Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 1132–1142

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Recognising speakers from the topics they talk about

Doris Baum

Fraunhofer IAIS, St. Augustin, Germany

Received 15 April 2011; received in revised form 27 April 2012; accepted 12 June 2012
Available online 19 June 2012

Abstract

We investigate how a speaker's preference for specific topics can be used for speaker identification. In domains like broadcast news or parliamentary speeches, speakers have a field of expertise they are associated with. We explore how topic information for a segment of speech, extracted from an automatic speech recognition transcript, can be employed to identify the speaker. Two methods for modelling topic preferences are compared: implicitly, based on speaker-characteristic keywords, and explicitly, by using automatically derived topic models to assign topics to the speech segments. In the keyword-based approach, the segments' tf-idf vectors are classified with Support Vector Machine speaker models. For the topic-model-based approach, a domain-specific topic model is used to represent each segment as a mixture of topics; the speakers' score is derived from the Kullback–Leibler divergence between the topic mixtures of their training data and of the segment.

The methods were tested on political speeches given in German parliament by 235 politicians. We found that topic cues do carry speaker information, as the topic-model-based system yielded an equal error rate (EER) of 16.3%. The topic-based approach combined well with a spectral baseline system, improving the EER from 8.6% for the spectral to 6.2% for the fused system.

Keywords: Speaker recognition; Topic classification; High-level features

1. Introduction

While speaker recognition systems traditionally rely on spectral features to recognise people by their voice, higher-level characteristics for capturing the manner of speaking also have been successfully implemented. Features capturing cues such as pronunciation and prosody can identify speakers and combine well with voice recognition (Kinnunen and Li, 2010; Ferrer et al., 2010; Kockmann et al., 2010; Hatch et al., 2005; Shriberg et al., 2005; Reynolds et al., 2000; Reynolds, 1995).

There are fewer studies on the use of lexical or semantic information for speaker recognition, possibly because this requires the use of automatic speech recognition (ASR) and thus the resulting system is not language independent. However, if a (manual or ASR) transcript is available for the speaker's segmentss, it is advantageous to use it for speaker identification.

For example, the speakers' names, mentioned during an introduction, may be spotted in the transcript (Canseco-Rodriguez et al., 2004; Canseco et al., 2005; Mauclair et al., 2006). When working on a domain like broadcast news, where speakers are often introduced or introduce themselves, patterns like "I am <name>" or "joining us now is <name>" can be identified and exploited for speaker recognition.

Even if the speaker's name is not mentioned, the transcript may carry idiolectal speaker information contained in frequently used words, idioms, and linguistic mannerisms. Idiolectal speaker recognition was first explored in (Doddington, 2001) with manual transcriptions: word-n-grams are used to capture characteristic phrases or phrase parts. The speaker idiosyncrasy of an n-gram is expressed by the log likelihood ratio between the frequency of the n-gram in the speaker and the background material. Subsequently,

E-mail address: dorisbaum@gmx.net

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.06.003

the approach was applied to real ASR transcripts and expanded by using Support Vector Machines (SVMs) with a linear kernel to model the speakers' n-gram counts instead of log likelihood ratio scoring (Shriberg et al., 2005). This resembles text classification with SVMs (Joachims, 1998) if the speakers are used as the categories, but differs in a number of aspects: error-prone ASR transcripts are used instead of written text, n-grams instead of single words; no stemming and stop-word-removal is applied, and a different term weighting scheme for normalising the n-gram counts is employed. (Joachims, 1998) uses term frequency-inverse document frequency (tf-idf) weighting, well known from text classification, while (Shriberg et al., 2005) employ rank normalisation (where each n-gram frequency value is replaced by its rank among the background documents).

Finally, not only the idiolect present in a speech segment but also its topic may contain speaker information. In appropriate domains, such as broadcast news, the speakers can be expected to talk about their interests or fields of expertise. A tennis player will mostly talk about sports while a minister will likely speak about politics. Thus, topic information can be used to identify speakers. In (Baum, 2009), we tested speaker-specific words as a means of recognising politicians by their fields of expertise. Like in (Doddington, 2001), log-likelihood ratio scoring was used, but the features were adapted to capture topic instead of idiolect. As is common in text categorisation, the frequencies of content word stems were used instead of word-n-grams frequencies, shifting the focus from characteristic phrases to topic marker words. However, topics are modelled only implicitly in this approach: speakers are associated with words, and the marker words for the speakers' fields of expertise will be characteristic for them.

In this paper, we explore how explicit topic modelling can be used for speaker recognition by linking the utterances' words to topics and topics to speakers. The notion of linking words to topics and topics to documents is well-known in text classification as probabilistic topic modelling. The popular Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Steyvers and Griffiths, 2007) models documents as probability distributions over topics and topics as probability distributions over words. The topics are the latent variables in this model and can be learned automatically from text corpora. This modelling has also been extended to include multiple authors for the documents (Rosen-Zvi et al., 2010). However, for our current experiments we assume that a speech segment contains only one speaker.

We describe and compare two methods for implicit and explicit usage of topic information for speaker recognition. The first, implicit, method is based on speaker-specific content keywords and uses SVM modelling (see Section 2.1). The second, explicit, employs LDA for topic modelling and assigns speakers to topics and topics to segments (see Section 2.2). For evaluation, we use a corpus of speeches from German parliament because in this domain the speakers can be expected to talk about their fields of expertise (see Section 3). The implementation and parameters of the proposed methods are detailed in Section 4, as well as the set-up of the idiolectal and spectral speaker recognition systems used as a comparison in the experiments. Section 5 discusses the experimental results obtained for the systems and their combinations. The conclusion and ideas for future work are given in Section 6.

2. Speaker recognition by topic

Speaker-specific topic preferences can be covered in various ways, especially if there is training material with topic annotations. However, this study focuses on two methods that are applicable *without* explicit topic annotations and reference transcripts of the speech segments. The text transcripts used as input by the system are generated by a large vocabulary continuous speech recognition (LVCSR) system. Both methods make use of the topic information present in the words of a segment. The first models topics only implicitly, by learning speaker specific keywords. The use of keywords is similar to (Baum, 2009), however, the modelling is different. The second method explicitly uses topics by employing unsupervised probabilistic topic models, as described in (Steyvers and Griffiths, 2007), to map words to topic clusters. With the topic models, it is possible to generate topic histograms for training and test documents. The histograms are then used as features for topic-model-based speaker recognition. Both methods use a text collection from the domain as background material, as a sufficiently large collection of speech segments from the domain may not be available.

The approaches themselves are described in the following sections while the implementation details (LVCSR system, stemmer, SVM modelling etc.) are specified in Section 4.

2.1. Keyword-based speaker recognition

Our keyword-based speaker recognition method uses weighted word frequencies as features for speaker modelling by Support Vector Machines (SVMs). It is based on text categorisation by SVMs (Joachims, 1998), where the categories are speakers. The modelling is similar to the word-n-gram SVM speaker recognition system described in (Shriberg et al., 2005), with the difference that the aim is to capture topic cues instead of idiolect, thus word stems are used instead of word n-grams. Also, tf-idf term weighting is employed instead of rank-normalisation.

2.1.1. Feature extraction

The word frequency features for speech segments are produced as depicted in Fig. 1: First, a word transcript of the segment is produced with an LVCSR system. Then, the text is pre-processed: Stop words are removed as they don't convey information about the topic at hand. The remaining words are reduced to their stem, which is considered to contain the topical information. After that, stems