Voice Command Recognition

Javier Fernández JAVIERFDR@GMAIL.COM

Alejandro Hernández ALEJANDRO.AJHR@GMAIL.COM

1. Problem Description

Speech recognition is a complex study field that focus on the translation of spoken words into text, commonly through the use of computer systems. To recognize words from audio brings many challenging tasks to deal with such as vocalization variations in terms of accent, pronunciation, volume and speed, reading vs. spontaneous speech, known vocabulary size, and environmental noise, among others. Speech recognition systems often require to gather and process great amounts of audio data and are dependent on the available languages and the quality of this data, and also it needs large corpora of text to build language models that allows the system to correct most mistakes and produce a text output of decent quality. Moreover, after these models are built and trained, both audio and text based commonly require high processing capacities from computer servers.

Speech recognition has many applications, and becomes particularly interesting when including it within embedded computing devices connected to the internet infrastructure, referred currently as *The internet of things (IoT)*. Having these devices the ability to understand human voice commands could improve the usability of this gadgets and also become its use more *natural*. An application of this can be seen, for example, for home automation devices such as centralized control of lighting, ventilation and air conditioning, home entertainment systems, pet feeding, home robot control and even security locks.
However, implementing a speech recognition system for IoT would have to deal with all the problems listed above, and would require to be consistently accurate in order to be feasible to be used by humans. It also requires technology and data processing time that is less likely that it could be managed by small to medium companies providing IoT gadgets.

A solution for this problem would need to be easily trained, very accurate for a set of specific commands accepted by each gadget, and implemented in a light-weight technology that could be either embedded on IoT objects or lightly deployed on computer servers.

2. Proposed Solution

A voice command recognition solution is proposed that can be implemented in IoT gadgets, requiring a simple training phase and being accurate for recognizing specific voice commands

accepted by each particular gadget. The purpose of the system is not to establish a dialogue but to convey an instruction, and according with that perspective the solution focuses in recognizing the voice tone differences that are more likely to represent a particular command. The audio is not translated to text, but matched with a learned model of voice tones referring to specific commands. A classification algorithm is used to learn patterns in the training audio data and outputs a specific command (class) from a given input. It is also a desirable quality of the system to be invariant with respect to the users, been able to recognize the same command from different people, provided they have a relatively close pronunciation of the word.

2.1. Goals

- To recognize specific commands from audio input.
- Robust recognition based on voice tone differences, minimizing false negatives.
- To develop and easy to train and fast to classify solution that is fit to run on embedded systems or light-weight servers.
- Good performance compared with existing open source speech recognition solutions.

2.2. General Description

The proposed solution is based on the following processing tasks:

- **Human voice recorded data**: That is obtained from human voice input recorded in electronical devices such as computers and mobile phones.
- **Signal Processing:** Obtaining a graphical representation of the input audio data in terms of signal amplitude, frequency and intensity, such as a spectrogram.
- **Voice tones differences detection:** using image processing to detect interesting feature points that best represent voice tones in the audio input.
- **Data transformation:** processing feature points to create a standard data structure among all samples.
- **Classification:** training a classification model from structured extracted data to assign a class (representing a command) from a given input.

Some of the main challenges of the proposed implementation are:

- The spectrogram must provide important graphical features that can be extracted through rather-simple image processing techniques. Is possible that noisy data would difficult the feature extraction process from images. Here the audio quality and the parameters used to obtain a short-window or long-window spectrogram are key decisions.

- The obtained features should be aggregated or grouped in a way that they form a representative data structure. The data through this representation should be consistently similar for the same commands and rather-different for different commands. This is what is wanted to be proved.
- Classification could be accurate in different levels such as: identifying the same command from different users, identifying commands just for a user (audio fingerprint), identifying commands recorded in different audio quality conditions (such as environmental noise or use of different devices).

3. Previous work

The problem in hand and its proposed solution have been worked with different approaches and focuses, and thus, there is a huge background of methods and technologies to be considered as previous work. However, the most relevant works in analysis of audio frequency patterns have been done for song analysis, thus the solution previous work will be highly inspired by such work and also by the most relevant (and open) implementation of speech recognition. Although the proposed solution is not based directly in these methods they are interesting for benchmarking comparison.

Below some interesting references related to this work are listed:

- An Industrial-Strength Audio Search Algorithm (1)
- An Overview of the SPHINX Speech Recognition System (2)
- PocketSphinx: A free, real-time continuous speech recognition system for hand-held devices (3)
- Parakeet: A Continuous Speech Recognition System for Mobile Touch-Screen Devices (4)
- Singing voice recognition based on matching of spectrogram pattern (5)

4. Proposed Experiments

In order to test the performance of the system, we must compare it with the current standard solution which is speech recognition software such as Sphinx. However, Sphinx is not limited or focused in solving the problem proposed by this work, so we will use pocketsphinx, which is a reduced version of sphinx adapted to this purpose.

Then we would like to test the robustness of the system to two kinds of variations: one is environmental noise, and the other is the subjective variation in the speech of an user. Concretely, the proposed experiments are:

- Test command recognition against pocketsphinx. A set of N recordings for each of C Commands.
- Perturbate examples with noise.
- Test with different qualities (subjective) of spoken commands (how well or accurate are the commands said)

References

[1] Avery Li-Chun Wang, *An Industrial-Strength Audio Search Algorithm.* 4th Symposium Conference on Music Information Retrieval, page 7–13, 2003.

[2] Kai-Fu Lee, Hsiao-Wuen Hon, and Raj Reddy *An Overview of the SPHINX Speech Recognition System* . IEEE Transactions on Acoustic Speech and Signal Processing, VOL. 38. NO. I, 1990.

[3] David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W Black, Mosur Ravishankar, and Alex I. Rudnicky *PocketSphinx: A Free, real-time continuous speech recognition system for hand-held devices.* , 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006.

[4] Keith Vertanen and Per Ola Kristensson *Parakeet: A Continuous Speech Recognition System for Mobile Touch-Screen Devices* ,
IUI '09 Proceedings of the 14th international conference on Intelligent user interfaces Pages 237-246, 2009.

[5] Khunarsal, P, Lursinsap, C. and Raicharoen, T. Alan W Black, Mosur Ravishankar, and Alex I. Rudnicky *Singing voice recognition based on matching of spectrogram pattern* , Neural Networks, 2009. IJCNN 2009.