Pattern Matching in Parameter Streams for ASR

Pattern matching task in ASR can be posed at the phoneme level, sylable level, word level or combinations of these. Here we look just at the phoneme matching task .

Acoustic parameters are extracted from the speech signal on a frame basis constructed from 5 to 20 ms time slices. Combinations of parameters, suitably scaled, are presented to the matcher as an acoustic vector time stream. The goal of the pattern matching task is to generate a ranked list of possible phonemes for each frame and then attempt to match these incoming phoneme sequences to phonemic representations of words in a dictionary. The imput AV stream and a phoneme stream from a word hypothesis are illustrated in Figure 1.

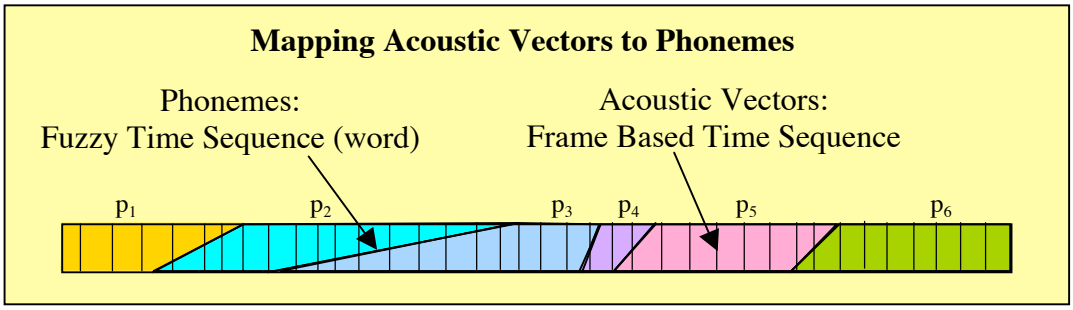

Figure 1: Stylised AV and phoneme streams for approximately 500ms of speech

A ranked list of possible phonemes is generated for each frame using liklihood tables accumulated from training data. A typical rank stream is illustrated in figure 2. A rank stream can be generated for each possible combination of acoustic parameters in the AV stream. From these rank streams we want to generate a ranked shortlist of possible phonemes.

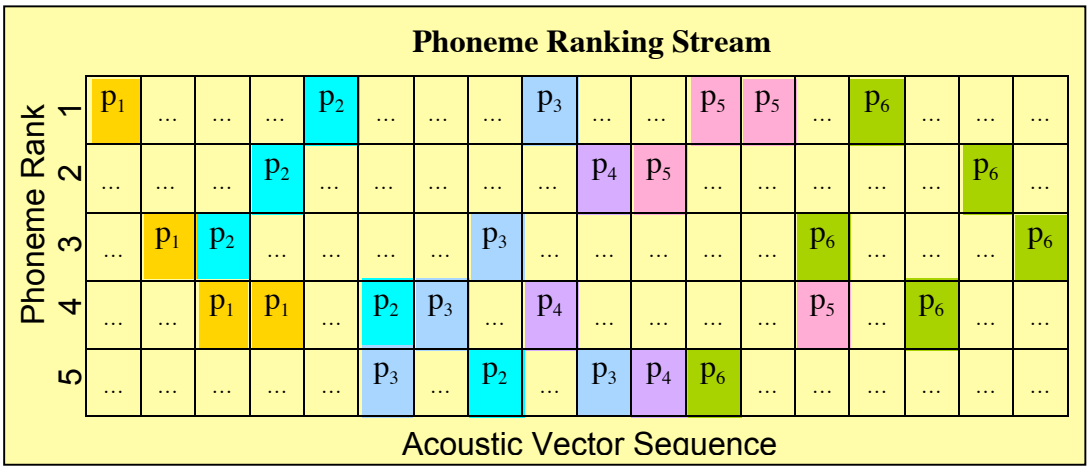

Figure 2: Time stream of phoneme rankings for input frames. For simplicity only relevant phonemes have been shown.

Once a shortlist has been generated for each frame the most likely candidate phoneme must be chosen using one-on-one discrimination tests between the candidates and through continuity constraints based on adjacent frames.