Flexible Stand-Alone Keyword Recognition Application Using Dynamic Time Warping

Miquel Ferrarons¹,²,⋆, Xavier Anguera¹, and Jordi Luque¹

¹ Telefonica Research, Edificio Telefonica-Diagonal 00, 08019, Barcelona, Spain
² Universitat Autonoma de Barcelona, Barcelona, Spain
{xanguera,jls}@tid.es

Abstract. We introduce a Query-by-Example (QbE) application for smart-phone devices that implements a recently proposed memory-efficient dynamic programming algorithm [1] for the task of keyword search. The application compares acoustic keywords with the audio input from the microphone and reacts to detected keywords with actions in the phone. These keywords are recorded by the user, who also defines what actions will be performed by each one. One of these keywords is defined to be a trigger keyword, which is used to wake up the system and thus reduce false detections. All keywords can be freely chosen by the user. In Monitor mode, the application stays listening to audio acquired through the microphone and reacts when the trigger + some keyword are matched. All processing is done locally on the phone, which is able to react in real-time to incoming keywords. In this paper we describe the application, review the matching algorithm we used and show experimentally that it successfully reacts to voice commands in a variety of acoustic conditions.

Keywords: Mobile search, dynamic time warping, query-by-example, keyword recognition.

1 Introduction

Currently smartphones are found everywhere. The usage of smartphones is not only driven by the need to make phone calls, as more and more people use them for activities like gaming, browsing the internet, reading, etc. Voice recognition is quickly gaining popularity and acceptance among smartphone users, probably due to the small keyboard footprint of these devices which makes it more complicated to type on it than to speak to it. Companies like Google [2] of Microsoft [3] have recently proposed powerful speech-enabled applications in the cellphone that are changing the general public miss-conception that speech recognition does not work. Driven by this trend, we decided to experiment with small-footprint but flexible voice-enabled command-and-control applications for the cellphone. In this paper we present one of these experiments. It corresponds to a keyword recognition application we developed for Android devices which

* M. Ferrarons was visiting Telefonica Research at the time this work was performed.
Flexible Stand-Alone Keyword Recognition Application

We implement a flexible Query-by-Example (QbE) algorithm to enable many functionalities in the phone by just giving spoken orders to it. QbE algorithms are used to search for matches of a given spoken query within a set of spoken utterances. In this case we use it to match pre-recorded acoustic keywords against online audio captured from the smartphone’s microphone. The proposed application is a proof of concept of a totally offline (no server connectivity required) speech enabled tool to control the telephone’s functionalities by voice. In addition, recorded keywords can be chosen by the user to be whatever word/sound he/she wants, and each one can be recorded multiple times to improve matching accuracy. There are two kinds of keywords, “trigger” and “action” keywords. A “trigger” keyword is recorded by the user to “wake up” the application (i.e. indicate that an “action” keyword will be spoken next). Then, “action” keywords are detected and associated to actions in the phone. In this proof of concept some actions have been implemented, like taking a picture, recording a voice note or picking up a call.

The rest of the paper is organized as follows: Section 2 describes the application implementation details, including the user interface (Subsection 2.1) and each of the two modes of operation (Subsections 2.2 and 2.3). Then, Section 3 describes and performs an evaluation of usage of the application among real users, to show that it does perform as expected. Finally, in Section 4 we draw some conclusions from the presented work, and draw some lines of future research.

2 System Description

The proposed application has two main modes of operation: Monitor mode, and Recording mode. In Recording mode the user records new acoustic keywords, assigns actions to them, and can set some parameters to each of these keywords. In Monitor mode the application constantly records the ambient sound and reacts accordingly when the user says any of the keywords recorded in the Recording mode. Next we describe the application user interface and how each of these modes works in more detail.

2.1 Application User Interface

We implemented the keyword recognition application in Android. Our development was tested using a Samsung GalaxyS phone, although any smartphone with similar (or superior) capabilities should be able to successfully run the application. Figure 1 shows three screenshots of the application as it is being used. Subfigure 1a shows the main screen for the Recording mode and Subfigure 1c shows the Monitor mode. The user can switch between the two modes by clicking the tab present on top of either of these screens.

Within the Recording mode the user sees the list of keywords that he has already recorded, and has the option (by clicking in “New Recording”) to record a new keyword. There is no limit to the number of different keywords that a user can record, although at some point the application might not be able to monitor...