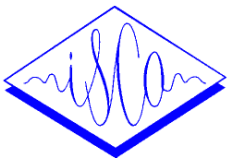

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

DISTANCE-BASED GAUSSIAN MIXTURE MODEL FOR SPEAKER RECOGNITION OVER THE TELEPHONE

R. D. Zilca
Research and Development Division
Amdocs Israel
8 Hapnina St. Raanana, ISRAEL
ranzilca@ieee.org

Y. Bistritz
Department of Electrical Engineering
Tel Aviv University
Tel Aviv 69978, ISRAEL
bistritz@eng.tau.ac.il

ABSTRACT

The paper considers text independent speaker identification over the telephone using short training and testing data. Gaussian Mixture Modeling (GMM) is used in the testing phase, but the parameters of the model are taken from clusters obtained for the training data by an adequate choice of feature vectors and a distance measure without optimization in the maximum likelihood (ML) sense. This distance-based GMM (DB-GMM) approach was evaluated by experiments in speaker identification from short telephone-speech data for a few feature vectors and distance measures. The selected feature vectors were Line Spectra Pairs (LSP) and Mel Frequency Cepstra (MFC). The selected distance measures were weighted Euclidean distance with IHM and BPL, respectively. DB-GMM showed consistently better performance than GMM trained by the expectation-maximization (EM) algorithm. Another notable observation is that a full covariance GMM (that is more comfortably trained by DB-GMM) always achieved significantly better performance than diagonal covariance GMM.

1. INTRODUCTION

Gaussian Mixture Modeling (GMM) provides a good approach to text-independent speaker recognition [1]. The testing phase is done by Maximum-Likelihood (ML) classification; therefore the trained parameters should be optimized in the ML sense. To meet this requirement, the Expectation-Maximization (EM) algorithm [2] was proposed to estimate the parameters of the Gaussian mixture probability density function in the training process [3]. However, in many practical situations too little available data causes the EM iterations to become ill conditioned or over fitted to the trained data. The problem is partly alleviated by limiting the model to diagonal covariance matrices rather than full covariance matrices.

The approach that we call Distance-Based GMM proposes to obtain parameters for the GMM model without optimization in the ML sense. Instead, the relative population, centers and spread of clusters of ("well chosen") feature vectors obtained from the training data by a ("well chosen") distance measure are used as the weights, average, and covariance (respectively) of the GMM. A VQ trained GMM was applied before to speaker verification from clean speech in [4] and shown to achieve almost the same speaker verification as a GMM trained by the EM algorithm. The somewhat heuristic separation of the principle of optimality used at the training and the testing processes may be justified in cases when it is not necessarily clear that the EM algorithm converges to the GMM that provide the best speaker recognition performance. For instance, when recognition is requested from a short duration of training data, the EM iterations may become ill conditioned and not converge to a numerically robust (especially a full covariance) GMM. Also, in a realistic application, when the speaker's voice is available via a communication channel, the training by EM iterations may reduce recognition rate by over-fitting the GMM to the specifics of the channel. Therefore, a DB-GMM not only guarantees a simpler training than the EM algorithm but it may also provide competing performance to EM-GMM for recognition of speakers over the telephone (and other communication channels) using short training data.

The paper reports some experiments in clos