Speech Communication 9 (1990) 119-128
North-Holland

DURATION IN CONTEXT CLUSTERING FOR SPEECH RECOGNITION

Joseph PICONE
Speech and Image Understanding Laboratory, Texas Instruments, Inc., Dallas, TX 75265, U.S.A.

Received 25 July 1989
Revised 17 January 1990

Abstract. Hierarchical clustering techniques have been shown to be a powerful tool in building speaker-independent reference models for Dynamic Time Warping (DTW) based speech recognition systems. In this paper, we introduce a clustering algorithm based on the standard KMEANS procedure that generates reference models for continuous density Hidden Markov Model (HMM) based systems by simultaneously considering spectral and duration information. Improved speech recognition performance using clustering is demonstrated on a digit recognition task using the TI/NBS studio quality connected digit database.

Zusammenfassung. Es ist bewiesen worden, daß die hierarchische Clusterbildung eine leistungsfähige Technik darstellt, um sprecherunabhängige Bezugsmodelle zu erstellen für Spracherkennungssysteme welche auf der Methode der dynamischen Verzerrung der Zeitachse (DTW) beruhen. In diesem Beitrag stellen wir einen Algorithmus zur Clusterbildung vor, welcher Bezugsmodelle erstellt für Systeme welche Markovmodelle (HMM) verwenden. Die Methode welche auf dem herkömmlichen KMEANS verfahren basiert, berücksichtigt sowohl spektrale wie zeitliche Daten. Es wird gezeigt, anhand einer Ziffererkennungsaufgabe, daß Clusterbildung die Erkennungsgenauigkeit verbessert.

Résumé. Il a été montré que des techniques d'analyse typologique hiérarchique constituent des outils puissants pour construire des références indépendantes du locuteur dans les systèmes de reconnaissance de la parole basés sur des méthodes de distorsion dynamique de l'échelle temporelle (DTW). Dans cet article, nous proposons un algorithme d'analyse typologique qui génère, en considérant simultanément le spectre et la durée, des patrons de références pour des systèmes basés sur des modèles markoviens (HMM) à densité continue. De meilleures performances de reconnaissance sont ainsi démontrées dans le cadre d'une tâche de reconnaissance de chiffres utilisant la base de données TI/NBS de chiffres connectés.

Keywords. Automatic speech recognition, speaker-independent continuous speech recognition, digit recognition, clustering.

1. Introduction

The performance of continuous density Hidden Markov Model (HMM) based speech recognition systems is strongly dependent on the choice of a good set of initial conditions, and on a sufficiently large training database. While re-estimation techniques, such as the Baum–Welch algorithm (Baum and Egan, 1963; Baum et al., 1970), have proven to be reliable methods to generate improved recognition models, these classes of training algorithms will not overcome fundamentally flawed HMM seed models. Good reference models are generated by developing good seed models. Seed model construction for continuous density HMMs has previously been an art. Appropriate hand-seeding of reference models often requires an extensive knowledge of the recognition unit, the application vocabulary, and even details of the application (such as the user interface).

The introduction of Markov modeling to speech recognition has resulted in a significant increase in recognition accuracy. One of the primary reasons for this increase relative to comparable Dynamic Time Warping (DTW) systems is that HMM technology generally does a better job of segmentation of the data. This is a result of the supervised training procedure (also referred to as HMM re-estimation) in which model optimization