Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 52 (2010) 236–245

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Speaker adaptation of language and prosodic models for automatic dialog act segmentation of speech

Jáchym Kolář [a,*], Yang Liu [b], Elizabeth Shriberg [c,d]

[a] Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic
[b] Department of Computer Science, University of Texas at Dallas, Richardson, TX, USA
[c] Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA
[d] International Computer Science Institute, Berkeley, CA, USA

Received 14 July 2009; received in revised form 20 October 2009; accepted 20 October 2009

Abstract

Speaker-dependent modeling has a long history in speech recognition, but has received less attention in speech understanding. This study explores speaker-specific modeling for the task of automatic segmentation of speech into dialog acts (DAs), using a linear combination of speaker-dependent and speaker-independent language and prosodic models. Data come from 20 frequent speakers in the ICSI meeting corpus; adaptation data per speaker ranges from 5 k to 115 k words. We compare performance for both reference transcripts and automatic speech recognition output. We find that: (1) speaker adaptation in this domain results both in a significant overall improvement and in improvements for many individual speakers, (2) the magnitude of improvement for individual speakers does not depend on the amount of adaptation data, and (3) language and prosodic models differ both in degree of improvement, and in relative benefit for specific DA classes. These results suggest important future directions for speaker-specific modeling in spoken language understanding tasks.

Keywords: Spoken language understanding; Dialog act segmentation; Speaker adaptation; Prosody modeling; Language modeling

1. Introduction

The general idea of model adaptation to a particular talker has successfully been used in the cepstral domain for speech recognition, for example by Gauvain and Lee (1994) and Gales (1998). However, less is known about speaker adaptation for spoken language *understanding*. This paper explores the question of speaker adaptation of generic models for a language understanding task. We focus on speaker-specific modeling for one spoken language understanding task, automatic dialog act (DA) segmentation of speech. This task is important since standard automatic speech recognition (ASR) systems output only a raw stream of words, leaving out important structural information such as locations of sentence or DA boundaries. Such locations are overt in standard text via punctuation and capitalization, but are “hidden” in speech. As shown by a number of studies, the absence of sentence or DA boundaries in speech transcripts causes difficulties for both humans and computers.

Effects on human sentence processing were studied by Jones et al. (2003), who demonstrated that sentence breaks are critical for readability of speech transcripts. Moreover, a lack of sentence segmentation can make the meaning of some utterances ambiguous. To take an extreme case, if an automatic speech recognizer outputs the stream of words “*no rooms are available*”, it is not clear what was said – whether it was “*No rooms are available.*” or “*No. Rooms are available.*” In this example, the two possible interpretations have completely opposite meaning. Such cases are

[*] Corresponding author. Address: University of West Bohemia, Department of Cybernetics, Univerzitní 8, 306 14 Plzeň, Czech Republic. Tel.: +420 377 63 2563; fax: +420 377 63 2502.

E-mail addresses: jachym@kky.zcu.cz (J. Kolář), yangl@hlt.utdallas.edu (Y. Liu), ees@speech.sri.com (E. Shriberg).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.10.005

relatively rare, but other forms of ambiguity can be much more frequent.

Lack of linguistic unit boundaries also causes significant problems for automatic processing. Many natural language processing (NLP) techniques (e.g., parsing, automatic summarization, information extraction, machine translation) are typically trained on well-formatted input, such as text, and fail when dealing with unstructured streams of words. For instance, Furui et al. (2004) reported that speech summarization improved when sentence boundaries were provided. In the area of parsing, Kahn et al. (2004) achieved a significant improvement in parsing performance when using a more accurate sentence boundary detection system. Furthermore, Matusov et al. (2007) showed that the use of automatically-detected sentence boundaries is beneficial for machine translation.

State-of-the-art approaches to DA segmentation typically use both lexical and prosodic information. Most prior work on this task has focused on identifying effective features or on developing advanced models. Such work has almost exclusively trained aggregate models, representing data pooled over speakers.

In this work, we investigate whether the speakers differ enough from each other in the production of DA boundaries to merit speaker-dependent modeling for this task. We perform speaker adaptation for both language and prosodic models, using a speech corpus of multiparty meetings. The meeting domain is chosen for several reasons. First, we are interested in spontaneous speech, since modeling of idiosyncratic lexical patterns for DA segmentation would not be meaningful for corpora of read speech. We also expect that idiosyncratic prosodic patterns are best seen in spontaneous speech. Second, as in any study of adaptation, it is essential to have enough data to adapt the general model to the specific one. Thus in our case the target domain should have speakers with plenty of speech data, and ideally data from different conversations for purposes of generalization.

For these reasons, we use data from a corpus comprising a series of naturally-occurring meetings. This corpus, like many real-world meetings, has recurring participants, presenting the opportunity for adapting models to the individual talkers. Furthermore, in this corpus, as in other meeting applications, the speakers are known beforehand and are recorded on separate channels. This allows us to focus on the question of inherent contributions from speaker-adaptive modeling, rather than confound results with the issue of speaker separation or recognition.

We ask several questions about speaker variation in lexical and prosodic patterns associated with DA boundaries. First, we ask whether speakers differ enough from overall (speaker-independent) models to benefit from model adaptation using a relatively small amount of their speech. Second, we explore whether the effectiveness of adaptation is correlated with the data amount available for adaptation. If this is not found to be the case, then it would suggest that speakers differ inherently in how well they are characterized by generic models. Third, we investigate whether adaptation performance is dependent on different DA types (for example statements versus questions). Finally, we compare speaker adaptation results for language modeling versus prosodic modeling.

The remainder of this paper is organized as follows. Section 2 surveys related work. Section 3 describes our language and prosodic models for DA segmentation, and the speaker adaptation approach. Sections 4 and 5 present our experimental setup and discuss the experiment results. Section 6 provides a summary and conclusions.

2. Related work

General (speaker-independent) methods for automatic detection of linguistic unit boundaries in speech have been studied quite extensively in the past decades. Several different approaches utilizing both textual (lexical or syntactic) and acoustic (prosodic) information have been proposed. The proposed techniques include hidden Markov models (HMMs) (Shriberg et al., 2000; Kim and Woodland, 2003), multilayer perceptrons (Warnke et al., 1997; Srivastava and Kubala, 2003), maximum entropy (Huang and Zweig, 2002; Liu et al., 2004), conditional random fields (Liu et al., 2005; Zimmermann, 2009), support vector machines (Akita et al., 2006; Magimai-Doss et al., 2007), and adaptive boosting (Zimmermann et al., 2006; Kolář et al., 2006b). Syntactic information has been used in (Roark, 2006; Favre et al., 2008). Domain adaptation for sentence boundary detection has been studied by Cuendet et al. (2006). However, basically no attention has been paid to *speaker adaptation* of lexical or prosodic models.

For related work, we should mention papers focusing on speaker-dependent modeling for general LM adaptation in speech recognition. Besling and Meier (1995) improved an automatic speech dictation system by speaker LM adaptation based on the LM fill-up method. Akita and Kawahara (2004) showed improved recognition performance using LM speaker adaptation by scaling the *n*-gram probabilities with the unigram probabilities estimated via probabilistic latent semantic analysis. Tur and Stolcke (2007) demonstrated that unsupervised within-speaker LM adaptation significantly reduced word error rate in meeting speech recognition.

Even less is known about speaker-specific variation in prosodic patterns, beyond basic F_0 normalization used by Shriberg et al. (2000). Studies in speech synthesis and speaker recognition have used prosodic variation successfully, but to our best knowledge, modeling stylistic prosodic variability for sentence or DA boundary detection has been mentioned only anecdotally in the literature (Ostendorf and Veilleux, 1994; Hirst and Cristo, 1998).

We have already presented preliminary results of this work in two conference papers (Kolář et al., 2006a; Kolář et al., 2007). Unlike the two earlier papers, this paper contains more results, analysis, and discussion. It also