NONLINEAR DISCRIMINANT FEATURE EXTRACTION FOR ROBUST TEXT-INDEPENDENT SPEAKER RECOGNITION

Yochai Konig, Larry Heck, Mitch Weintraub, and Kemal Sonmez

Speech Technology and Research Laboratory
SRI International
Menlo Park, CA 94025

RÉSUMÉ

Cet article propose une méthode basée sur l'analyse discriminative non-linéaire pour extraire et sélectionner un ensemble de vecteurs acoustiques utilisés pour l'identification de locuteurs. L'approche consiste à mesurer et grouper un grand nombre de mesures acoustiques (correspondant à plusieurs trames de données consécutives), et à réduire la dimensionalité du vecteur résultant au moyen d'un reseau de neurones artificielles. Le critère utilisé pour optimiser les poids du réseau consiste à maximiser une mesure de la séparation entre les locuteurs d'une base de données d'apprentissage. L'architecture du reseau est telle que l'une de ses couches intermédiaires représente la projection des vecteurs acoustiques d'entrée sur un espace de dimensionalité inferieure. Après la phase d'apprentissage, cette partie du réseau peut etre isolée et utilisée pour projeter les vecteurs acoustiques d'une base de données de test. Les vecteurs acoustiques projetés peuvent alors être classifiés. Combiné à un classificateur cepstral, le classificateur utilisant ces nouveaux vecteurs acoustiques réduit de 15% le taux d'erreur de classification de la base de données définie par NIST en 1997 pour l'évaluation des systèmes de reconnaissance du locuteur.

ABSTRACT

We study a nonlinear discriminant analysis (NLDA) technique that extracts a speaker-discriminant feature set. Our approach is to train a multilayer perceptron (MLP) to maximize the separation between speakers by nonlinearly projecting a large set of acoustic features (e.g., several frames) to a lower-dimensional feature set. The extracted features are optimized to discriminate between speakers and to be robust to mismatched training and testing conditions. We train the MLP on a development set and apply it to the training and testing utterances. Our results show that by combining the NLDA-based system with a state of the art cepstrum-based system we improve the speaker verification performance on the 1997 NIST Speaker Recognition Evaluation set by 15% in average compared with our cepstrum-only system.

1. INTRODUCTION

Our goal is to extract and select features that are more invariant to non-speaker-related conditions such as handset type, sentence content, and channel effects. Such features will be robust to mismatched training and testing conditions of speaker verification systems. With current feature sets (e.g., cepstrum) there is a big performance gap between matched and mismatched tests [8] even after applying standard channel compensation techniques [4]. In order to find these features, the feature extraction step should be directly optimized to increase discrimination between speakers, and to filter out the non-relevant information.

Our proposed solution is to train a multilayer perceptron (MLP) to nonlinearly project a large set of acoustic features to a lower-dimensional feature set, such that it maximizes speaker separation. We train the MLP on a development set that includes several realizations of the same speakers under different conditions. We then apply the learned transformation (MLP in feed-forward mode) to the training and testing utterances. Finally, we use the resulting features for training the speaker recognition system, e.g., Bayesian adapted Gaussian mixture system [9].

We begin by reviewing related studies in Section 2. We describe the proposed feature extraction technique in Section 3. The Development database is described in Section 4. In Section 5, we report the experimental results on the 1997 NIST evaluation set. We continue with analysis of the results in Section 6. Finally, we conclude and describe directions for future work in Section 7.

2. RELATED STUDIES

The related studies to the NLDA technique can be divided into two main categories: robust speaker verification systems, and data-driven feature extraction techniques. Previously proposed approaches to increase robustness to mismatched training and testing conditions, especially to handset variations, include handset-dependent background