More than just identity: speaker recognition and speaker profiling using the GBR-ENG database

Linda Gerlach[1]*, Finnian Kelly*[2]*, and Anil Alexander*[2]

[1]*Institut für Germanistische Sprachwissenschaft, Philipps-Universität Marburg, Germany*

Gerlach8@students.uni-marburg.de

[2]*Oxford Wave Research Ltd., Oxford, United Kingdom*

{finnian|anil}@oxfordwaveresearch.com

In addition to their spoken content, recordings of speech contain information about the identity of the speaker, along with other speaker characteristics such as gender, language, and accent. Additional speaker characteristics that are more difficult to discern from speech include age, height, and certain health conditions. In recent years, there has been much research interest in exploring the extent to which these various forms of information can be extracted from the voice (often referred to as speaker profiling) (Kelly et al., 2017). In this study, we use a recently recorded database of British English (GBR-ENG, 2019), to estimate how well speaker identities can be modelled using state-of-the-art automatic speaker recognition approaches, and also to explore the extent to which various speaker characteristics can be discerned automatically. Speaker recognition performance was evaluated in two different recording conditions (landline and mobile), using two generations of automatic speaker recognition algorithms within the VOCALISE (Alexander et al., 2016, Kelly et al., 2019) automatic speaker recognition system (namely i-vectors and x-vectors). An automatic speaker profiling approach was then evaluated on the tasks of gender, language, and age estimation.

The GBR-ENG database

The GBR-ENG database contains 6000 telephone recordings from 600 speakers. The recordings consist of 3-6 minutes of English speech, from either a landline or mobile telephone conversation, and were made across three regions in England (North, South, Midlands). Along with the speaker identities, metadata such as speaker gender[1], age, and region of upbringing is also available.

Speaker Recognition Experiments

A subset of 3349 mobile and 2134 landline telephone GBR-ENG recordings (from 534 and 387 speakers respectively) was selected for speaker recognition testing. All possible within and across condition comparisons were evaluated using VOCALISE i-vector and x-vector systems (with PLDA scoring). The x-vector system was observed to provide significant performance gain over the i-vector system in all conditions, achieving Equal Error Rates (EERs) of 0.94%, 1.68%, and 3.30% for landline-landline, mobile-mobile, and landline-mobile conditions respectively (the EERs for the i-vector system were 3-4 times greater). A subset of 236 mobile GBR-ENG recordings from 50 speakers (with no speaker overlap with the test set) was subsequently used to apply condition adaptation (Kelly et al., 2019), resulting in performance improvements for both i-vector and x-vector systems.

Speaker Profiling Experiments

A speaker profiling experiment to estimate speaker gender, age, and language was evaluated using a VOCALISE i-vector based approach. A set of 630 GBR-ENG landline recordings from 118

[1] In this paper, we use the term ‘gender’ to refer to the biological sex of a speaker.

speakers was selected as a training set, with a separate set of 210 GBR-ENG landline recordings from 39 speakers reserved for testing. The speakers in the training set were balanced across gender and dialect region. The median age of the training speakers was 33, with a range of 18—60. A speaker profiling system was trained using several thousand NIST SRE recordings (independent of GBR-ENG), balanced across gender and distributed across an age range of 18—89. A Support Vector Machine (SVM) classifier was trained with i-vectors extracted from this set, along with their gender, language, and age labels. A second speaker profiling system was trained by supplementing the large NIST set with the 630 GBR-ENG training recordings. Both profiling systems were applied to classify the gender of the 210 GBR-ENG test recordings; the independent NIST-only system and the system trained on the merged set of NIST and GBR-ENG recordings both achieved 100% accuracy (using a reduced training set of only GBR-ENG data resulted in an accuracy of 97.14%). Age estimation resulted in a mean absolute error of 7.07 years with the independent training set, and 6.51 years with the merged training set. All of the GBR-ENG speech is in English, with variation in regional accent. Language was identified with an accuracy of 99.52% with the independent training set; however, further classification of English into the three dialect regions was not successful with the language recognition training pipeline. Future work will evaluate the effectiveness of x-vectors within the SVM-based speaker profiling approach, and will consider alternative approaches that exploit DNNs to predict speaker characteristics such as gender, language, accent, age and height directly from features.

References

Alexander, A., Forth, O., Atreya, A. A. and Kelly, F. (2016). *VOCALISE: A Forensic Automatic Speaker Recognition System supporting Spectral, Phonetic, and User-Provided Features.* Odyssey 2016.

GBR-ENG database (2019). *A telephonic speech database collected for the UK Government for evaluating speech technologies.* Further details on application.

Kelly, F., Forth, O., Atreya, A. A., Kent, S. and Alexander, A. (2017). *What your voice says about you – Automatic Speaker Profiling using i-vectors,* International Association of Forensic Phonetics and Acoustics (IAFPA) conference 2017, Split, Croatia.

Kelly, F., Forth, O., Kent, S., Gerlach, L., Alexander, A. (2019). *Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors*. Audio Engineering Society (AES) Forensics Conference 2019, Porto, Portugal.