What your voice says about you: Automatic Speaker Profiling using i-vectors

Finnian Kelly, Oscar Forth, Alankar Atreya, Samuel Kent, and Anil Alexander
Research and Development, Oxford Wave Research Ltd., Oxford, U.K.
{finnian|oscar|alankar|sam|anil}@oxfordwaveresearch.com

In forensic and investigative speech analysis tasks dealing with large volumes of recordings, triage by human experts may not be feasible. The ability to automatically extract information such as a speaker's gender[1], age and spoken language would support the rapid assessment of audio recordings in such cases. Additionally, this information could be used within an automatic speaker recognition framework, to inform the selection of a reference population for example. In this paper, we explore the automatic estimation of speaker gender, age and spoken language from telephone quality speech using an i-vector framework (Dehak et al. 2011a).

In the i-vector approach, a speech segment is converted into a compact, fixed-length representation, in which most of the *important* variability is retained. For speaker recognition, the variability of interest is speaker identity. However, other information carried by the speech signal is also encoded in the i-vector (Dehak et al. 2011a, Bahari et al. 2014, Ranjan et al. 2015).

We use the NIST Speaker Recognition Evaluation (SRE) data 2004-2008 for our experiments. From this large pool of speech, we extracted subsets of multilingual conversational telephone speech, balanced across gender, and containing as broad an age range as possible. The VOCALISE speaker recognition software (Alexander et al. 2014) was used to extract i-vectors from all selected recordings.

Gender Recognition and Age Estimation

A pool of 9000 NIST SRE recordings, balanced across gender and distributed across an age range of 18-89, was divided into train and test partitions in the ratio 2:1. There were 1000 unique speakers across both partitions, with no speaker overlap. After extracting i-vectors for the train set, we trained support vector machine (SVM) models for gender classification and age regression using the known gender and age labels. The SVM models were then applied to generate gender and age labels for the test set. In the case of gender recognition, an equal error rate (EER) of 2.4% and an accuracy rate of 97.7% were obtained. For age estimation, a Mean Absolute Error (MAE) of 7.50 years (males: 8.09, females 7.15) was obtained.

Language Recognition

We considered a set of ten widely-spoken languages for our tests, namely, Arabic, Bengali, Chinese (Mandarin), Chinese (Cantonese), English, Hindi, Japanese, Russian, Spanish and Thai. A pool of 2000 NIST SRE recordings, balanced across gender and each of the languages of interest, was divided into train and test partitions. There were 500 unique

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

speakers across both partitions, with no speaker overlap. We applied linear discriminant analysis (LDA) to the train and test i-vectors based on the 10 language classes to reduce their dimension and enhance separability. To accommodate for gender and regional variations within the languages, we applied a k-means clustering to the train i-vectors. Using Cosine-Similarity based scoring of the train and test i-vectors, we obtained an average language recognition accuracy of 85.05% and a mean EER of 8.22%. Figure 1 contains a visualisation of language i-vectors and a confusion matrix of per-language recognition accuracy rates.

In this work, we have demonstrated how key speaker meta-data such as gender, language and age may be estimated automatically from telephone speech. Initial results indicate that this approach can be successfully extended to unconstrained public sources such as Youtube recordings.

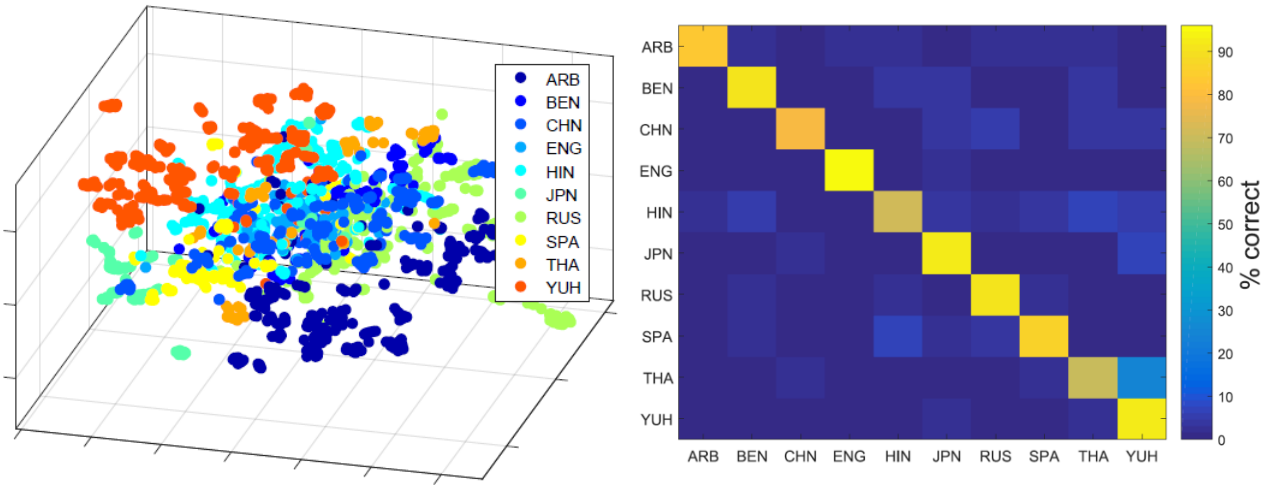

Figure 1: **Left**: An unsupervised t-SNE (van der Maaten and Hinton 2008) projection of the 400-dimensional i-vectors (pre-LDA) into 3 dimensions. Each point is an i-vector and each colour indicates a different language. **Right**: A confusion matrix of language recognition accuracy (the colour-bar indicates accuracy in percent).

References

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2011a). Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788-798.

Dehak, N., Torres-Carrasquilo, P. A., Reynolds, D., and Dehak, R. (2011b). Language recognition via i-vectors and dimensionality reduction. *Interspeech 2011*, Florence, Italy, pp. 857-860.

Bahari, M. H., McLaren, M., Van hamme, H., and van Leeuwen, D. (2014). Speaker age estimation using i-vectors. *Engineering Applications of Artificial Intelligence,* vol. 34, pp. 99-108.

Ranjan, S., Liu, G, and Hansen, J. H. L. (2015). An i-Vector PLDA based gender identification approach for severely distorted and multilingual DARPA RATS data. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) 2015*, Scottsdale, AZ, pp. 331-337.

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. *Speaker Odyssey 2016*, Bilbao, Spain.

van der Maaten, L. J. P. and Hinton, G. E. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605.