Chapter IX

Applications of Kernel Theory to Speech Recognition

Joseph Picone, Mississippi State University, USA
Aravind Ganapathiraju, Mississippi State University, USA
Jon Hamaker, Mississippi State University, USA

Abstract

Automated speech recognition is traditionally defined as the process of converting an audio signal into a sequence of words. Over the past 30 years, simplistic techniques based on the design of smart feature-extraction algorithms and physiological models have given way to powerful statistical methods based on generative models. Such approaches suffer from three basic problems: discrimination, generalization, and sparsity. In the last decade, the field of machine learning has grown tremendously, generating many promising new approaches to this problem based on principles of discrimination. These techniques, though powerful when given vast amounts of training data, often suffer from poor generalization. In this chapter, we present a unified framework in which both generative and discriminative models are motivated from an information theoretic perspective. We introduce the modern statistical approach to speech recognition and discuss how kernel-based methods are used to model knowledge at each level of the problem. Specific methods discussed include kernel PCA for feature extraction and support vector machines for discriminative modeling. We conclude with some emerging research on the use of kernels in language modeling.
Introduction

The goal of a speech-recognition system is to provide an accurate and efficient means of converting an audio signal to text typically consisting of a string of words. The audio signal is often sampled at a rate between 8 and 16 kHz. The signal is converted to a sequence of vectors, known as features, at a rate of 100 times per second. These features, denoted by \( O = \{O_1, O_2, ..., O_k\} \), are referred to as observations. The observations are then ultimately mapped to a sequence of words, denoted by \( W = \{W_1, W_2, ..., W_l\} \), by integrating knowledge of human language into a statistical modeling framework. The dominant approach to achieving this signal-to-symbol conversion is based on hidden Markov models (HMMs; Jelinek, 1998; Rabiner & Juang, 1993). A speech-recognition system today is typically just one component in an information-retrieval system that can perform a wide range of human-computer interactions including voice mining, dialog, and question answering (Maybury, 2005). Historically, speech recognition has focused on maximizing the probability of a correct word sequence given the observations, denoted by \( P(W|O) \), using generative models. However, in this chapter, we will explore a relatively new class of machines that attempt to directly minimize the error rate using principles of discrimination.

There are several subtle aspects of this problem that make it a challenging machine learning problem. First, our goal is to produce a machine that is independent of the identity of the speaker or the acoustic environment in which the system operates. This requires a learning machine to infer characteristics of the signal that are invariant to changes in the speaker or channel, a problem often described as robustness (Hansen, 1994). Second, the duration of a word can vary in length even for the same speaker, which requires a learning machine to be able to perform statistical comparisons of patterns of unequal length. Third, the pronunciation of a word, which is often represented as a sequence of fundamental sound units referred to as phones, can vary significantly based on factors such as linguistic context, dialect, and speaking style (Jurafsky & Martin, 2000). Fourth, and perhaps most importantly, state-of-the-art speech-recognition systems must learn from errorful transcriptions of the words. Systems are typically trained from transcriptions that contain only the words spoken rather than detailed phonetic transcriptions, and often these word transcriptions have error rates ranging from 1% to 10%. Practical considerations such as these often require careful engineering of any learning machine before state-of-the-art performance can be achieved. Nevertheless, in this chapter, we will focus primarily on the core machine learning aspects of this problem.

An Information Theoretic Basis for Speech Recognition

Given an observation sequence, \( O \), a speech recognizer should choose a word sequence such that there is minimal uncertainty about the correct answer (Valtchev, 1995; Vertanen, 2004). This is equivalent to minimizing the conditional entropy:
20 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:

www.igi-global.com/chapter/applications-kernel-theory-speech-recognition/24825?camid=4v1

This title is available in InfoSci-Books, InfoSci-Medical, Communications, Social Science, and Healthcare, Computer Vision and Image Processing. Recommend this product to your librarian:

www.igi-global.com/e-resources/library-recommendation/?id=1

Related Content

A New Image Distortion Measure Based on Natural Scene Statistics Modeling

www.igi-global.com/article/new-image-distortion-measure-based/68001?camid=4v1a

Counting People Using Blobs and Contours
Shafraz Subdurally, Devin Dya and Sameerchand Pudaruth (2013). International Journal of Computer Vision and Image Processing (pp. 1-16).

www.igi-global.com/article/counting-people-using-blobs-and-contours/87247?camid=4v1a

Probabilistic Modeling for Detection and Gender Classification

www.igi-global.com/article/probabilistic-modeling-for-detection-and-gender-classification/111474?camid=4v1a
Both Hands’ Fingers’ Angle Calculation from Live Video
International Journal of Computer Vision and Image Processing (pp. 1-11).
www.igi-global.com/article/both-hands-fingers-angle-calculation/72312?camid=4v1a