TUULS of the forensics trade: Assessing the viability of pooling heterogeneous speech corpora for automatic speaker comparison purposes

Dominic Watt, Carmen Llamas*, Peter French*†, Almut Braun* and Duncan Robertson**
**University of York / †JP French Associates*
dominic.watt | *carmen.llamas* | *peter.french* | *almut.braun* | *d.robertson@york.ac.uk*

The Use and Utility of Localised Speech Forms in Determining Identity (TUULS; ESRC ES/M010783/1) project has two key strands: one sociophonetic, one forensic. The sociophonetic part seeks to uncover patterns of phonetic variation in three cities in north-east England (Newcastle, Sunderland, Middlesbrough), according to speakers' sex, age, and level of routinised geographical mobility, a measure of the amount of contact they have with other localities through e.g. travelling to work or for leisure purposes (Britain 2016). Owing to their relatively limited levels of such contact, we hypothesise that less mobile individuals are more likely to retain conservative, locality-specific pronunciations. More mobile individuals, by contrast, will be harder to locate geographically via their speech because it exhibits fewer narrowly localised forms.

In this paper we discuss the implications of these pronunciation differences for the two main types of work in forensic speech science. The first is speaker profiling. Here, we examine the speech of an unknown individual recorded committing an offence so as to help the authorities identify a suspect by narrowing the pool of potential perpetrators. Knowing how speech patterns vary across small geographical areas is crucial to this task. The second, and principal, type of work undertaken by forensic speech scientists is speaker comparison. This involves assessing the similarities and differences between speech recordings, most commonly a recording of a crime in progress (e.g. a threatening phone call) and a recording of a suspect in police interview. Alongside assessing the samples for *similarity*, we must also consider the *typicality* of these resemblances in the relevant population. The similarity and typicality statistics we use can be most objectively computed using detailed and up-to-date reference databases. Regrettably, existing resources are often limited by their small scale, the spottiness of their geographical coverage, and/or by being outdated.

Automatic Speaker Recognition (ASR) systems are seeing increasing use in forensic speaker comparison work. Based as they are on distances between spectral coefficients which bear only indirect relation to the segmental properties of accents with which human analysts tend to concern themselves, it is reasonable to suppose that ASR algorithms are largely indifferent to speaker accent in the sociophonetic sense. What, then, is to stop us from pooling data sets, irrespective of the accents represented in each database? Does the performance of the ASR system suffer when corpora of highly dissimilar accents, e.g. SSBE and Newcastle English, are combined (cf. Hughes & Foulkes 2017)? In our presentation we demonstrate some of the effects, and consider the implications, of merging heterogeneous speech corpora in this way.

References

Britain, D. (2016). Sedentarism and nomadism in the sociolinguistics of dialect. In Coupland, N. (ed.). *Sociolinguistics: Theoretical Debates*. Cambridge: CUP, pp. 217-241.

Hughes, V. & Foulkes, P. (2017). What is the relevant population? Considerations for the computation of likelihood ratios in forensic voice comparison. *Proceedings of Interspeech 2017*, Stockholm, Sweden, August, pp. 3772-3776.