Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 28 (2014) 1083–1114

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Feature analysis for discriminative confidence estimation in spoken term detection☆

Javier Tejedor[a,*], Doroteo T. Toledano[b], Dong Wang[c], Simon King[d], José Colás[a]

[a] *Human Computer Technology Laboratory, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain*
[b] *ATVS-Biometric Recognition Group, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain*
[c] *Center for Speech and Language Technologies, Tsinghua University, Beijing 100084, PR China*
[d] *Centre for Speech Technology Research, University of Edinburgh, UK*

Received 2 July 2012; received in revised form 27 February 2013; accepted 16 September 2013
Available online 4 October 2013

Abstract

Discriminative confidence based on multi-layer perceptrons (MLPs) and multiple features has shown significant advantage compared to the widely used lattice-based confidence in spoken term detection (STD). Although the MLP-based framework can handle any features derived from a multitude of sources, choosing all possible features may lead to over complex models and hence less generality. In this paper, we design an extensive set of features and analyze their contribution to STD individually and as a group. The main goal is to choose a small set of features that are sufficiently informative while keeping the model simple and generalizable. We employ two established models to conduct the analysis: one is linear regression which targets for the most relevant features and the other is logistic linear regression which targets for the most discriminative features. We find the most informative features are comprised of those derived from diverse sources (ASR decoding, duration and lexical properties) and the two models deliver highly consistent feature ranks. STD experiments on both English and Spanish data demonstrate significant performance gains with the proposed feature sets.

Keywords: Feature analysis; Discriminative confidence; Spoken term detection; Speech recognition

1. Introduction

1.1. Spoken term detection

The enormous amount of speech information now stored in audio repositories motivates the development of automatic audio indexing and spoken document retrieval methods. Spoken term detection (STD), defined by NIST as *searching vast, heterogeneous audio archives for occurrences of spoken terms* (NIST, 2006), is a fundamental building block of such systems (Mamou and Ramabhadran, 2008; Can et al., 2009; Vergyri et al., 2007; Akbacak et al., 2008; Szöke et al., 2008, 2008; Thambiratmann and Sridharan, 2007; Wallace et al., 2010; Jansen et al., 2010; Parada et al.,

☆ This paper has been recommended for acceptance by Murat Saraclar.

* Corresponding author. Tel.: +34 914976216.
E-mail addresses: javier.tejedor@uam.es, javiertejedornoguerales@gmail.com (J. Tejedor), doroteo.torre@uam.es (D.T. Toledano), wangdong99@mails.tsinghua.edu.cn (D. Wang), simon.king@ed.ac.uk (S. King), jose.colas@uam.es (J. Colás).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.09.008

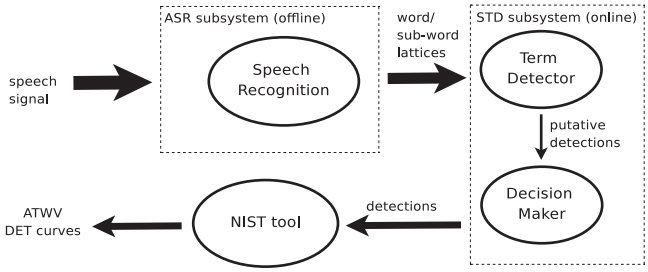

Fig. 1. The standard STD architecture and evaluation.

2010; Chan and Lee, 2010; Chen et al., 2010; Motlicek et al., 2010), and its development has been strongly influenced by NIST STD evaluations (NIST, 2006, 2013).

The standard STD architecture is comprised of two main stages: indexing by the Automatic Speech Recognition (ASR) subsystem, then search by the STD subsystem, as depicted in Fig. 1. The ASR subsystem transforms the input speech into word or sub-word lattices. The STD subsystem comprises a *term detector* and a *decision maker*. The term detector searches for putative occurrences of the query terms in the word/sub-word lattices – it hypothesizes detections – and the decision maker then decides whether each detection is reliable enough to be considered as a hit or should be rejected as a false alarm (FA). A tool provided by NIST is used for performance evaluation. It must be noted that the ASR subsystem must run just once and therefore the STD subsystem cannot make use of the speech signal directly.

Searching the output of a Large Vocabulary Continuous Speech Recognition (LVCSR) system, i.e., word lattices, has been shown to work well when the query terms are only composed of in-vocabulary (INV) words, since these will be in the LVCSR system vocabulary and therefore will occur in the word lattices. However, as noted by Logan et al. (2000), about 12% of users' queries typically contain out-of-vocabulary (OOV) words, which will never be found in the word lattices, because they do not appear in the LVCSR system vocabulary. Common approaches to solve this problem usually involve producing sub-word (typically phone/phoneme) lattices with the ASR subsystem, and then searching for sub-word representations of the enquiry terms (Saraçlar and Sproat, 2004; Mamou et al., 2007; Can et al., 2009; Szöke et al., 2006; Wallace et al., 2007; Parlak and Saraçlar, 2008). Other sub-word units are possible, such as syllables (Meng et al., 2007), graphemes (Wang et al., 2008; Tejedor et al., 2008) or multi-grams (Pinto et al., 2008; Szöke et al., 2008).

In STD, a *confidence score* is assigned to each putative occurrence detected in the lattice, which reflects the possibility of it being a real occurrence. A widely used confidence score that can be derived from the lattice is defined as follows:

$$c_f = \frac{\sum_{\pi_\alpha,\pi_\beta} p(O|\pi_\alpha, K_{t_s}^{t_e}, \pi_\beta) P(\pi_\alpha, K_{t_s}^{t_e}, \pi_\beta)}{\sum_{\varsigma} p(O|\varsigma) P(\varsigma)} \tag{1}$$

where $K_{t_s}^{t_e}$ denotes a detection of K, which is a partial path that starts at t_s and ends at t_e and corresponds to the pronunciation of term K. c_f is the confidence of $K_{t_s}^{t_e}$. π_α and π_β denote paths before and after K respectively, with π_α starting from the beginning of the audio and π_β ending at the end of the audio. ς in the denominator represents any full path in the lattice. Note that a particular term occurrence may correspond to a group of overlapped detections $\{K_{t_s^i}^{t_e^i}\}$. In that case, the detection group is treated as a single detection, and c_f is derived by a certain merging scheme (Wang et al., 2011). In this work, we simply choose the best confidence of the group members as c_f.

Based on the confidence scores, the decision maker determines which putative occurrences are reliable enough to be called *detections*. If a detection actually appears in the audio, it is called a *hit*. Otherwise, it is called a *false alarm*. Any occurrence of the query term in the audio that is not hypothesized by the STD system is called a *miss*.

To evaluate STD system performance, NIST defines an evaluation metric called *actual term weighted value (ATWV)* (NIST, 2006), which integrates the hit rate and false alarm rate into a single metric and then averages over all search terms:

$$ATWV = \frac{1}{|\Delta|} \sum_{K \in \Delta} \left(\frac{N_{hit}^K}{N_{true}^K} - \beta \frac{N_{FA}^K}{T - N_{true}^K} \right) \tag{2}$$