Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 58 (2014) 81–90

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Likelihood ratio calculation for a disputed-utterance analysis with limited available data

Geoffrey Stewart Morrison [a,*], Jonas Lindh [b], James M Curran [c]

[a] Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, UNSW Sydney, NSW 2052, Australia

[b] Division of Speech and Language Pathology, Department of Clinical Neuroscience and Rehabilitation, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Box 452 SE405 30, Gothenburg, Sweden

[c] Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

Received 5 March 2013; received in revised form 13 November 2013; accepted 14 November 2013
Available online 25 November 2013

Abstract

We present a disputed-utterance analysis using relevant data, quantitative measurements and statistical models to calculate likelihood ratios. The acoustic data were taken from an actual forensic case in which the amount of data available to train the statistical models was small and the data point from the disputed word was far out on the tail of one of the modelled distributions. A procedure based on single multivariate Gaussian models for each hypothesis led to an unrealistically high likelihood ratio value with extremely poor reliability, but a procedure based on Hotelling's T^2 statistic and a procedure based on calculating a posterior predictive density produced more acceptable results. The Hotelling's T^2 procedure attempts to take account of the sampling uncertainty of the mean vectors and covariance matrices due to the small number of tokens used to train the models, and the posterior-predictive-density analysis integrates out the values of the mean vectors and covariance matrices as nuisance parameters. Data scarcity is common in forensic speech science and we argue that it is important not to accept extremely large calculated likelihood ratios at face value, but to consider whether such values can be supported given the size of the available data and modelling constraints.

Keywords: Forensic; Likelihood ratio; Disputed utterance; Reliability; Hotelling's T^2; Posterior predictive density

1. Introduction

In forensic speech science the amount of data available from casework samples is often limited. In forensic-speech-science research reports and presentations, and in forensic-speech-science casework reports, we have sometimes seen very large likelihood-ratio values presented, with those values being derived from calculations based on small amounts of data. For reasons we will explain in the present paper, we think that the likelihood-ratio values reported could not be supported by the small amount of data available, but it appears that the authors were not aware of this problem and simply reported the calculated values. In the present paper we illustrate this problem using data from an actual disputed-utterance case on which one of us worked. We chose this example as an extreme example in which the problem should be obvious, but one should also be aware that, because by necessity likelihood ratios in forensic speech science often have to be calculated on the basis of small amounts of data, the problem may also have a substantial impact on calculated likelihood-ratio values in cases where it is not so obvious. The same problem may also occur in other branches of forensic science.

In a disputed-utterance analysis the task of the forensic scientist is to calculate the strength of evidence in the form of a likelihood ratio answering the question: How probable are the observed acoustic properties of the disputed utterance if the speaker had said what the prosecution

* Corresponding author.
E-mail address: geoff-morrison@forensic-evaluation.net (G.S. Morrison).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.11.004

claims they said versus if the speaker had said what the defence claims they said (Morrison and Hoy, 2012). The present paper describes the calculation of a forensic likelihood ratio on the basis of relevant data, quantitative measurements, and statistical models.[1] The data are taken from an actual case in which the amount of data available to train the models was limited. An initial analysis using single-Gaussian models is described, and the problem with this approach in this case is discussed. The problem is due to the fact that not only is there a small amount of training data but the data point for the disputed utterance is also far out on the tail of one of the modelled Gaussian distributions. Revised analyses are then described. One attempts to take account of the sampling uncertainty of the mean vectors and covariance matrices via the use of a Hotelling's T^2 distribution. The other makes use of a Bayesian posterior-predictive-density analysis which integrates out the values of the mean vectors and covariance matrices as nuisance parameters.

Copies of the data and the Matlab (Mathworks, 2010) and R (R Development Core Team, 2013) scripts used to make the calculations described in this paper are available from http://geoff-morrison.net/#DispUtLimDat.

2. Background to the case

In 2008 one of the authors of the current paper, JL, was asked to perform a disputed-utterance analysis in a Swedish murder case [n° B1293-07 of Hovrätten för nedre Norrland, 2008-02-26]. A word on an audio recording of a police interview with an eye witness (a female Swedish speaker) was disputed as being either the pronoun "dom" [dɔm] (*they*) or the name "Tim" [tʰɪm]. The recording also contained 29 undisputed tokens of "dom" and 16 undisputed tokens of "Tim" spoken by the same speaker. The fact that all the tokens came from the same recording meant that there were no recording-channel or speaking-style mismatch issues.

The original speaker was not cooperative at the point in time when the disputed-utterance analysis was performed, and no other recordings of this speaker were provided, hence no additional data from this speaker could be used in the analysis.

3. Acoustic analysis

In 2008 JL measured the voice onset time (VOT) and first and second formant (F1 and F2) values in the disputed token and each of the tokens of the undisputed words ("dom" and "Tim"). The recording had been made on an analogue tape and was digitised with a sampling rate of 16 kHz and 16 bit quantisation. Details of the recording equipment were not provided at the time and cannot be ascertained now. JL subjectively characterised the recoding quality as poor but not extremely bad. The signal to noise ratio was 28 dB.

Measurements were made using Praat (Boersma and Weenik, 2008). The VOT of each token was measured three times (analyses below will be based on the mean of the three VOT measurements for each token). Prevoicing and plosive bursts were clearly visible on a waveform display. No substantial speaking rate differences were observed during the relevant potions of the recording.[2] F1 and F2 were measured at a single point in the middle of the vowel in each token.[3] Formants were measured using the Burg autocorrelation linear-predictive-coding (LPC) algorithm (Anderson, 1978). The resulting formant values were overlayed on a broadband spectrogram and the maximum frequency below which to search for formants adjusted if the initial results were obviously erroneous. Otherwise the settings recommended in the Praat documentation for female speakers were used.[4]

[1] We work in a paradigm which requires the use of the likelihood-ratio framework, has a strong preference for the calculation of likelihood ratios on the basis of relevant data, quantitative measurements, and statistical models, and requires empirical assessment of the degree of validity and reliability of the analytical procedures under conditions reflecting those of the case under investigation (we will refer to this as the "new" paradigm). Although numerous papers have been published describing the new paradigm and describing approaches to forensic voice comparison conducted within the new paradigm, as far as we are aware Morrison and Hoy (2012) is the only previously published paper on disputed utterances conducted within this paradigm. The arguments in favour of the use of the new paradigm for disputed-utterance analysis are made in Morrison and Hoy (2012) and are not repeated here since this is not the focus of the present paper. The present paper assumes the new paradigm and addresses a problem which only arises if one's approach is based on data and statistical models. This problem is the focus of the present paper. We think that this is an important problem which potentially affects not just disputed-utterance analysis, but also forensic voice comparison and other branches of forensic science. The disputed-utterance case we describe in this paper happens to provide a good example of the problem. Our primary objective is to make researchers and practitioners aware of this problem and our secondary objective is to describe some potential solutions. The old paradigm for disputed-utterance analysis is based on listening to the speech on the audio recording and making a subjective judgement – the problem which we address in the present paper does not arise in that paradigm.

[2] Swedish has devoicing of voiced plosives following voiceless consonants, e.g., in "samt dom" *and them*. Only phonetically voiced tokens of [dɔm] were analysed (and included in the count of 29 tokens). The disputed utterance was not in a context which would result in devoicing.

[3] Note that the units of analysis are the words "dom" [dɔm] and "Tim" [tʰɪm], not the vowels [ɔ] and [ɪ] – any nasalisation effect on the vowel due to the following bilabial nasal is therefore automatically included in the formant measurements. Likewise for any differential reduction effect in the personal pronoun "dom" versus the proper noun "Tim".

[4] Recommended settings in the Praat documentation for female speakers for "Sound: To Formant (burg)...": Maximum frequency 5500 Hz. Maximum number of formants 5 (i.e., 10 LPC coefficients). Window length 0.025, "Praat uses a Gaussian-like analysis window with sidelobes below −120 dB.... the actual Gaussian window duration is 0.050 seconds. This window has values below 4% outside the central 0.025 seconds, and its frequency resolution (−3 dB point) is 1.298/(0.025 s) = 51.9 Hz" (Praat manual). Pre-emphasis 50 Hz "frequencies below 50 Hz are not enhanced, frequencies around 100 Hz are amplified by 6 dB, frequencies around 200 Hz are amplified by 12 dB, and so forth" (Praat manual).