Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 37 (2016) 11–23

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Speaker-adapted confidence measures for speech recognition of video lectures☆

Isaias Sanchez-Cortina*, Jesús Andrés-Ferrer, Alberto Sanchis, Alfons Juan

MLLP, DSIC, Universitat Politècnica de València, Camí de Vera, s/n, 46022 València, Spain

Received 28 November 2014; received in revised form 19 October 2015; accepted 21 October 2015
Available online 21 November 2015

Abstract

Automatic speech recognition applications can benefit from a *confidence measure* (CM) to predict the reliability of the output. Previous works showed that a *word-dependent naïve Bayes* (NB) classifier outperforms the conventional word posterior probability as a CM. However, a discriminative formulation usually renders improved performance due to the available training techniques.

Taking this into account, we propose a *logistic regression* (LR) classifier defined with simple input functions to approximate to the NB behaviour. Additionally, as a main contribution, we propose to adapt the CM to the speaker in cases in which it is possible to identify the speakers, such as online lecture repositories.

The experiments have shown that speaker-adapted models outperform their non-adapted counterparts on two difficult tasks from English (videoLectures.net) and Spanish (poliMedia) educational lectures. They have also shown that the NB model is clearly superseded by the proposed LR classifier.

Keywords: Confidence measures; Speech recognition; Speaker adaptation; Log-linear models; Online video lectures

1. Introduction

Significant advances in the field of *Automatic Speech Recognition* (ASR) have been achieved over the last decades. Nowadays, automatic transcriptions of spontaneous speech in moderately noisy environments have reached an accurate enough quality (Rousseau, 2011; Sundermeyer et al., 2011; Swietojanski et al., 2013). This quality can be even better when ASR systems are adapted to specific scenarios (Leggetter and Woodland, 1995; Gales, 1998; Gauvain and Lee, 1994; Digalakis et al., 1995; Wiesler et al., 2014; Martinez-Villaronga et al., 2013). Nonetheless, ASR is still far from producing error-free transcriptions and, consequently, its performance in many applications is not completely satisfactory.

☆ This paper has been recommended for acceptance by James Glass.

* Corresponding author. Tel.: +34 663673215.
E-mail addresses: isanchez@dsic.upv.es (I. Sanchez-Cortina), jandres@dsic.upv.es (J. Andrés-Ferrer), josanna@dsic.upv.es (A. Sanchis), ajuan@dsic.upv.es (A. Juan).

http://dx.doi.org/10.1016/j.csl.2015.10.003
0885-2308/

To further improve the usefulness and performance of the current technology, researchers have proposed to compute a normalised score or *confidence measure* (CM) to indicate the reliability of the ASR output. This score has been computed at different levels: phoneme, word, phrase or sentence. Nevertheless, CM at the word level has been the main focus in the literature due to its usefulness for the vast majority of applications (Wessel et al., 2001; Kim and Ko, 2005; Gao et al., 2009; Bardideh et al., 2007; Wang et al., 2010; Junfeng and Yeping, 2011; Yadav and Patil, 2013).

One widely used word-level CM has been word posterior probability (Wessel et al., 2001). From then on, many works have focused on combining word posterior with additional sources of knowledge. The combination has been addressed as a classification problem in the vast majority of the works. Most well-known classifier algorithms have been tried: linear, Gaussian mixtures, neural networks, decision trees, support vector machines, etc. For further reference, a still good comprehensive survey can be found in Jiang (2005).

In the framework of CM as a classification problem, significant improvements were achieved by means of a combination of word-dependent (specific) and word-independent (generalised) *naïve Bayes* (NB) classifiers (Sanchis et al., 2012). Nonetheless, NB is learned by means of a generative criterion, the *maximum likelihood estimate* (MLE), which involves some issues. In particular, MLE overfits due to the unseen data. This issue was addressed in NB work by using a complex *backing-off* smoothing technique. But still, MLE aims at modelling the distribution underlying a given sample, which does not guarantee the solution to be the best suited for classification. Indeed, better fitted criteria may improve overall performance. For instance, the *maximum mutual information* (MMI) (Heigold et al., 2010) aims at better discriminating between classes without explaining the data. This criterion has been widely exploited in the literature for the *maximum entropy* (ME) models (Guiasu and Shenitzer, 1985; Yu et al., 2011).

Nevertheless, despite the success of MMI training in many applications, there is no direct relationship between maximising the MMI and minimising the probability of classification error. Instead, there are better suited criteria, which guarantee the minimisation of the *classification error rate* (CER) such as the *minimum classification error* (MCE) or the *mean squared error* (MSE). Therefore, we propose a *logistic regression* (LR) model to be learnt by means of the MSE to surpass NB performance.

On the other hand, speaker model adaptation has proved to be very effective for the improvement of recognition performance (Leggetter and Woodland, 1995; Gales, 1998; Gauvain and Lee, 1994; Digalakis et al., 1995). However, adaptation of the CMs to the speaker is nowadays unexplored. There is an increasing number of interesting scenarios in which CMs can be very useful and information about the speaker is available, such as the online lecture repositories. These repositories usually count with a large number of speeches delivered by a reduced number of speakers. Improving CM performance in these academic repositories is highly motivated since manual transcription is not affordable for such a large amount of speeches. Moreover, ASR performance is usually poor due to the amount of technical concepts, very different native and non-native accents, etc. In this scenario, *interactive speech transcription* (IST) guided by CMs can help in massively producing acceptable transcripts for large amounts of videos with limited manual effort (Silvestre-Cerda et al., 2013).

Motivated by the scenario depicted above, we propose to adapt the CM models to the speaker in an attempt to improve CM classification and IST performance. To do so, we formulate the speaker adaptation to extend both the published NB and the proposed LR models.

The rest of the content is organised as follows: the inclusion of speaker dependence into the NB model is described in Section 2. Section 3 proposes the LR model and formulates its corresponding speaker-dependent version. Section 4 describes the evaluation of the proposed models on two challenging tasks based on ASR transcripts from videoLectures.net and poliMedia repositories. Comparative results are presented including also conditional random field (CRF) models (Seigel, 2013; Seigel and Woodland, 2011; Fayolle et al., 2010). Section 5 proves that the increased CM performance results in better amended transcripts for videoLectures.net when integrated into an IST application. Finally, Section 6 raises the conclusions.

2. Speaker-adapted naïve Bayes classifier

In this section, we introduce a speaker-adapted confidence estimator model. The model is designed to extend the *naïve Bayes* (NB) approach that was successfully applied to speech recognition (Sanchis et al., 2012) as well as to machine translation (Sanchis et al., 2007). Thus, let us first briefly recall the speaker independent NB model.