Phoneme Response Sequence Recognition Using Hidden Markov Model

X. Dong and H. K. Kwan
Department of Electrical and Computer Engineering
University of Windsor
401 Sunset Avenue, Windsor, Ontario, Canada N9B 3P4
kwan1@uwindsor.ca

Abstract

In this paper, phoneme response sequence recognition using the Hidden Markov Model (HMM) is presented. The Self-Organizing Map and the Learning Vector Quantization are used to organize the mel-frequency cepstral coefficients feature vectors of short and long phonemes segmented from speech samples to obtain the phoneme maps. The optimal phoneme response sequence of a speech sample displayed on the phoneme maps are determined by the Viterbi algorithm. Such a phoneme response sequence is presented as an observation sequence to the HMMs representing the 0-9 digits to determine the most-likely digit-voice utterance. Simulation results are given.

1 Introduction

Automatic speech recognition is a challenging area of research due to the dynamic nature of speech. Hidden Markov Models (HMMs) are popular for building speech recognition systems [1]-[4]. Some recent studies in speech recognition involve a combination of artificial neural networks (ANNs) and the HMMs to form hybrid systems [5]-[6]. This approach benefits from the discriminative nature of ANNs and the temporal modeling capability of HMMs. In this study, the Self-Organizing Map (SOM) and Learning Vector Quantization (LVQ) [7] are used to self-organize phoneme feature vectors to form phoneme maps for speech recognition. Transitions between phonemes of human utterances are usually smooth and continuous, which will result indistinct and obscure feature vectors at these phoneme transitions due to tone variations. These indistinct and obscure feature vectors lead to incorrect classification or recognition results. In this study, the Viterbi search algorithm [4] is used to reduce the effect of this variation associated with the obscure phoneme transitions by searching the optimal phoneme response sequence. Consequently, the phoneme response sequence obtained from the phoneme maps become more stable and similar for different utterances of the same speech, by the same person and even by different persons. With such phoneme response sequences, the problem of the robust speech recognition is now transformed into an invariant phoneme response sequence recognition problem. Given an observation of phoneme response sequence, the HMMs are used to determine the most-likely speech utterance. Simulation results on the 0-9 digit-voice recognition problem are presented. Instead of using the HMMs, an alternative approach is to use the fuzzy neural network advanced in [8] to learn and recognize phoneme response sequences of 0-9 digit voices in the form of a 2-dimensional pattern recognition problem [9].
10 digit-voices from a new Group D of 8 different persons consisting of 5 male and 3 female speakers, the speaker-independent recognition result obtained by the HMMs is given in Table 3.

5 Summary

In this study, the 0-9 digit-voice recognition problem has been expressed in the form of phoneme response sequence recognition through the use of the HMMs. The methodology is based on the use of SOM and LVQ for organizing the phoneme maps, the use of the Viterbi search algorithm to determine the optimal phoneme response sequences, and the HMMs for determining the most-likely voice utterance out of 0-9 digits.

References


