Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 40 (2016) 23–45

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

A study of speaker clustering for speaker attribution in large telephone conversation datasets

Houman Ghaemmaghami [a,*], David Dean [a], Sridha Sridharan [a], David A. van Leeuwen [b]

[a] *Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia*
[b] *Center for Language and Speech Technology, Radboud University Nijmegen, Netherlands*

Received 19 June 2015; received in revised form 17 November 2015; accepted 28 March 2016
Available online 9 May 2016

Abstract

This paper proposes the task of speaker attribution as speaker diarization followed by speaker linking. The aim of attribution is to identify and label common speakers across multiple recordings. To do this, it is necessary to first carry out diarization to obtain speaker-homogeneous segments from each recording. Speaker linking can then be conducted to link common speaker identities across multiple inter-session recordings. This process can be extremely inefficient using the traditional agglomerative cluster merging and retraining commonly employed in diarization. We thus propose an attribution system using complete-linkage clustering (CLC) without model retraining. We show that on top of the efficiency gained through elimination of the retraining phase, greater accuracy is achieved by utilizing the farthest-neighbor criterion inherent to CLC for both diarization and linking. We first evaluate the use of CLC against an agglomerative clustering (AC) without retraining approach, traditional agglomerative clustering with retraining (ACR) and single-linkage clustering (SLC) for speaker linking. We show that CLC provides a relative improvement of 20%, 29% and 39% in attribution error rate (AER) over the three said approaches, respectively. We then propose a diarization system using CLC and show that it outperforms AC, ACR and SLC with relative improvements of 32%, 50% and 70% in diarization error rate (DER), respectively. In our work, we employ the cross-likelihood ratio (CLR) as the model comparison metric for clustering and investigate its robustness as a stopping criterion for attribution.

© 2016 Elsevier Ltd. All rights reserved.

Keywords: Speaker attribution; Linking; Diarization; Complete-linkage clustering; Joint factor analysis; Cross-likelihood ratio

1. Introduction

The increasing number of spoken audio archives around the world has brought about a need for technologies capable of automatically annotating and indexing large datasets of this type, with respect to speaker identity. Speaker diarization is vital to this task and can reveal *'Who spoke when?'* in a given recording (Chen and Gopalakrishnan, 1998). A speaker diarization system can be used in a variety of applications, such as automatic speech recognition (ASR), surveillance, forensics and information retrieval. For example, in ASR applications, diarization can be used to provide system speaker labels along with output text transcriptions (Kenny, 2005). In addition, it can be used prior to ASR to obtain

[*] Corresponding author at: Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia. Tel.: +61 7 3138 2113.

E-mail address: houman.ghaemmaghami@qut.edu.au (H. Ghaemmaghami).

http://dx.doi.org/10.1016/j.csl.2016.03.005
0885-2308/© 2016 Elsevier Ltd. All rights reserved.

speaker information that may be employed to enhance speech transcriptions through speaker adaptation of ASR models. In surveillance applications, speaker diarization can play a vital role in processing of audio recordings by eliminating the need for a person to manually annotate audio records. In forensic applications, it can be employed for speaker identification tasks. In the broadcast industry, it can be used to facilitate the manual transcription of subtitles, or for information retrieval applications, allowing for automatic indexing of individual spoken audio documents and enabling the end user to browse an audio document by a specific speaker identity (Barras et al., 2006), for example, that of a reporter.

Viet-Anh et al. (2011) and Yang et al. (2011) proposed extending diarization to process multiple recordings and referred to this process as cross-show speaker diarization. Speaker diarization is traditionally applied within a recording (Barras et al., 2006; Wooters and Huijbregts, 2008), and is not typically designed to identifying speakers across temporally-independent recordings. These recordings can contain inconsistencies in their environments or the voice of speakers due to aging or health complications. For this reason, we employ the term *speaker linking*, as first used by van Leeuwen (2010) and adopted by others (Bourlard et al., 2013; Ferras and Bourlard, 2012; Ghaemmaghami et al., 2013; Vaquero et al., 2011), to refer to the task of determining speaker identities across independent recordings. We employ speaker diarization and speaker linking to annotate an input dataset and refer to the combination of the two tasks as *speaker attribution* (Ghaemmaghami et al., 2011, 2013).

Speaker attribution refers to the combined tasks of speaker diarization and speaker linking. It is used to annotate recordings within a spoken audio archive, through determining instances of the same speakers within and across multiple recordings (Ghaemmaghami et al., 2012). To conduct attribution on a set of audio files, speaker diarization must first be applied to the individual recordings in order to obtain speaker-homogeneous segments for each file (Ghaemmaghami et al., 2012). This requires a first stage, referred to as speaker change detection or speaker segmentation (Chen and Gopalakrishnan, 1998; Ghaemmaghami et al., 2015; Moraru et al., 2003), which is carried out to obtain segments of audio that are hypothesized to have speech from only one speaker identity. This is often a difficult task and can severely impact the proceeding stages of diarization if impure segments are selected. These segments are then clustered to match parts of the audio that are produced by the same speaker within a recording (Morgan and Bourlard, 1995; Mori and Nakagawa, 2001). This speaker segment clustering stage can also impact overall system performance and requires accurate speaker modeling of short speech durations and efficient clustering of these segments. The clusters of segments obtained from various recordings, and ideally belonging to unique speakers, are then linked across multiple inter-session recordings within an audio archive using speaker linking (Ghaemmaghami et al., 2011; van Leeuwen, 2010). As the size of the analyzed audio archive increases, so too does the demand for greater efficiency in conducting attribution. It is thus necessary to carry out the diarization and linking tasks in a more efficient manner and without sacrificing accuracy.

Recent techniques in diarization draw heavily from state-of-the-art speaker recognition approaches for conducting speaker modeling and comparison. Some commonly utilized techniques that have successfully been applied to diarization include maximum a posteriori (MAP) adaptation using a Gaussian mixture model (GMM) universal background model (UBM) (Reynolds et al., 2000), joint factor analysis (JFA) with session compensation and speaker variability modeling, as proposed by Kenny et al. (2010), and probabilistic linear discriminant analysis (PLDA) modeling in the i-vector space (Vaquero et al., 2011). It is also necessary to compare speaker models; thus, various scoring techniques, such as the cosine distance (Vaquero et al., 2011), cross-likelihood ratio (CLR) (Barras et al., 2006; Meignier et al., 2006) and Bayes' factor (Wang et al., 2010), have been proposed and commonly used for diarization. Other methods, such as variational Bayesian learning, have been utilized for diarization with promising results but have not been as popular as the JFA or i-vector approach (Valente et al., 2010). These techniques have provided the means for carrying out efficient and reliable speaker modeling; however, clustering of speaker models in diarization is commonly carried out using the traditional method of agglomerative merging with retraining (Viet Bac and Fohr, 2007; Wooters and Huijbregts, 2008). This approach is not practical and becomes more inefficient as the length of the analyzed recording increases. In addition, the *hard* many-to-one decision applied through the retraining stage, following an incorrect cluster merge, can bring about irreversible speaker errors that are carried through the entire diarization process.

Speaker linking was first proposed by van Leeuwen (2010) for processing large spoken datasets, with the aim of linking together speaker-homogeneous segments based on speaker identity. In van Leeuwen's study, the inefficiency issues associated with agglomerative cluster merging and retraining were overcome through a method of eliminating the retraining phase of the agglomerative clustering approach. This resulted in greater efficiency; however, it was not

Download English Version:

https://daneshyari.com/en/article/6951550

Download Persian Version:

https://daneshyari.com/article/6951550

Daneshyari.com