Abstract

On Deep Learning strategies to address
Automatic Speech Recognition (ASR) for dysarthric speech

by Rosanna Turrisi

This thesis explores deep learning techniques to improve Automatic Speech Recognition (ASR) for people affected by dysarthria. Dysarthria is a widely spread motor disorder causing high speech unintelligibility and, often, also motor control abnormalities. Hence, ASR-based technologies may represent the only possibility for dysarthric individuals to interact with other people or machines. Unfortunately, traditional ASR systems fail in presence of dysarthric speech. For instance, we tested Google Speech API and IBM on a subset of the TORGO dataset. These provide more than 80% of WER, while the human error is 30%. One of the main issues is that the ASR model cannot capture the inter-speaker variability, due to the small and limited dysarthric speech corpora. To overcome this issue, we propose three possible strategies.

Firstly, we investigate the use of speech production knowledge as additional information in the ASR system. As the articulatory features (AFs) are difficult to collect, especially for dysarthric speakers, we move a step backward and first study deep learning based methods to synthesize AFs for audio-only corpora.
Specifically, we propose the use of phonetic features in addition or substitution to the acoustic ones in the standard Acoustic Inversion (AI) mapping, with the aim of improving its generalization across datasets. Then, we introduce unsupervised methods to synthesize AFs that leverage phonetic features, to extract raw articulatory information, and acoustic vectors, to capture complex phenomena such as the coarticulation.
We finally integrate the synthetic AFs as secondary target or additional input in the ASR model. After a preliminary study on the TIMIT corpus showing encouraging results on phone classification, we evaluate the ASR performance on CHiME-4. The first and the second strategies provide a Word Error Rate (WER) relative reduction of 1.9 % and 5.4%, respectively, over the traditional ASR system.

Secondly, we consider the scenario in which we have access to multiple labelled datasets (*sources*) and we want to learn a classifier for an unlabelled dataset (*target*). Such a problem is known in machine learning as *multi-source domain adaptation.* We propose an algorithm, named Multi-Source Domain Adaptation via Weighted Joint Optimal Transport (MSDA-WDJOT), that aims at finding simultaneously an Optimal Transport-based alignment between the source and

target distributions and a re-weighting of the sources distributions, based on their similarity with the target distribution. We then employ MSDA-WJDOT in two real-world applications: dysarthria detection and spoken command recognition. In the first case, we assume to have multiple labelled noisy datasets containing dysarthric and healthy speech and we adopt MSDA-WJDOT to learn a binary classifier for an unlabelled noisy dataset. The proposed approach outperforms all the competitor models, improving the detection accuracy of 0.9 % over the best one. In the second case, MSDA-WJDOT is used to perform dysarthric speaker adaptation in a voice command recognition system. This provides an accuracy relative improving of 21% and 12% over the baseline and the best competitor model, respectively.

Finally, we focus on contexts in which only a small vocabulary needs to be recognized. This allows to simplify the problem to spoken command recognition. Towards this direction, we collected a dysarthric speech corpus containing commands related to the task of making a call. This is the richest Italian dysarthric speech corpus to date and it can be used to train a Command Recognizer and develop a smartphone Contact application. Last but not least, we introduce the AllSpeak project in which it has been developed an Android application for people affected by Amyotrophic Lateral Sclerosis. Specifically, this App is based on a Voice Command Recognition that recognizes commands related to basic needs (e.g., “I am thirsty") even when the speech intelligibility is almost vanished.