# Machine Learning in Assistive Technology: a Solution for People with Dysarthria

Davide Mulfari, Gabriele Meoni, Luca Fanucci
University of Pisa
Italy

## ABSTRACT

Nowadays, dysarthric speech processing represents a challenge in assistive technology contexts. In this paper, we investigate the use of machine learning in conjunction with convolutional neural networks to implement a speaker dependent solution that is capable to detect just a few number of predefined keywords. The proposed system has been trained with utterances from Italian users with severe and mild dysarthria and it is configurable according to specific users' preferences.

## CCS CONCEPTS

• **Human-centered computing** → **Accessibility**; *Accessibility systems and tools*;

## KEYWORDS

Machine Learning; Assistive Technology; Automatic Speech Recognition; Dysarthria

**ACM Reference Format:**
Davide Mulfari, Gabriele Meoni, Luca Fanucci. 2018. Machine Learning in Assistive Technology: a Solution for People with Dysarthria. In *International Conference on Smart Objects and Technologies for Social Good (Goodtechs '18), November 28–30, 2018, Bologna, Italy.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3284869.3284928

## INTRODUCTION

*Dysarthria* is a collective term for a wide range of neuromotor speech disorders resulting from abnormalities in the strength, speed, range, steadiness, tone or accuracy of movements required for control of the respiratory, phonatory, resonatory, articulatory and prosodic aspect of speech production. These conditions are often associated with various kinds of disabilities, for example quadriplegia, while lead to very low intelligibility of the users' speaking, so the disabled voice is difficult to understand by both humans (especially for unfamiliar listeners) and machines. Indeed, the speech of a person with dysarthria may vary considerably depending on time of day, level of stress, level of fatigue and presence / absence of several environmental factors. Moreover, substantial variability exists among speakers with dysarthria because of difference in severity level and involvement of various aspect of the speech production system [2]. It prevents people with speech disabilities from taking advantage of Automatic Speech Recognition (ASR) solutions on computer-based systems in many application contexts, e.g., smart home automation. In this paper, we introduce a speaker dependent ASR solution tailored to the specifics of dysarthric users which leverages the knowledge from strategies in the field of machine learning: our effort is to define a model, based on Convolutional Neural Networks (CNNs), capable of recognizing just a few number of keywords uttered by users with spastic dysarthria. Specifically, a speech dataset has been arranged and currently it contains 1.000 contribution (e.g.,audio files) of three male adults with dysarthria saying twelve different words. As described in the rest of paper, during our research Google's TensorFlow has been employed as machine learning framework, supporting both the training and the inference processes.


*Goodtechs '18, November 28–30, 2018, Bologna, Italy*

ACM ISBN 978-1-4503-6581-9/18/11.
https://doi.org/10.1145/3284869.3284928

## 1 TRAINING PROCESS

The training process is of critical importance for our project and currently is based on deep learning principles. Within speech processing field, we have considered a particular type of deep neural networks [3], that are CNNs. A CNN is composed of one or more convolutional layers pooling or subsampling layers, and fully connected layers. The aim of these layers is to extract simple representations at high resolution from the input data, and then converting these into more complex representations, but at much coarser resolutions within subsequent layers; therefore, CNNs are mainly designed for image classification problems. During our research activity, we have implemented a convolutional architecture with two convolutional layers. Here, several key layers are: Convolutional (Conv) layer (multiple convolution filters to obtain different features), Pooling layer (down-sampling by taking max operation to reduce the amount of parameters and computation in the network, and hence control overfitting), Dropout layer (only keep a neuron active with some probability p, or set it to zero otherwise to control overfitting), Linear low-rank (Lin) layer (perform linear multiplication and addition to transfer the output of Conv layer to discrete nodes, reduce parameters and computation, control overfitting), and Fully-connected (FC) layer (preserve full information, or make the final softmax prediction). The training procedure of the speech model requires Google's TensorFlow, which is a machine learning system that operates at large scale and in heterogeneous environments. In addition, TensorFlow supports both large-scale training and inference: it efficiently uses hundreds of

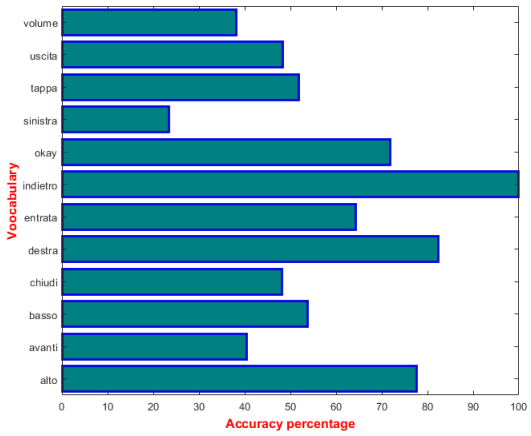


**Figure 1: Bar chart summarizing results.**

powerful (GPU-enabled) servers for fast training, and it runs trained models for inference in production on various platforms, ranging from large distributed clusters in a datacenter, down to running locally on mobile devices [1]. During our experiments, a single TensorFlow instance has been deployed on one server with a dedicated graphical processing unit; this configuration allowed us to employ CNNs in order to classify audio data (utterances) within precise classes. The process leverages knowledge from the field of image classification, by converting audio i.e., a one-dimensional continuous signal across time, into a 2D spatial problem. TensorFlow's utilities solve that issue by defining a window of time our speech commands should fit into, and converting the audio signal in that window into an image. This is done by grouping the incoming audio samples into short segments, just a few milliseconds long, and calculating the strength of the frequencies across a set of bands. Each set of frequency strengths from a segment is treated as a vector of numbers, and those vectors are arranged in time order to form a two-dimensional array. This array of values can then be treated like a single-channel image, called spectrogram [4].

## 2 EARLY EXPERIMENTS

With the collaboration of dysarthic users who have been previously involved in the training stage, we conducted initial experiments to investigate the accuracy level of the trained model. For this purpose, a separate data set for testing has been defined: it consist in 100 audio files, each recording contains just one keyword and it lasts for 2500 milliseconds. Figure 1 summarizes our results by using a bar graph. On y-axis, we report each class (single word) from our ASR vocabulary, while the accuracy percentage is shown on x-axis. Globally, we have appreciated a 57,5 % accuracy level: it is a promising result toward the creation of a vocal interface allowing the user with dysarthria to control a computer-based system using speech commands.

The last part of the paper summarizes a comparison, in terms of accuracy, between our ASR software and Dragon NaturallySpeaking from Nuance Communications, that is a popular speech recognition tool for desktop computers creating an individual voice profile for each user of the Dragon platform. The voice profile contains information about the unique characteristics of each person's voice along with a customized set of words, known as a vocabulary, and user-specific information including software settings and personalized voice commands. During our experiments, Dragon NaturallySpeaking version 13 Premium was installed and configured on a Lenovo Z50 notebook computer. Hence, a person with severe dysarthria and quadriplegia performed all the training exercises required by the software in order to generate a custom voice profile. In particular, training for Dragon required the recitation of a few short paragraphs to allow the system to calibrate and adjust to the speaker's voice. Then, the system prompted its user to read six passages before using the system. Additional training exercises were done to train Dragon considering single isolated words within our ASR Italian reduced vocabulary. In order to estimate the accuracy level, the disabled collaborator uttered each word for twenty times. No utterance were recognized by the Dragon NaturallySpeaking. In terms of accuracy, global results were very poor and were not comparable with the performance of ASR solution based on TensorFlow. These results have demonstrated that Dragon may not be intended for recognizing speech in presence of severe dysarthria, while our machine learning approach may be preferred to detect predefined keywords. We believe that it is an interesting step toward the creation of assistive technology systems for people with speech disabilities.

## 3 CONCLUSION

A speaker - dependent ASR system for users with dysarthria has been presented in this paper. Its key feature is to bring together machine learning technologies and convolutional neural networks. This approach is designed for detecting just a few number of keywords within a reduced vocabulary. Initial experiments showed promising results thanks to a dedicated training procedure based on TensorFlow framework. In future works, we plan to better investigate the proposed approach with the collaboration of many Italian disabled students attending our university.

## REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning.. In *Osdi*, Vol. 16. 265–283.

[2] Karen Hux, Joan Rankin-Erickson, Nancy Manasse, and Elizabeth Lauritzen. 2000. Accuracy of three speech recognition systems: Case study of dysarthric speech. *Augmentative and Alternative Communication* 16, 3 (2000), 186–196.

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[4] Tara N Sainath and Carolina Parada. 2015. Convolutional neural networks for small-footprint keyword spotting. In *Sixteenth Annual Conference of the International Speech Communication Association.*