Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 28 (2014) 997–1017

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Fast vocabulary acquisition in an NMF-based self-learning vocal user interface

Bart Ons [*], Jort F. Gemmeke, Hugo Van hamme

Department ESAT-PSI, KU Leuven, Leuven, Belgium

Received 11 April 2013; received in revised form 7 March 2014; accepted 14 March 2014
Available online 25 March 2014

Abstract

In command-and-control applications, a vocal user interface (VUI) is useful for handsfree control of various devices, especially for people with a physical disability. The spoken utterances are usually restricted to a predefined list of phrases or to a restricted grammar, and the acoustic models work well for normal speech. While some state-of-the-art methods allow for user adaptation of the predefined acoustic models and lexicons, we pursue a fully adaptive VUI by learning both vocabulary and acoustics directly from interaction examples. A learning curve usually has a steep rise in the beginning and an asymptotic ceiling at the end. To limit tutoring time and to guarantee good performance in the long run, the word learning rate of the VUI should be fast and the learning curve should level off at a high accuracy. In order to deal with these performance indicators, we propose a multi-level VUI architecture and we investigate the effectiveness of alternative processing schemes. In the low-level layer, we explore the use of MIDA features (Mutual Information Discrimination Analysis) against conventional MFCC features. In the mid-level layer, we enhance the acoustic representation by means of phone posteriorgrams and clustering procedures. In the high-level layer, we use the NMF (Non-negative Matrix Factorization) procedure which has been demonstrated to be an effective approach for word learning. We evaluate and discuss the performance and the feasibility of our approach in a realistic experimental setting of the VUI-user learning context.

Keywords: MIDA; Phone posteriorgram; NMF; Fast learning; Vocabulary acquisition

1. Introduction

Command-and-control (C&C) speech recognition allows users to interact with systems like domestic devices, assistive technology, computers, smart-phones or other mobile devices. The user speaks a command or a phrase to control different functions in the environment like the central heating or the light units in the house, to retrieve information on their smartphone or to navigate through a menu on a computer. C&C applications are especially useful

[*] Corresponding author. Tel.: +32 16321071.
E-mail address: Bart.ons@esat.kuleuven.be (B. Ons).

http://dx.doi.org/10.1016/j.csl.2014.03.004

for people with a physical disability affording them handsfree control of their wheel chair, the positioning of their bed or other independent living aids.

In most speech driven C&C applications, the spoken commands are restricted to a predefined list of phrases described by a restricted grammar and vocabulary. The size of the vocabulary ranges from a few to a few hundred words and the grammars are mainly rule-based. Although the targeted VUI application allows a developer to consider many interaction scenarios beforehand, the use of a VUI is not always successful when the interaction oversteps the clear boundaries of the lexicon, the grammars or the dialogue models. Even in less restrictive frameworks, such as in the now popular Siri speech recognition application for the iPhone, performance degrades rapidly if the acoustic models do not match the speech material used to train the system, for example on accented or dysarthric speech. The goal of this paper is to investigate a VUI model which is able to associate any utterance to a C&C action allowing command and control usability by deviant speech as well.

Over the past decade, various approaches have been proposed for adaptation to unexpected circumstances in real-life situations. For instance, Paek and Chickering (2007) proposed a statistical model for mobile devices that tracks the past of the user's behaviour in order to predict commands. In Heinroth et al. (2012), grammars were able to adapt dynamically to real-life communication making interactions more natural. In Potamianos and Narayanan (1998) and Kuhn et al. (2000), speaker-independent acoustic models were adapted to speaker-dependent models allowing for better recognition of the user-specific vocalizations. There are plenty more studies that paved the way to more natural interaction with machines and devices by means of human-centred design and user adaptation. For instance, in a study of Parker et al. (2006) a robust speech recogniser was developed to adapt to dysarthric speech as well. In the "Speech Training And Recognition for Dysarthric Users of Assistive Technology" (STARDUST) project (Parker et al., 2006), the problem was tackled by adaptation in two directions: a training package assisting dysarthric speakers to improve the recognition likelihood of their utterances (users adapting to speech recognition systems) and speech recognition systems having greater tolerance to variability of dysarthric vocalizations (speech recognition models adapting to users) were developed.

However, all these approaches have in common that these systems are still based on acoustic and language models that are trained beforehand and adapted through interaction to the spoken utterances of the user. While these methods focus on adaptation, we focus on *grounding*: learning both vocabulary and acoustics directly from the user during the usage of the VUI. The grounding process (Clark and Schaefer, 1989) refers to the process by which common ground or meaning is built between the user and the system. Situated in the "Adaptation and Learning for Assistive Domestic Vocal Interfaces" (ALADIN) project (van de Loo et al., 2012; Gemmeke et al., 2013), we aim to design a VUI that learns to understand speech by mining the speech input from the end user and the changes that are provoked on a device.

The VUI should learn to understand classes referring to devices, actions or properties by using cross-situational evidence and learning the statistical regularities between two modalities, namely, the spoken utterances of the user and the feedback coming from the device(s). Supervision coming from the device is weak in the sense that the information provided to the VUI consists of signals referring to states and actions in a machine without any chronological information, orthographic nor phonetic transcriptions. Earlier studies have demonstrated that multi-modal Non-negative Matrix Factorisation (NMF) is a useful tool to learn weakly co-occurring regularities over two modalities in order to find the intra- and inter-modality patterns. For instance, in Caicedo et al. (2012), NMF is used to generate multimodal image representations that integrate visual and text features for image collections guided by ratings, comments and tags on the web. Akata et al. (2011) used a similar approach and called it multiview clustering to cluster images and predict image labels. Similar to NMF-based keyword discovery in Driesen et al. (2012a), we use NMF to learn co-occurrences between acoustic feature vectors emerging from the spoken utterances and semantic label vectors describing the action properties.

In order for a self-learning approach to be useful as a VUI, the learning process should be *fast*. At the same time, after sufficient training tokens have been presented, the *accuracy* should be high. The contribution of this work is twofold. First, we investigate to what extent the learning speed and accuracy can be improved by using more advanced feature representations in NMF. We use phone classifiers to create phone confidence measures to replace the conventional acoustic input in NMF learning (Driesen, 2012; Sun, 2012). In addition to phone classifiers, we also evaluate a speaker-dependent version of soft Vector Quantization (soft VQ), which is a data-driven and probabilistic procedure to cluster the acoustic data of the speaker. We tested the usefulness of this data-driven approach for small training sets as user-specific data is expected to be scarce in the beginning of the VUI usage.