Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 23 (2009) 165–175
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Evolutionary minimization of the Rand index for speaker clustering

Wei-Ho Tsai [a,*], Hsin-Min Wang [b,1]

[a] Department of Electronic Engineering and Graduate Institute of Computer and Communication Engineering, National Taipei University of Technology, Taipei 10608, Taiwan
[b] Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan

Received 15 July 2007; received in revised form 9 May 2008; accepted 9 May 2008
Available online 17 May 2008

Abstract

We propose an effective method for clustering unknown speech utterances based on their associated speakers. The method jointly optimizes the generated clusters and the required number of clusters by estimating and minimizing the Rand index. The metric reflects the clustering errors that arise when utterances from the same speaker are placed in different clusters; or when utterances from different speakers are placed in the same cluster. One useful characteristic of the Rand index is that its value only reaches the minimum when the number of clusters is equal to the size of the true speaker population. We approximate the Rand index by a function of the similarity measures between utterances and then use a genetic algorithm to determine the cluster in which each utterance should be located, such that the function is minimized. Our experiment results show that this novel speaker-clustering method outperforms conventional methods that use the Bayesian information criterion to determine the required number of clusters.

Keywords: Genetic algorithm; Rand index; Speaker clustering

1. Introduction

Motivated by the need for effective methods to index and archive the burgeoning amount of spoken data being generated universally, recent research on automatic classification of speech samples based on speakers' voice characteristics has been extended from the traditional supervised problem of speaker identification/verification (Campbell, 1997) to an unsupervised paradigm (Makhoul et al., 2000). Basically, the paradigm involves two tasks: segmenting an audio recording into speech utterances that contain only one speaker's voice (Siegler et al., 1997; Johnson, 1999; Zhou and Hansen, 2000), and grouping utterances from the same speaker

* Corresponding author. Tel.: +886 2 27712171x2257; fax: +886 2 27317120.
E-mail addresses: whtsai@ntut.edu.tw (W.-H. Tsai), whm@iis.sinica.edu.tw (H.-M. Wang).
[1] Tel.: +886 2 27883799x1714; fax: +886 2 27824814.

0885-2308/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2008.05.001

into one cluster (Gish et al., 1991; Jin et al., 1997; Solomonoff et al., 1998; Chen and Gopalakrishnan, 1998; Reynolds et al., 1998). The tasks can be addressed jointly by a process called *speaker diarization* (Tranter and Reynolds, 2006; Ben et al., 2004; Tranter, 2005; Zhu et al., 2005; Sinha et al., 2005). It is hoped that, by locating utterances from the same speaker, the human effort required to index speech data can be greatly reduced, i.e., from having to listen to every audio recording to only checking a few utterances in each cluster. In this paper, we concentrate on the latter problem, referred to as *speaker clustering*. Assume that we have a collection of N speech utterances, each of which is from one of P unknown speakers, where $N \geqslant P$, and P is also unknown. The aim of speaker clustering is to partition the N utterances into M clusters such that $M = P$ and each cluster only contains utterances from one speaker.

Since no prior information regarding the speakers involved and the speaker population size is available in most practical applications, a common strategy used to solve the speaker-clustering problem involves three steps: characterizing the voice similarities between utterances, generating clusters based on those similarities, and determining the optimal number of clusters. The most popular speaker-clustering method employs hierarchical agglomerative clustering (HAC) (Gish et al., 1991; Jin et al., 1997; Solomonoff et al., 1998; Chen and Gopalakrishnan, 1998; Reynolds et al., 1998; Johnson and Woodland, 1998; Faltlhauser and Ruske, 2001; Ajmera et al., 2002; Moh et al., 2003; Liu, 2005). This approach generates a cluster tree by sequentially merging the utterances deemed similar to each other. Then, the tree is cut using the Bayesian information criterion (BIC) (Schwarz, 1978; Chen and Gopalakrishnan, 1998; Zhou and Hansen, 2000) to retain the appropriate number of clusters. Although various modifications to the method have been proposed (Ben et al., 2004; Tranter, 2005; Zhu et al., 2005; Sinha et al., 2005), most of them focus on improving the combination of speaker clustering and speaker segmentation in a speaker diarization system. There is a dearth of research on improving the performance of speaker clustering per se.

As noted in our previous works (Tsai and Wang, 2005, 2006), one major drawback of most speaker-clustering systems is the problem of error propagation in HAC. Specifically, although HAC merges the most similar utterances sequentially, it is possible that, in some merging operations, utterances by different speakers may be mis-grouped into the same cluster. In such cases, the utterances cannot be separated in subsequent merging operations; hence, the mis-grouping errors will proliferate as more clusters are merged. To resolve this problem, we proposed clustering methods that maximize the within-cluster homogeneity of speakers' voice characteristics by jointly considering all the clusters to be generated, instead of by the cluster-by-cluster technique used in HAC. However, like most speaker-clustering systems, our approach followed the principle of BIC-based methods by determining the optimal number of clusters after completion of the cluster generation process. Since the back-end determination of the optimal number of clusters trusts the front-end cluster generation process completely, the inevitable errors generated by the front-end can propagate to the back-end, which may lead to inaccurate estimation of the speaker population size.

To overcome the above-mentioned limitations, in this paper, we propose a new clustering method that simultaneously optimizes the generated clusters and the required number of clusters by estimating and then minimizing the Rand index (Rand, 1971; Hubert and Arabie, 1985; Solomonoff et al., 1998). The metric indicates clustering errors that place utterances from the same speaker in different clusters, or place utterances from different speakers in the same cluster. A useful characteristic of the Rand index is that its value only reaches the minimum when the number of clusters is equal to the true size of the speaker population. We approximate the Rand index by a function of the similarity measures between utterances, and use a genetic algorithm (Goldberg, 1989) to determine the cluster in which each utterance should be located, such that the function is minimized. The resulting clusters are thus optimized in a global fashion, rather than in the pair-by-pair manner used in HAC-based methods. Furthermore, the number of clusters derived by minimizing the approximated Rand index naturally reflects the speaker population size.

The remainder of the paper is organized as follows. In Section 2, we explain our motivation for studying the problem of speaker clustering and describe the performance assessment method used in this study. Section 3 introduces the proposed method for estimating and minimizing the Rand index, whereby the resulting partition of utterances approaches an optimal state in terms of within-cluster homogeneity and the number of clusters. Section 4 details our experiment results. Then, in Section 5, we present our conclusions and indicate the direction of our future work.