Speaking2Me: Unsupervised Speaker Identification for Smartphones using Deep Learning

Jon Baker and Christos Efstratiou
J.Baker@kent.ac.uk,C.Efstratiou@kent.ac.uk
University of Kent, United Kingdom

1 MOTIVATION

Allowing people with sociability issues to monitor their levels of interactivity with other people can be of great value to help them manage their condition. In particular there is specific interest from people who suffer from social anxiety, or bipolar disorder, to detect social encounters and understand how they might contribute in the development of their condition. Technical solutions for passive tracking of social interactions fall broadly into two categories: (i) instrumented environments with sensing capabilities that can track social interactions within the instrumented space [2], or (ii) mobile sensing of social interactions where all the relevant parties need to use a particular type of technology (e.g. wearable social tracking tags [1]). Both of these approaches have limitations that do not make them suitable for the continuous tracking of an individual's social life throughout their daily lives. The first approach only works in certain environments (e.g. workplace), while the second limits the data capture to social interactions between users who use specialized technology. Indeed, in order for an individual user to be able to track their own social life, technical solutions should work in any environment, without the need for their social peers to use certain technology.

In this work we consider the detection of speech between individuals, where the participating parties can be accurately detected through the analysis of their voice patterns. Existing techniques in speaker identification tend to rely on supervised machine learning techniques. However, supervised techniques do not scale well, as they put a significant burden on the user to manually generate training data sets for speakers that need to be detected. In our work we consider the development of a speaker recognition solution that can run on a single user's mobile device, and accurately match voice samples with previously encountered individuals. Speaking2Me, is a system for unsupervised speaker identification using deep learning. It employs an autoencoder-based deep learning model. That model can generate an *encoding* feature vector for every voice sample, which maintains close similarity with encoded vectors of voice samples of the same individual, and significant dissimilarity with voice samples from other people. The Speaking2Me system achieves high levels of accuracy in matching voice samples of people, without the need to pre-train the system with samples from the specific people.

2 APPROACH

In Speaking2Me we produce an autoencoder model using a large public dataset of voice samples (VoxCeleb [3] - 1251 speakers) with a wide range of voices recorded in a range of environments. The encoder of this model can be used to generate an encoded feature vector that represents the characteristics of a particular voice. When feeding the encoder with previously unknown voice samples we can estimate a similarity metric between voices, and detect when voices of the same person have re-occurred in the input stream. Through this we can cluster similar voice occurrences under the same individual, and thus estimate when a social interaction with that individual has occurred. Through this process, Speaking2Me can identify social interactions with new people or the re-encounter of a previously known social contact without any intervention from the user. In a real-world deployment the Speaking2Me user can assign their own labels to such detect people if they need to.

The VoxCeleb [3] dataset, contains speech signals from YouTube videos which are shot in a large number of multi-speaker acoustic environments, including: red carpet events, outdoor stadiums and quiet studio interviews. More crucially, the data contains noise from real-world background noise, including: chatter, laughter, overlapping speech and room acoustics. For feature selection, we picked the Mel-frequency cepstral coefficients as the most appropriate input for speech processing. Training is produced using audio samples of 3 seconds window size with 20% overlap sliding window (identified experimentally to achieve high accuracy in speaker identification). 20 MFCCs were extracted from the 3 second window and their deltas, to create an input vector of 40 in size. The features were used as the input for a basic 1-dimensional convolution neural network (CNN) autoencoder, which was 2 layers deep and made use of PReLU activation. After each convolution layer, we applied Maxpooling to encode the input data, thus the input is reduced by 75% in size. An early prototype autoencoder was produced using a subset of the Voxceleb data set consisting of 40 peoples' voices (we selected voices of all users who's name starts with 'E').

3 RESULTS

We tested the autoencoder's effectiveness against an set of unknown 20 classes with no crossover from train/validation sets. To obtain initial results, we implemented a Euclidean distance classifier, where all samples are compared to a target class and the smallest Euclidean distance is chosen as the predicted class. We produced a box chart of all the distances for speaker *id*10003 compared to 19 other classes, and we can clearly visualize samples of the same true label being clustered together with the target class, and all other samples obtaining a much higher distance metric. This result demonstrates the feasibility of differentiating voices of different people with high level of confidence, using an autoencoder trained with the VoxCeleb dataset.

REFERENCES

[1] Chloë Brown, Christos Efstratiou, Ilias Leontiadis, Daniele Quercia, Cecilia Mascolo, James Scott, and Peter Key. 2014. The architecture of innovation: Tracking face-to-face interactions with ubicomp technologies. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 811–822.

[2] Hande Hong, Chengwen Luo, and Mun Choon Chan. 2016. SocialProbe: Understanding Social Interaction Through Passive WiFi Monitoring. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MOBIQUITOUS 2016)*. ACM, New York, NY, USA, 94–103. https://doi.org/10.1145/2994374.2994387

[3] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612* (2017).

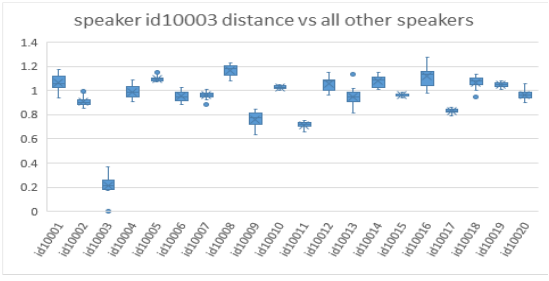

Figure 1: Similarity of voice samples of a single individual against all speakers, after processing through the autoencoder model.