Connected digit speech recognition has an important role to apply in many speech applications such as voice- telephone dialing. The major problem in connected digit speech recognition system is to detect the boundary of each word in digit string. Moreover, the language model can not be used to enhance the system precision, like the other speech recognition systems, such as dialogue speech recognition system. Nevertheless the speaker-independent speech recognition system is more practical than the speaker-dependent one. Thus this research presents the Speaker-Independent Thai Connected Digit Speech Recognition using Hidden Markov Model.

The system consists of five steps, which are speech signal preprocessing for preparing speech signal into speech frame, speech feature extraction for distilling the necessary information of speech and representing as the speech feature parameters, pattern training for learning speech feature in each digit, pattern classification for comparing the similarity between reference models and speech input, and decoding process for finding the best sequence result. This research focuses on speech feature extraction, pattern training and pattern classification. Speech feature extraction consists of 6 features, which are the Mel Frequency Cepstral Coefficients (MFCC), delta MFCC, delta-delta MFCC, energy, delta energy and delta-delta energy.
To find appropriate speech features, we varied feature parameters from 2 to 6 features. In pattern training and pattern classification, the Continuous Density Hidden Markov Model (CDHMM) is used to train the speech patterns and to recognize the speech input. Finally the Viterbi search algorithm is applied for decoding process. The important parameters of CDHMM are the number of states and the number of Gaussian mixtures.
In this research we adjusted these two parameters to find an appropriate value. The experimental number of states for CDHMM and the number of Gaussian mixtures are starting from 5 and 2 respectively.

For the experiment, there are two sets of samples; training and testing sets. In training set, we use data from 100 speakers (50 females, 50 males) for 2000 utterances. Each utterance is seven digit lengths. In testing set, 40 speakers (20 females, 20 males) are used. The best model is the model that uses 5 mains of speech feature (41 values) with 9 states and 8 Gaussian mixtures. The average word accuracy is 74.25%.

Amarin Deemagarn

Student's signature

Asanee Kawtrakul 20 /07 /2007

Thesis Advisor's signature