those most valuable features for each phonetic category. Eventually, a fixed search range is used in the boundary refinement procedure. As a matter of fact, this statistical approach does not necessarily work well for all phonetic categories; we thus apply a heuristic method additionally to refine the boundaries in some phonetic transition categories which tend to be more error-prone. Since the proposed approach combines statistics and heuristic concepts, we refer to it as a hybrid approach.

In addition, we also proposed another refinement method which is based on a score prediction concept. Each candidate boundary can be evaluated via the proposed score predictive model (SPM). If the predicted score of a candidate boundary is higher, its location will be possibly closer to the true boundary. The detailed description of the two methods (hybrid and SPM) shall be elaborated in Chapter 5 and Chapter 6. The flowchart of automatic phonetic segmentation on speech corpora is depicted in Fig. 1.2.

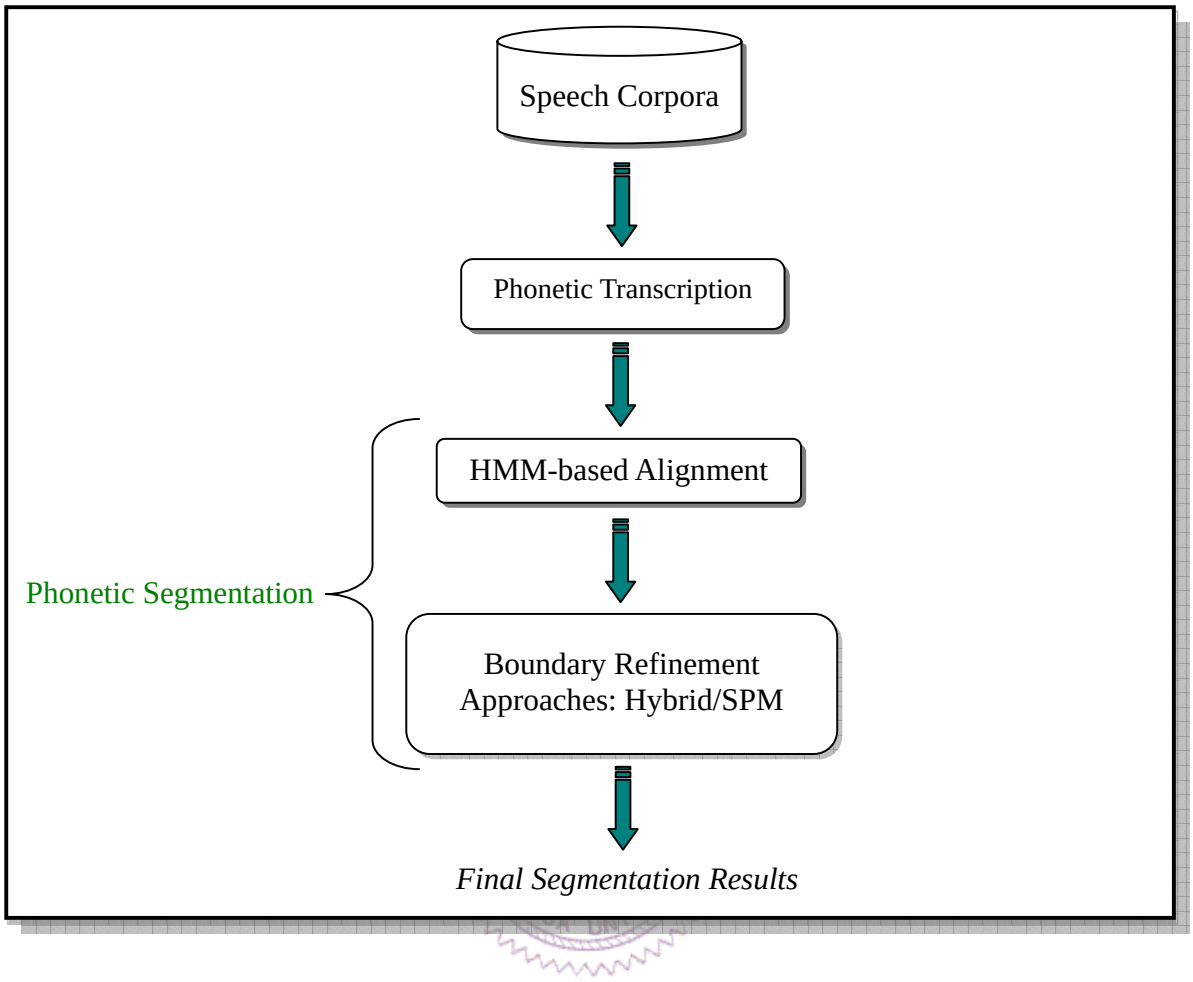

Fig. 1.2. The flowchart of automatic phonetic segmentation on speech corpora**.**

Similarly, employing the same phonetic segmentation methods on singing voice corpora should be also feasible. Unfortunately, the performance of HMM-based segmentation on singing voice corpora is not as satisfactory as that on speech corpora. This is probably due to several aspects of the physical differences between the singing voice and speech. 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. Furthermore, there is no readily available HMM-based recognizer specifically

designed for singing voices.

In fact, if the initial segmentation identified by HMM is not satisfactory, then the corresponding post-processing will not be able to refine the boundaries effectively. Related experiments conducted by Park *et al.* [27] indicated that it is very difficult to cope with the problem of large segmentation errors using a boundary refinement post-processor. Therefore, there is a need to improve the results of the HMM-based recognizer for phonetic segmentation. Unfortunately, this seems to be a difficult task according to the literature. For instance, Kawai and Toda [28] found that only a slight improvement in segmentation was obtained by using various acoustic model structures (such as changing numbers of states in a model, or numbers of Gaussian mixtures in a state) and acoustic model units (such as moving from monophones to biphones, triphones, and so on). Moreover, in a similar study [29], Lee showed that the automatic segmentation performance is slightly degraded when an embedded-reestimation procedure is employed that uses an utterance and its transcription for sentence-based training.

Based on the above mentioned viewpoints, it is difficult to significantly improve the performance of the HMM-based recognizer. Fortunately, the recording artist was required to sing a song by following the corresponding melody information, from a music score, when the singing voice corpus was collected. Consequently, we could perform phonetic segmentation by aligning the singer's pitch information with the corresponding melody information based on dynamic time warping (DTW). In past studies, DTW has been adopted successfully in melody recognition [30][31]. In particular, we have used DTW to perform note segmentation with good performance in our previous study on automatic singing voice rectification [32]. Hence, we can use DTW with pitch features as an

alternative for tackling the problem of phonetic segmentation. However, our experiments demonstrated that the performance of DTW is not as high as anticipated. One of the reasons is that if two neighboring syllables have the same pitch it is quite difficult to identify the boundary via DTW. On the other hand, it turns out that this situation can be handled by the HMM with MFCCs features, as long as the two syllables have a different pronunciation. Similarly, co-articulation between two syllables usually poses a difficult problem for HMM, but it is never a problem for DTW as long as they have different pitch. In other words, the two initial estimates identified by HMM and DTW seem to be complementary to each other and should be integrated for better performance. Hence in this study we try to use the proposed hybrid/SPM approaches to integrate the two initial estimates obtained by HMM and DTW. Fig. 1.3 displays the flowchart of the automatic phonetic segmentation on speech corpora.

This dissertation is organized as follows. Chapter 2 introduces the related work concerning the automatic phonetic segmentation. Chapter 3 describes the preprocessing of corpus-based TTS/SVS systems. Chapter 4 explains the construction of the HMM-based and DTW-based alignment. Chapter 5 elaborates the boundary refinement based on a hybrid approach. Chapter 6 introduces the boundary refinement based on a score predictive model concept. Chapter 7 presents conclusions and discusses future work.

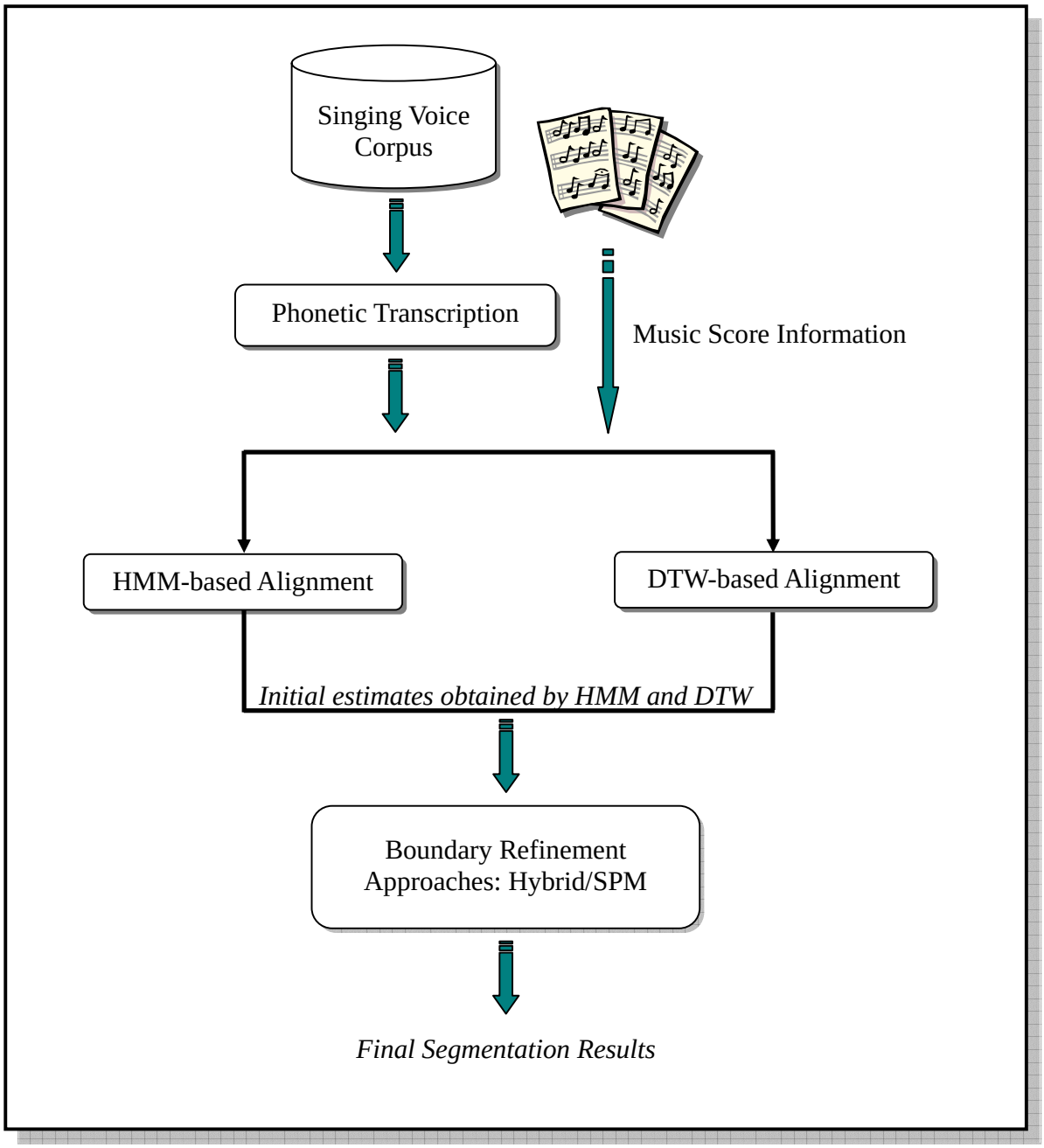

Fig. 1.3. The flowchart of automatic phonetic segmentation on singing voices.