Chapter 82
Comparison of Tied-Mixture and State-Clustered HMMs with Respect to Recognition Performance and Training Method

Hiroyuki Segi
NHK (Japan Broadcasting Corporation) Science & Technology Research Laboratories, Japan

Shoei Sato
NHK (Japan Broadcasting Corporation) Science & Technology Research Laboratories, Japan

Kazuo Onoe
NHK (Japan Broadcasting Corporation), Science & Technology Research Laboratories, Japan

Akio Kobayashi
NHK (Japan Broadcasting Corporation) Science & Technology Research Laboratories, Japan

Akio Ando
University of Toyoma, Japan

ABSTRACT

Tied-mixture HMMs have been proposed as the acoustic model for large-vocabulary continuous speech recognition and have yielded promising results. They share base-distribution and provide more flexibility in choosing the degree of tying than state-clustered HMMs. However, it is unclear which acoustic models to superior to the other under the same training data. Moreover, LBG algorithm and EM algorithm, which are the usual training methods for HMMs, have not been compared. Therefore in this paper, the recognition performance of the respective HMMs and the respective training methods are compared under the same condition. It was found that the number of parameters and the word error rate for both HMMs are equivalent when the number of codebooks is sufficiently large. It was also found that training method using the LBG algorithm achieves a 90% reduction in training time compared to training method using the EM algorithm, without degradation of recognition accuracy.

DOI: 10.4018/978-1-5225-1759-7.ch082

1. INTRODUCTION

Speech-recognition systems are of particular interest in Japan because real-time keyboard entry in the Japanese language is complicated by the need to select the correct characters among homonyms.

Remarkable advances have been made in speech-recognition technology in recent years. One example is the simultaneous subtitling system for Japanese television broadcast programs developed by Kobayashi (2013), which uses speech recognition to make real-time captions for use by the hearing impaired. Another example is the transcription system using speech recognition developed by Kawahara (2012) that is currently deployed in the Japanese Parliament.

These systems employ Hidden Markov Models (HMMs) that were proposed long time ago and many speech recognition systems are still using HMMs even now (Hofmann, 2012; Liu, 2013; Ogawa, 2012; Singh, 2012; Siu, 2012). The continuing development of large-scale speech databases has made it possible to use large amounts of data to train HMMs (Itou, 1998; Maekawa, 2000; Segi, 2010). As the volume of data increases, it is possible to increase the number of parameters without losing the estimation accuracy, and highly accurate speech recognition can be realized by introducing more complex structures in HMMs. For example, a state-clustered HMM (Hwang, 1996; Onoe, 2003; Young, 1994) has been proposed in which base (usually Gaussian) distributions and weights are shared within individual clusters. In addition, a tied-mixture HMM (Nguyen, 1995; Sankar, 1998; Lee, 2000), in which base distributions and weights can be shared separately, has been reported to produce favorable results.

However, the performance of different acoustic models has not yet been compared using the same training data. Moreover, the established training methods for HMMs, the Linde-Buzo-Gray (LBG) algorithm (Linde, 1980) and the Expectations-Maximization (EM) algorithm (Dempster, 1977), have not been compared. Although these training methods are proposed long time ago and discriminative training (Povey, 2002) is used in recent years, these training methods are used now to make initial models for discriminative training (Delcroix, 2013).

The current paper compares the recognition performance of the respective HMMs and training methods under the same conditions. Section 2 describes previous comparisons of state-clustered and tied-mixture HMMs. Section 3 describes the differences between state-clustered and tied-mixture HMMs. Section 4 describes HMM training methods. Section 5 discusses recognition tests as follows: section 5.1 describes test conditions; section 5.2 compares training methods for state-clustered and tied-mixture HMMs; section 5.3 compares recognition performance and processing speeds with different numbers of tied-mixture HMM weights; section 5.4 compares state-clustered and tied-mixture HMMs with different numbers of base distributions included in the codebook; and section 5.5 compares state-clustered and tied-mixture HMMs with the same overall number of base distributions but different codebook sizes. Section 6 summarizes this work.

2. PREVIOUS WORK

State-clustered and tied-mixture HMMs have been compared in previous studies. Huang (1993) showed that a tied-mixture HMM had better recognition performance than a state-clustered HMM, although the training method differed between the two. Sankar (1998) compared a state-clustered HMM (with 937 clustered states and 32 base distributions per state) with a tied-mixture HMM (with 38 codebooks and 789 base distributions per codebook), and found that when the overall numbers of base distributions