Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2018) xxx-xxx
www.elsevier.com/locate/csl

Feature-space SVM adaptation for speaker adapted word prominence detection☆

Q2 Andrea Schnall[a,*], Martin Heckmann[b]

[a] Control Methods and Robotics Lab, TU Darmstadt, Landgraf-Georg Str. 4, Darmstadt 64283, Germany
Q3 [b] Honda Research Institute Europe GmbH, Carl-Legien-Str. 30, Offenbach/Main 63073, Germany

Received 6 September 2017; received in revised form 28 April 2018; accepted 1 June 2018
Available online xxx

Abstract

Prosodic cues such as the word prominence play a fundamental role in human communication, e.g., to express important information. Since different speakers use a wide variety of features to express prominence, there is a large difference in performance between speaker dependently and speaker independently trained models. To cope with these variations without training a new speaker dependent model, in speech recognition speaker adaptation techniques such as feature-space Maximum Likelihood Linear Regression (fMLLR) turned out to be very useful. These methods are developed for GMM-HMM based classifiers under the assumption that the data can be well modeled via the mixture of a few Gaussian distributions. However, in many cases these assumptions are too restrictive. In particular a discriminative classifier such as an SVM often yields far superior results to a GMM. Therefore, we propose a new adaptation method, which adapts the data to the radial basis function kernel of the SVM. To avoid overfitting we apply two regularization terms. The first is based on fMLLR and the second is an L_1 regularization to enforce a sparse transformation matrix. We analyze the method in the context of speaker adaptation for word prominence detection, with varying amounts of adaptation data and different weights of the regularization terms. We show that our novel method clearly outperforms fMLLR-GMM and fMLLR-SVM based adaptation.

Keywords: Prosody; Speaker adaptation; FMLLR; SVM; Prominence,

1. Introduction

The information in speech is not only contained in the lexical meaning of the spoken words or even in the syntax or sentence structure. Significant amounts of essential, meaning-determining information is conveyed via prosody, the way we say words. Prosody consists of rhythm, stress and intonation. This is not part of the vocabulary or the grammar, which are typically captured within speech recognition. In human-human communication, prosody can help to structure the utterance in a phrase, emphasize the novel information, differentiate between questions and statements but also show the emotion of the speaker, sarcasm and irony. Humans are also able to understand if there

☆ This paper has been recommended for acceptance by Roger Moore.
* Corresponding author.
E-mail address: aschnall@rmr.tu-darmstadt.de (A. Schnall), martin.heckmann@honda-ri.de (M. Heckmann).

https://doi.org/10.1016/j.csl.2018.06.001

are different possible meanings of an utterance indicated by prosodical changes (Shriberg, 2005). Consequently, prosody can convey a lot of valuable information for interactive systems, but still most of these systems do not include any prosodic cues (Stolcke et al., 2000; Hirschberg et al., 2006). One reason might be that it is not easy to build a powerful detection system, since there is a large variation between different speakers. Yet without considering prosodical information, speech-processing systems stand no chance to reliably process complex human interactions (Rosenberg, 2009).

In light of these considerations, prosodical information should be used for human−machine interaction systems as well, in order to improve their performance and make their use more intuitive. One prosodic cue humans use is the prominence of a word, which we define here as a greater perceived strength of one word in a sentence (Streefkerk, 1997). The determination of word prominence plays an important role in human communication. By strongly increasing the prominence of a word, for example, a correction can be indicated (Shriberg, 2005). In state-of-the-art speech interaction systems, this way of over pronunciation often leads to even more miss-recognitions (Hirschberg et al., 2004; Swerts et al., 2000). Speakers also use prosodic cues to mark important information such that the important words stand out from their environment.

Different approaches to extract and represent this information have been proposed. For the related and more investigated pitch accent detection often non acoustic features are used (Gregory and Altun, 2004), i.e., syntactic, phonological, or acoustic features in combination with text-based features (Levow, 2008). Yet in our scenario we want to only focus on the acoustic variations underlying word prominence. Concerning different learning approaches, in most approaches supervised learning is applied (Moniz et al., 2014; Rosenberg et al., 2015), but some are based on semi-supervised learning (Jeon and Liu, 2012), and unsupervised learning methods (Tamburini and Caini, 2005; Kalinli and Narayanan, 2007; Kakouros and Räsänen, 2016). While a speaker dependently trained classifier for detecting word prominence can already reach quite good results, the performance drops if a speaker independent system is used. For real applications it is difficult to use a speaker dependently trained model due to the high amount of labeled data which is needed to train such a model. To overcome this problem, we use a speaker independent system and adapt it with a small amount of speaker dependent data to obtain an accuracy comparable to that of a model trained on a specific speaker. This small amount of labeled data could be collected prior to the first use of a system equipped which our algorithm or the system could perform an unsupervised adaptation by performing the adaptation based on data collected during the interaction with high classification condense.

Our experiments show that support vector machines (SVM) are a better choice for the classification than e.g. a Gaussian mixture model based classification or, in case of a low amount of training data, Deep Neural Networks (DNNs) (Schnall and Heckmann, 2016b). Therefore, we investigate speaker adaptation methods for SVM-based classifiers. A standard adaptation method in speech processing, the feature-space maximum likelihood linear regression (fMLLR) (Gales, 1998), showed not to be successful in combination with an SVM, since the assumption that the data is well described as the superposition of a small number of Gaussian kernels in a Gaussian mixture model is questionable. Another issue might be that fMLLR does not incorporate discriminative information. In this paper we propose a method, based on the radial basis function parameters of the trained SVM model. In the adaptation these parameters are used to incorporate the information of the decision boundary. Subsequently, to avoid overfitting and reach better results with a low amount of adaptation data, we use a regularization to constrain the model. More precisely, we use an fMLLR cost term to regularize the SVM adaptation as well as an L_1 regularization. Preliminary results of these methods have been presented in Schnall and Heckmann (2016c,a). In this paper we substantially extend the experiments and perform a more in depth analysis of the results.

The next section will present the state of the art methods. Afterward, in Section 3, we give an overview of the fMLLR (3.1) followed by the introduction of our new adaptation methods, the feature-space SVM adaptation (3.2) and the Gaussian-regularized SVM adaptation (3.3). Section 4 presents the database used followed by a description of the features in Section 5. Following a short overview on the experimental setup in Section 6 is the presentation of the results in Section 7. We will end with a discussion of our findings in Section 8 and a subsequent conclusion in our final section.

2. State of the art

Speaker adaptation techniques try in general to reach speaker dependent classification performance with only a small amount of speaker specific data. The methods can be divided in speaker normalization such as Vocal Tract

Please cite this article as: A. Schnall, M. Heckmann, Feature-space SVM adaptation for speaker adapted word prominence detection, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.06.001