Chapter 2.13
Building Sequence Kernels for Speaker Verification and Word Recognition

Vincent Wan
University of Sheffield, UK

ABSTRACT

This chapter describes the adaptation and application of kernel methods for speech processing. It is divided into two sections dealing with speaker verification and isolated-word speech recognition applications. Significant advances in kernel methods have been realised in the field of speaker verification, particularly relating to the direct scoring of variable-length speech utterances by sequence kernel SVMs. The improvements are so substantial that most state-of-the-art speaker recognition systems now incorporate SVMs. We describe the architecture of some of these sequence kernels. Speech recognition presents additional challenges to kernel methods and their application in this area is not as straightforward as for speaker verification. We describe a sequence kernel that uses dynamic time warping to capture temporal information within the kernel directly. The formulation also extends the standard dynamic time-warping algorithm by enabling the dynamic alignment to be computed in a high-dimensional space induced by a kernel function. This kernel is shown to work well in an application for recognising low-intelligibility speech of severely dysarthric individuals.

INTRODUCTION

In recent years, support vector machines (SVMs) have become an important tool in speaker verification and speech recognition. This chapter describes the development of sequence kernels in these domains. The following paragraphs introduce the speaker verification techniques established prior to the advent of sequence kernels. We then assess the impact of sequence kernels in speaker and speech recognition.

In text-independent speaker verification, the aim is to determine from a sample of speech

whether a person's asserted identity is true or false; this constitutes a binary classification task well suited to SVM discriminative training and generalisation. The robust classification of speech signals of variable duration comprises one of the principal challenges in this area. Classifiers such as *multilayer perceptrons*, *Gaussian mixture models* (GMMs), and *vector quantisers* do not process variable-length sequences directly. Traditionally, speaker verification applications depended on modeling the distribution of cepstral input vectors (e.g., mel frequency cepstral coefficients) using GMMs; variable-length sequence scoring was achieved by computing the average log likelihood score of the input vectors over the length of the test utterance.

The GMM (see Bishop, 1995) is a well-known modeling technique that was applied to speaker verification by Reynolds (1992). Let $f(x_A)$ denote the score for an utterance of speech x_A that is represented as a sequence of L frames $x_A=\{x_{A1}, \ldots, x_{AL}\}$ where x_{Ai} is a vector of cepstral features and A enumerates the utterances. In speaker verification, each frame x_{Ai} is scored separately by the GMM of the asserted speaker, and the utterance score is the mean of the frame log likelihood scores:

$$f(x_A) = \frac{1}{L}\sum_{i=1}^{L} \log P(x_{Ai} \mid M_{\text{ml}}), \qquad (1)$$

where M_{ml} is the model of the asserted speaker created using the maximum likelihood criterion. If $f(x_A)$ is greater than a predetermined threshold, then the speaker's asserted identity is confirmed. An improvement on this approach incorporates a (Gaussian mixture) universal background model (UBM), U, which is trained on a large number of speakers. The improved scores are the ratio of the speaker model's likelihood to the UBM's likelihood:

$$f(x_A) = \frac{1}{L}\sum_{i=1}^{L} \log P(x_{Ai} \mid M_{\text{ml}}) - \log P(x_{Ai} \mid U). \qquad (2)$$

A further refinement replaces M_{ml} with a better model M_{ad} created by adapting U to the speaker using Maximum a Posteriori Probability (MAP) adaptation (Reynolds, 1995):

$$f(x_A) = \frac{1}{L}\sum_{i=1}^{L} \log P(x_{Ai} \mid M_{\text{ad}}) - \log P(x_{Ai} \mid U). \qquad (3)$$

Early SVM approaches by Schmidt and Gish (1996) and then by Wan and Campbell (2000) replaced the GMM estimate of the log likelihood in equation (1) with the raw output of polynomial and Radial Basis Function (RBF) kernel SVMs. The success of this approach, however, was limited since it proved difficult to train SVMs on a large set of cepstral input vectors: Using all available data resulted in optimisation problems requiring significant computational resources that were unsustainable. Employing clustering algorithms to reduce the data also, unfortunately, reduced the accuracy. Furthermore, this type of training ignores the concept of an utterance: This is important since discriminative training may discard information that is not considered useful for classification, thus training discriminatively on the (frame-level) input vectors may inadvertently discard information that could prove useful for classifying the sequence.

The solution was to map each complete utterance onto one point in the *feature space* using sequence kernels (Campbell, 2002; Fine, Navratil, & Gopinath, 2001; Wan & Renals, 2002) described in the next section. They enable SVMs to classify variable-length sequences directly, simultaneously incorporating the notion of an utterance, enabling sequence-level discrimination, and effectively reducing the number of input vectors (now sequence-level input vectors) by several orders of magnitude compared to the frame-level approach. Sequence kernels have been so effective in reducing error rates that they are now incorporated into many state-of-the-art speaker verification systems.