Exploring the strength of evidence of long-term formants in bilingual speakers

Didier Meuwly\(^1\)\(^,\) Willemijn Heeren\(^3\) and Annabel Bolck\(^1\)

\(^1\)Netherlands Forensic Institute, WISK, The Netherlands
\(^2\)University of Twente, EWI, SCS, The Netherlands
\(^3\)Leiden University Centre for Linguistics, Leiden University, The Netherlands

d.meuwly@utwente.nl; a.bolck@nfi.minvenj.nl
w.f.l.heeren@hum.leidenuniv.nl

Introduction

In forensic speaker recognition the phonetic acoustic approach is used to extract and compare speaker dependent features from audio recordings (trace and reference specimens). Heeren et al., (2014), for example, have presented a statistical analysis describing the intra and intervariability of the long-term formants F2 and F3 for 12 bilingual (Dutch/Turkish) male speakers. These results were obtained for speakers selected from a corpus of forensic utterances, the NFI-FRITS database. The forensic contribution of the long-term formants has been studied for different languages (Rose, 2006; Zhang et al. 2008) and using different approaches (Becker et al. 2008, Gold et al. 2013). But the availability of data from a forensic corpus of bilingual male speakers offers an opportunity to explore the strength of evidence of the long-term formants F2 and F3 for this condition.

Method

In forensic science the evaluation of the strength of evidence is based on the Bayes Theorem of conditional probabilities. It is computed as a likelihood ratio (LR), the ratio of the probability of the observations in the light of a pair of mutually exclusive alternative hypotheses H1 and H2 (Evett, 1998). In our case the strength of evidence of the long-term formants F2 and F3 is computed for the hypotheses H1: the utterances from which F2 and F3 are extracted originate from one single male bilingual speaker (Dutch/Turkish) and the hypothesis H2: the utterances from which F2 and F3 are extracted originate from two different bilingual male speakers (Dutch/Turkish).

Different approaches are described in the literature to compute LRs. They can be classified as feature- or score-based approaches (Lindley, 1977; van Leeuwen and Brümmer, 2007). When a large number of features are extracted from each utterance score-based methods are frequently used as a way to reduce the multi-dimensional feature space and estimate the model parameters in one dimension (Gonzalez-Rodriguez, 2006). When a limited number of features is extracted from each utterance, like the phonetic acoustic features F2 and F3, a feature-based approach allows for a better estimation of the model parameters with a limited amount of data (Bolck et al., 2009).

The ‘LR wizard’ is a statistical tool developed at the Netherlands Forensic Institute to compute both score and feature-based LRs. This tool offers parametric and non-parametric methods to model univariate and multivariate parameters, and several graphical representations like the Tippett plot and the Empirical Cross Entropy (ECE) plots are available to describe the performance of the method and the calibration of the LRs computed. The method used to perform the experiments is described in (Bolck et al., 2009). It is a source-independent approach, the mean is source-specific but we have not used a covariance matrix depending on all sources.
Results and discussion

The LR wizard has been used to compute the strength of evidence for F2 and F3 from utterances in Dutch only, in Turkish only and when no condition is applied to the language. The performance of the method is presented in Figure 1.

<table>
<thead>
<tr>
<th>Tippett plot</th>
<th>Dutch</th>
<th>Turkish</th>
<th>Both languages</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
<td><img src="image3.png" alt="Graph" /></td>
<td><img src="image4.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

**Figure 1** Tippett plots describing the strength of evidence for F2 and F3 from test and reference specimens from the same source in Dutch only, in Turkish only and when they are in different languages.

The results show a quantifiable strength of evidence (higher than 1) when the trace and reference specimens are from the same language (Dutch or Turkish) but not when they are in different languages. The current results do not provide evidence to support the hypothesis that, for bilingual male speakers (Dutch or Turkish), F2 and F3 are speaker dependent features when no condition is set on the language. Currently our best explanation is that the strength of evidence is weak when the trace and reference specimens are from the same language (Dutch or Turkish), and it becomes even weaker when they are from different languages because the language has an influence on the F2 and F3. The variation of the features becomes larger over the languages, reducing the strength of evidence. In the future the data will be further investigated with a feature-based approach and investigated with a score-based approach in order to improve and compare the performance and calibration.

References


