Clustering of unknown voices

Hanna Ruch, Andrea Fröhlich & Martin Lory, Zurich Forensic Science Institute

Introduction

A common request in a forensic phonetician's routine may be the following: How many different speakers do the recordings contain? This question may arise, for instance, in cases where the police wiretapped telephone lines to record phone frauds such as the so-called grandparent scam.

The question above can be addressed using the auditory-phonetic approach, by which the different voices on the recordings are analysed auditorily based on a protocol. Following this approach the voices are analysed and described for voice quality, segmental and supra-segmental features, para-linguistic characteristics, pausing behaviour, and syntax, among others. Features are normally compared against a standard variety, a spoken norm, or a neutral voice, and are made note of on a protocol. Voice descriptions are then compared against each other, and voices which share the same or a major number of relevant features are grouped together. A further possibility to cluster voices is using the phonetic-acoustic approach, which was not conducted for the present dataset.

This procedure becomes very time-consuming and expensive in cases with a large number of recordings. In these cases, approaches based on automatic speaker comparison and (statistical) cluster analysis can be an appropriate alternative.

In this paper we present such an approach based on the automatic speaker comparison system VOCALISE (Kelly et al 2019) and a number of different cluster analysis methods. The modelling technique of the newest version of VOCALISE, called xVocalise, is based on deep neural networks, an advanced machine learning technique (for details see Kelly et al 2019). The procedure described in this paper was applied to a real case in which seven voices on five different recordings had to be grouped according to their similarity.

Method

In a first step an auditory-phonetic analysis and auditory clustering were carried out for the seven voice samples. The files were then pre-processed in such a way that each interval contained only one (hypothetically) different voice. Clipping and non-human noise such as ring tones were removed.

The pre-processed (cleaned) recordings were then analysed automatically using VOCALISE (version 2019A-XVector) based on x-vectors, which is the current state-of-the-art approach in automatic speaker comparison. The comparison between all possible recording pairs was conducted using Linear Discriminant Analysis (LDA) and cosine distance (for details see Kelly et al. 2019) and resulted in a cosine similarity matrix.

Statistical analysis was conducted using R (R Development Core Team 2020). The following clustering methods were applied to the similarity matrix:

- Clustered heatmap
- Multidimensional scaling (MDS) and k-means clustering

Additionally, the x-vectors (a high-dimensional representation of each recording, i.e. each voice) were used as input to the following clustering methods:

- Principal Component Analysis (PCA)
- t-distributed stochastic neighbour embedding (t-SNE)

Results

The statistical methods generally confirmed the results of the auditory analysis (three different speakers). As an example, Figure 1 shows the results of the clustered heatmap. The three blueish clusters indicate that there are three groups of similar voices. According to the visualisation, Speaker3 is more similar to Speaker2 than to Speaker1. This result reflects the auditory analysis, where Speaker2 and Speaker3 were also perceived to be more similar.

The results of the different clustering methods also showed that variation within one hypothetical speaker can be considerable. A closer look at the data suggested that not only recording quality but also the communicative situation and dialogue partner affects the result of the automatic system and, ultimately, the goodness of fit of the clustering.

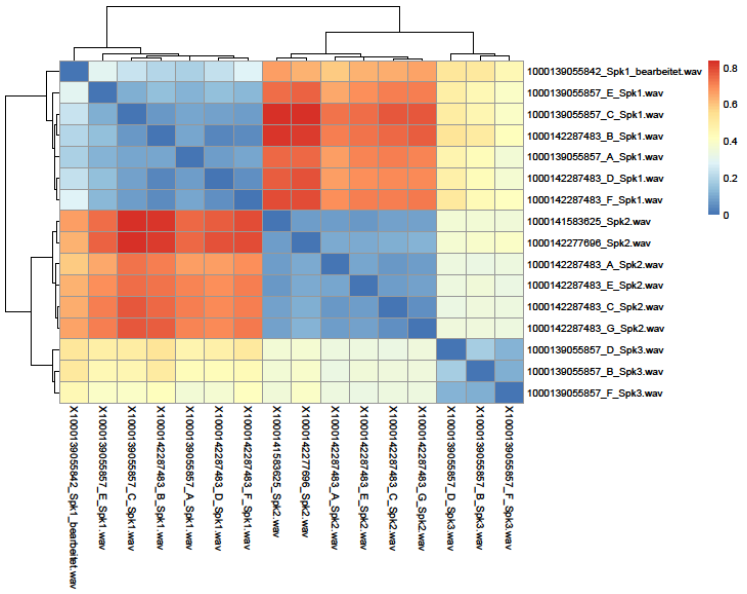

Figure 1: Clustered heatmap showing the cosine distance between all recording pairs. Blueish colours indicate similarity, reddish colours indicate dissimilarity between pairs of voices. The labels Spk1, Spk2, Spk3 stand for the result of the auditory clustering; A-F stand for different intervals (voice samples) within a recording in which two questioned speakers alternated.

Discussion and future directions

For the present dataset and the concrete application in casework the clustered heatmap appeared to be the most appropriate method because it is less abstract and can be explained in a more straight-forward way than the other techniques.

Seven different voices within a dataset is still a rather manageable number, which in could easily be analysed using the auditory approach. Given that the speakers in the dataset are unknown, the results of the two approaches - auditory and automatic/statistical approach - a method validation was not possible. In an on-going project we are currently improving, testing, and validating the methods on a larger, non-forensic dataset with telephone recordings of known speakers. Results are expected to be ready for the AISV conference.

References

Kelly, Finnian, Oscar Forth, Samuel Kent, Linda Gerlach, and Anil Alexander (2019). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors, Audio Engineering Society (AES) Forensics Conference 2019, Porto, Portugal.

R Development Core Team (2020). R: A language and environment for statistical computing. URL: http//www.r-project.org