Forensic phonetics and speaker characteristics

Francis Nolan

The forensic task which phoneticians are most generally asked to perform is speaker identification. Usually the police want to know if a suspect is the speaker on an incriminating recording. Given the plasticity of the voice, and the variability of the relationship between speech and the 'machine' producing it (the speaker), phoneticians often shy away from identification and prefer to talk in terms of speaker comparison. Whatever the terminology, the holy grail of research in this area is a set of properties which are recoverable from the (often degraded) speech signal and which vary more between speakers than within the speech of one individual.

During its early forensic application, in the third quarter of the twentieth century, phonetic speaker comparison was essentially an extension of dialectology, with a focus on the auditory analysis of consonants and, more particularly vowels – given the latter's greater robustness in telephone transmission. A lack of appreciation of the potential for perceptual equivalence in acoustically distinct signals, and of the size of relevant populations, undoubtedly led to an overestimation of the reliability of the method. As acoustic phonetic analysis became commonplace, the emphasis in speaker comparison shifted in the last quarter of the century to the 'auditory-acoustic' method, combining the trained ear with measurements of parameters such as fundamental frequency and vowel formants. The two approaches provide, to an extent, complementary information. The auditory-acoustic method remains the dominant paradigm in many jurisdictions, including the UK and Germany. But in this century, a third era has arrived with the availability of powerful techniques for automatic speaker verification/identification, relying largely on spectral information resulting from the vocal tract and coded (standardly) as Mel Frequency Cepstral Coefficients (MFCCs). Under ideal conditions, and with an appropriate reference population to draw on, these methods can estimate the likelihood that the incriminating voice sample was produced by the suspect as opposed to another speaker in the population.

The papers in this Discussant Session have in common that they look beyond the usual sources of speaker specific information, whether in perceived pronunciation, the acoustic cues (such as formants) used to characterise linguistic categories, or the vocal-tract related MFCCs of automatic methods. French et al. explicitly compare both MFCCs and an alternative spectral characterisation in terms of long term formant distributions (LTFDs) with a componential auditory analysis of voice quality. The latter 'VPA' analysis is based on Laver's [1] articulatory settings. Only supralaryngeal settings are included. The MFCC and LTFD analyses are shown to be highly correlated and similar in speaker-discriminatory power; the auditory VPA analysis less correlated, and somewhat less powerful. It is suggested that the VPA analysis may be capturing partially complementary information about the vocal tract.

Chan investigates the speaker characterising potential of tonal patterns in Mandarin and in Cantonese. The tones of the former, with its less crowded tone system, in general provide better discrimination of speakers. If f0 level is normalised, contour shape retains appreciable discriminative power. Additionally, Chan takes account of the effect of tonal environment, showing that coarticulation to 'conflicting' tones on either side of the target syllable causes a significant reduction in discrimination in Cantonese, but not in Mandarin.

With the paper by Kolly et al. we are reminded that the act of speaking involves more than the words spoken. They look not at episodes of speech but at the bits in between – the pauses. It is shown that as a speaker switches between native Swiss German and L2 English and French there is a degree of consistency in the number and duration of pauses in read sentences, as well as variation between speakers. Along somewhat

similar lines, Braun and Rosin show that there is considerable variation between speakers in the distribution of different types of hesitation phenomena, and that the distribution shows stability across recordings of spontaneous speech on three different days.

All four of these papers remind us of the complexity of the speech event, and the multiplicity of ways in which speaker-specific information may be imprinted on it. Speech is both cognitive and physical, and is also crucially listener-facing, and so subject to complex constraints. The Gestalt of a 'voice' is multifaceted, as is the variation within it. These papers highlight individuality in less obvious, less researched features of the voice. The major challenge will be to find ways of exploiting these features in speaker comparison, whether within the auditory-acoustic method of speaker comparison, or as an adjunct to automatic methods.

Reference

[1] Laver, J. (1980) The Phonetic Description of Voice Quality. Cambridge: Cambridge University Press