Chapter 2. Related Work

In this chapter, we introduce several studies concerning automatic phonetic segmentation which usually includes two fundamental steps, HMM-based explicit segmentation and a post-processing boundary refinement.

2.1. HMM-based Phonetic Segmentation

Studies on automatic phonetic segmentation are closely related to previous knowledge of automatic speech recognition. The HMM-based approaches play an important role in automatic phonetic segmentation [33]-[35]. Generally speaking, context-dependent HMM can provide better representation of the spectral movements in phonetic transitions than context-independent HMM, whereas context-independent HMM possibly has more precise segmentation than context-dependent HMM [36]. This is possibly due to that context-dependent HMM are trained with realizations of phones in the same context, the HMM has no any information to discriminate between the phone and its context accordingly. On the other hand, since context-independent HMM are trained with realizations of phones in different contexts, in theory the HMM should be able to discriminate between the phone and its context.

Based on these viewpoints, it seems better to train a context-independent HMM recognizer instead of a context-dependent HMM one for automatic phonetic segmentation. Toledano *et al.* [37] conducted several experiments to make a performance comparison between the two kinds of HMM training (context-dependent vs. context-independent). They indicated that for smaller tolerances (5-10 ms) context-dependent HMM with fewer

Gaussians behave better. For medium tolerances (15-30 ms) context-independent HMM with fewer Gaussians are better. For large tolerances (>35 ms) context-dependent HMM with more Gaussians have better results. Finally, they determined to train a context-dependent HMM recognizer to phonetic segmentation. The choice is reasonable because that if most of cases had small initial segmentation errors (ex. <50 ms), the following boundary refinement would be able to perform well. In fact, a plenty of researchers (e.g. [22]-[25][29][34]) choose to train a context-dependent HMM recognizer rather than context-independent one for automatic phonetic segmentation.

2.2. Boundary Refinement

Once the initial phoneme boundaries are labeled by a HMM-based recognizer, the subsequent task is to refine these initial boundaries to correct locations as possible via a boundary refinement procedure. There have been many studies in the literature concerning automatic phonetic segmentation of speech corpora during the past few years. Generally speaking, these studies can be classified into the rule-based and the statistics-based methods.

For the rule-based methods, a few human defined correction rules were used to refine the phoneme boundaries of speech corpora [22][23][38][39]. In [22], Toledano *et al.* tried to mimic human labeling using a set of fuzzy rules. In [23], Chou *et al.* proposed a speaker-dependent based HMM model plus simple boundary correction rules for Mandarin Chinese. In [38], Houben believed that acoustic-phonetic knowledge in combination with the phonetic transcription can be helpful in determining the label positions. In [39], K. Hatazaki *et al.* presented a method for phoneme segmentation by an expert system utilizing

spectrogram reading strategy and knowledge. There is no denying that it is somewhat difficult to design these systems mentioned above without the aid of human experts. On the other hand, it is not easily scaled up due to the fact that different phonemes from various languages have individual rules to be identified.

For the statistics-based methods, there have been numerous studies proposed in the past few years [20][21][24]-[29][40]. For example, in [40], van Hermert concluded that using both implicit and explicit segmentation together can segment more precisely and produce better speech quality than using explicit segmentation alone. Here an implicit segmentation algorithm splits up the utterance into segments on the basis of the degree of similarity between the frequency spectra of neighboring frames, but an explicit algorithm does the same thing based on the degree of similarity between the frequency spectra of the frames in the utterance and reference spectra.

In [21], Bonafonte *et al.* took Gaussian probability density distribution as a similarity measure. Wang *et al.* [24] proposed a post-refining method with fine contextual-dependent GMMs and employed CART to cluster acoustically similar GMMs, so that the GMM for each leaf node is reliably trained by the limited manually labeled boundaries. The acoustic feature used for training is referred to as the super vector demonstrated in Fig. 2.1. It is noted that the basic acoustic feature is m-dimensional MFCCs alone. Here m is usually set to 39. It is likely inadequate because that the method uses MFCCs alone to refine the boundaries of all categories of phonetic transitions. For example, if a boundary is of the case “silence + fricative”, other simple features, such as energy, may outperform MFCCs.

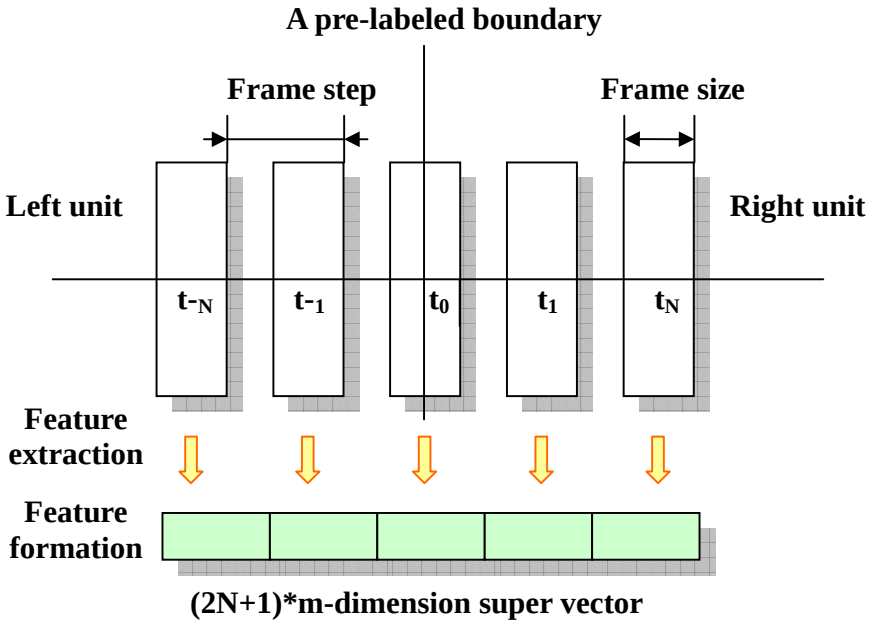

Fig. 2.1. According to [24], the representation of the super vector for a boundary.

On the other hand, a multi-layer perceptron (MLP) was also employed to achieve an improved accuracy of phonetic segmentation [25][26][27][29] . For example, Lee [29] proposed a boundary detector based on MLPs. To increase the accuracy of phonetic segmentation, several specialized MLPs were individually trained based on phonetic transition. The optimum partitioning of the entire phonetic transition space and the corresponding MLPs were constructed from the standpoint of minimizing the overall deviation from the hand-labeling position. Fig. 2.2 shows the overall block diagram of Lee's MLP-based method.

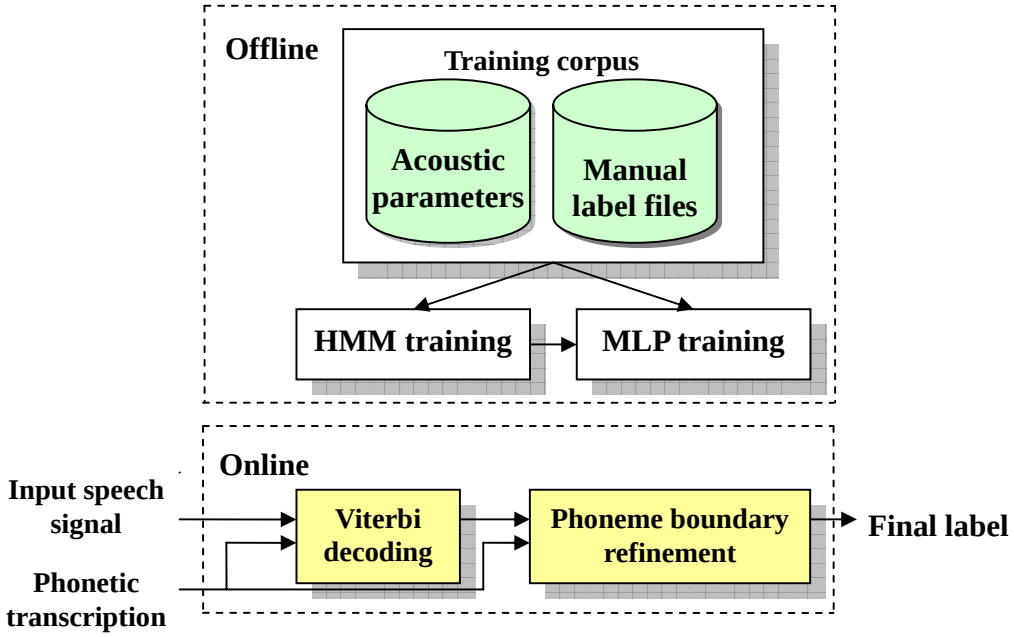

Fig. 2.2. Block diagram of an MLP-based phone boundary refining system.