Conversational Style Mismatch: its Effect on the Evidential Strength of Long-term F0 in Forensic Voice Comparison

Phil Rose[1], *Cuiling Zhang*[2,3]

[1]Australian National University Emeritus Faculty, [2]School of Criminal Investigation, Southwest University of Political Science & Law, Chongqing, China, [3]Chongqing Institutes of Higher Education Key Forensic Science Laboratory, Chongqing, China.

cuilingzhang@hotmail.com, philjohn.rose@gmail.com

Abstract

We describe a speaker verification experiment to investigate the effect of mismatch in conversational style on the strength of evidence furnished by long-term fundamental frequency in forensic voice comparison. Non-contemporaneous recordings of informal conversations, simulated police interrogations, and information exchanges from 90 male Chinese speakers were compared within the likelihood ratio framework. Evaluation with C_{llr} and error rates shows rather poor strength and weight of evidence for matched comparisons, which degrades still further with mismatched comparisons. This suggests that long-term F0 should only be used forensically in conjunction with other features.

Index Terms: forensic voice comparison, likelihood ratio, long-term F0, Chinese, validation, weight of evidence

1. Introduction

In forensic speaker identification, the expert typically compares questioned and known voice samples to help determine whether the questioned voice has come from the known speaker. Usually the questioned sample is from an offender and the known sample from a suspect, and the beneficiary is a fact-finder (judge or jury), an investigating authority (police), or legal representative [1]. Although not included in the most recent general report on the scientific validity of feature comparison methods in forensic science [2], technical forensic speaker identification also typically relies on such *feature comparison* methods [3], whether the features be the mel-frequency cepstral coefficients of automatic approaches [4], or more transparent, but less powerful, acoustic-phonetic properties like formant frequencies in so-called forensic-semi-automatic speaker recognition [5].

Features used to help identify voices forensically should ideally be common and easy to extract, and relatively immune to channel distortion. But most importantly, of course, the feature must be demonstrably effective in discriminating same-speaker speech samples from different-speaker speech samples, and have been shown to do so under the conditions of the forensic case in which they are being used:

> Without actual *empirical* evidence of the ability of a forensic feature-comparison method to produce conclusions at a level of accuracy appropriate to its intended use under circumstances reasonably related to this use, an examiner's conclusion that two samples are likely to have come from the same source is *completely meaningless*." [6, pp. 1-2].

One popular acoustic-phonetic feature in forensic semi-automatic speaker recognition – its mean and standard deviation values reportedly used by 94% and 72% of forensic voice comparison experts world-wide [7] – is fundamental frequency (F0), the acoustic reflex of the rate of vibration of the vocal cords. This is because of promising results in early speaker recognition research; and also because F0 is (relatively) easily measurable and there is usually lots of it. It is relatively immune to channel distortion (for although the fundamental and H_2 might be attenuated by phone transmission, many harmonics remain in higher frequency ranges to permit estimation of the so-called missing F0. The use of intonational F0 in a real case, and equally importantly the validity of the method, is documented in [5].

As well as the many linguistic uses of F0, which encodes tone, intonation and stress, many non-linguistic factors, like state of health, are also known to affect it. This multiplicity of factors has an adverse effect on its between- to within-speaker variance ratio by increasing the latter. Since the inherent strength of forensic speaker recognition features relies primarily on their ratio of within- to between-speaker variance, one would not expect particularly good strength of evidence (SoE) from global F0 properties, and this has been demonstrated in several studies, e.g. [1,8,9]. These studies have, however, also used arguably ecologically less than valid material – or at least material less likely to occur in real cases. For example, contemporaneous recordings were used in [8], and monologs of varying, but atypically long, duration in [1, 9]. Such conditions also have the potential to overestimate the SoE of uncontrolled global F0. Finally, one relationship between suspect and offender recordings which is commonly found in real-world case-work, but which does not seem to have been tested, is mismatch in formality between the conditions under which the suspect and offender voice recordings are obtained. Typically, the suspect's recordings are taken from a formal police interview, whereas the offender's recordings are from informal conversational exchanges. This paper's aim is to test, within a likelihood ratio framework, how well F0 from natural speech performs, in particular under these mismatched conditions.

The likelihood ratio (LR) framework [10] was recently endorsed as best practice in forensic automatic and semi-automatic speaker recognition by the *Board of the European Network of Forensic Science Institutes*, representing 58 laboratories in 33 countries [11]. As far as this paper's aim is concerned, the LR framework has two merits. Demonstrating that a forensic speaker identification method actually works is called *validation*. The LR framework allows a system to be validated in a forensically realistic manner [12], and the discriminability of forensic speaker recognition systems has in fact been tested with it now for nearly two decades. Secondly, a likelihood ratio also quantifies the SoE of a particular feature or system, and from this, as will be shown, an estimate of the weight of evidence, and expected weight of evidence, can also

easily be derived [13].

2. Procedure

2.1. Database

We used a database of 90 male Chinese speakers recorded in 2011 for the purpose of aiding forensic voice comparison research and practice. The speakers were recruited from the *Chinese Police College of Criminal Investigation* 中国刑事警察学院 in Shenyang and all spoke varieties of North-Eastern Mandarin. The database was structured according to the protocol in [14], which was designed to elicit realistic (i.e. non-contemporaneous, natural speech) recordings for testing forensic voice comparison. Speakers were recorded on two occasions separated by about a month. Three different conversational tasks were recorded in sound-proofed rooms at 44.1 kHz and 16 bit resolution: a conversation, a simulated police interview, and a co-operative exercise where both speakers were given different copies of a badly transmitted fax and had to work out its contents. For the conversation and the fax tasks, speakers were paired, and communicated on a landline phone while being recorded on separate channels using lapel mikes. For the interview, each speaker was interviewed separately by a research assistant with internship experience in interrogating suspects, and was again recorded on a lapel mike. Thus there were six recordings per speaker, labeled C1 (for *conversation recording* 1), I1 (*interview recording* 1), F1 (*fax recording 1*), and C2, I2, F2.

2.2. Front-end

For convenience in extracting F0, the quiescent portions of each speaker's recordings were removed, and the remaining non-silent portions saved as separate short .wav files using the 'sound file cutter-upper' *Matlab* code [15]. A *Praat* script was written to cycle through and inspect each putative vocalization and downsample it to 8k (to eliminate fricative and aspiration noise which can affect automatic F0 extraction). The inspection was used to estimate appropriate settings for the F0 extraction (which were between 30 Hz and 300 Hz) and a voicing threshold (0.3). *Praat* extracted F0 with these settings every centisecond, using autocorrelation.

Not all non-silent portions of the recordings were deliberate utterances, of course. Visual inspection also showed that some of the short .wav files contained extraneous noise with artefactually extracted F0, or brief hesitation phonation with badly extracted F0, and these were excluded by rejecting any short .wav file with a duration less than 35 csec. In Chinese speech, the shortest utterances (usually a monosyllabic CV or CVC word) will probably be slightly less than this; but it is obviously more important to exclude F0 measurements from non-speech than include all speech. It was also noted that longer stretches of speech sometimes contained hesitation phenomena a little before or after a longer utterance, with badly extracted F0 between the utterance and the hesitations. Such examples were excluded by ensuring that the extracted F0 consisted of at least 60% of the overall duration of the portion.

R code was then written to retain only extracted F0 values, and combine the remaining portions into a single file from which a density could be estimated. Speakers differed slightly in the amount they spoke during their tasks. Consequently, differing amounts of voiced speech were obtained, ranging from ca. 60 to 630 seconds, with a mean amount of net voiced speech per speaker of ca. 220 seconds. The different tasks also resulted in different amounts of net voiced speech, with the interview having the most (mean = ca. 280 sec.) and the fax the least (mean = ca. 150 sec.). A small difference also obtained between the two non-contemporaneous recordings, with the second showing a slightly narrower range.

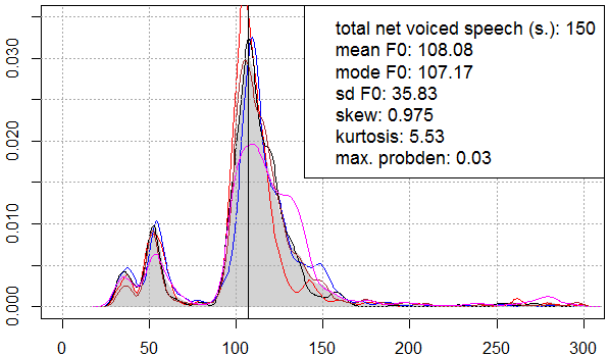

Figure 1: *Empirical kernel density distribution of long-term F0 from speaker 35's C1 recording. Coloured lines = 30 sec. replicates. X-axis = F0 (Hz), y-axis = probability density. Vertical line = modal F0.*

Each speaker's F0 data for each recording was then divided into 30-second chunks to act as replicates from which within-speaker variance could be estimated – an important part of the procedure for estimating multivariate likelihood ratios. Figure 1 shows data thus prepared for a single speaker. The grey distribution represents a kernel density for the 15000 F0 measurements from his 150 sec. voiced speech. The coloured lines show the distributions of the five different 30 sec. replicates from this. The majority of the speaker's F0 values are concentrated in a typically mildly positively skewed distribution with a modal value of 107 Hz. The speaker also has quite a lot of much lower F0 values centered at about 50 Hz which come from creaky phonation. Quite a lot of speakers show this bimodality.

2.3. Parameterisation

Previous LR-based discriminations using LTF0 distributions [1,9] have used six parameters from the F0 distribution: mean, mode, standard distribution, skew, kurtosis and maximum probability density. These were separately estimated from each speaker's 30-second replicates. One of the advantages of the MVLR formula used in this analysis is to be able to take the inevitable correlations between such parameters into account, and a principle components decorrelation, which will result in slightly degraded performance, was considered otiose.

2.4. Back-end

The multivariate kernel-density (MVKD) likelihood ratio formula [16] was used to compare the parameterized acoustics from each speaker's first recording with their second recording, and with the parameters of the other speakers' first recordings to get scores for same-speaker and different-speaker comparisons. (Only one set of different-speaker comparisons was used: between a first speaker's first recording and a second speaker's second.) The scores were then converted to likelihood ratios with logistic regression calibration using the *Focal* tool-kit [18]. A leave-one-out cross-validation was used as a strategy against overfitting,

whereby the test data were removed from the reference data for the estimation of the covariance matrices. A leave-one-out cross-validation was also used in the estimation of the logistic-regression coefficients for calibration.

The performance of an LR-based detection system like this, equivalently its validity, is currently assessed by the information-theoretic log likelihood ratio cost C_{llr} [19]. C_{llr} relates to the average amount of information the system provides to its end-user. Positive C_{llr} values below unity – the smaller the C_{llr} the better – indicate that the system has the capability of reducing the user's uncertainty in the hypothesis. Error rates for same- and different-speaker comparisons were also calculated.

With three different tasks, six different comparisons are possible: three with matched tasks, where speakers are compared using data from the same task, e.g. fax - fax; and three mismatched tasks, with comparisons based on speech from different tasks, e.g. fax - interview. For each of these mismatched tasks, two comparisons were possible from reversing the task (i.e. fax – interview and interview – fax). Nine comparisons were thus made. We are above all interested in seeing how well F0 performs with the realistically constituted data in the mismatched condition when conversation speech is compared with interview speech, which is the common type of comparison between offender and suspect in forensic reality.

3. Results

Table 1 gives the results – C_{llr}s and error-rates – for the three matched and three unmatched comparisons. The *expected weights of evidence* in the right column are explained in section 4. It can be seen first of all that, with the exception of the fax-fax comparison, all C_{llr} values are high, ranging from 0.75 to 0.87. This indicates that the long-term F0 is on average providing some information to reduce the user's uncertainty, but not much. The better performance of the fax-fax comparison (C_{llr} = 0.53) might be due to factors contributing to smaller within-speaker variation like more constrained subject matter, or less emotion. Finally, it can be seen that mismatched conditions do result in less information, with the forensically most relevant comparison, between conversation and interview, having one of the two worst performances (C_{llr} = 0.87).

Table 1. *Results. ER = error rate (%), SS/DS = same-/different-speaker comparison, EWoE= expected weight of evidence (decibans), conv = conversation, int = interview.*

comparison	C_{llr}	ER_{SS}	ER_{DS}	EWoE
matched				
conv-conv	0.79	16.7	36.5	2.0
int-int	0.75	15.6	33.5	2.4
fax-fax	0.53	11.1	23.5	4.2
mismatched				
conv-fax	0.87	20	43.5	1.2
fax-conv	0.80	14.4	37.7	2.3
conv-int	**0.87**	**18.9**	**42.1**	**1.4**
int-conv	0.85	20	40.5	1.4
int-fax	0.86	21.1	42.0	1.2
fax-int	0.85	21.1	41.7	1.3

Figure 2 shows the Tippett plot for the forensically relevant conversation-interview comparison. Same-speaker LRs increase towards the right; different-speaker LRs towards the left. The slight rightwards displacement of the equal error-rate point, which of course results in exaggeration of the different-speaker error-rate and attenuation of same-speaker error-rate, is typical but mysterious and may be related to the logistic-regressive nature of the calibration.

It can be appreciated that figure 2 depicts a feature with a very poor strength of evidence: the maximum $\log_{10}$LR observed for same-speaker comparisons was 0.46 – a LR of ca. 3 – and the mean same-speaker LR was 1.7. Even under the most advantageous conditions for the prosecution, i.e. with evidence of maximum strength and suspect just one of two who could have said the incriminating speech, the posterior probability would be just ca. 75% in favour of the suspect being the unknown speaker. Given less favourable priors – say suspect one of five possible perpetrators – and the average LR for the comparison, the posterior would be [1.7 /(1.7+4) =] ca. 30%, suggesting rational belief in a different-speaker hypothesis.

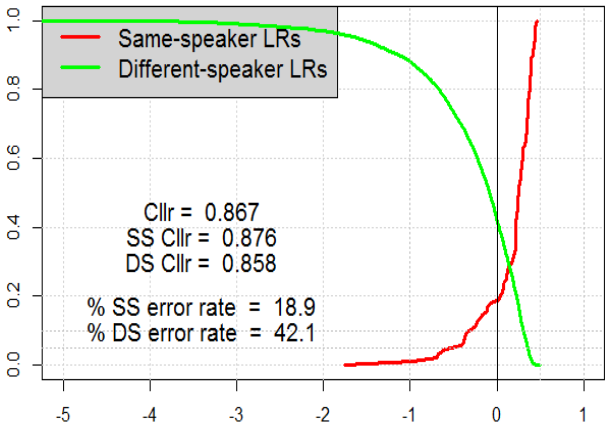

Figure 2: *Tippett plot for 90 speakers' conversation-interview comparison. X-axis = $\log_{10}$LR greater than ..., y-axis = cumulative proportion of different-speaker comparisons ~ 1-cum.prop. of same-speaker comparisons.*

4. Strength and Weight of Evidence

In English, the two properties commonly metaphorically predicated of evidence – *strength* and *weight* – both highlight the important fact that evidence can have different, continuously varying values (another metaphor!). The strength of evidence E in favour of a proposition H like *these two speech samples have come from the same speaker* is properly measured by the likelihood ratio: the ratio of the probability of getting E assuming H is true to the probability of getting E assuming it is not: $P(\mathrm{E}|\mathrm{H})/P(\mathrm{E}|\sim\mathrm{H})$ [10, pp.95ff.]. The odds form of Bayes' theorem at (1) makes it clear that the strength of evidence is also the Bayes' Factor, by which the prior odds in favour of the hypothesis [O(H)] have to be multiplied to get the posterior odds in favour of the hypothesis once the evidence is adduced [O(H|E)] [13, p.88].

$$\frac{O(\mathrm{H}|\mathrm{E})}{O(\mathrm{H})} = \frac{P(\mathrm{E}|\mathrm{H})}{P(\mathrm{E}|\sim\mathrm{H})} \qquad (1)$$

The additive property of evidence suggested by *weight* can be modeled by simply taking the logarithm of the Bayes' Factor [13, p.89]. Turing suggested a base of 10, calling the unit a *ban,* with a *deciban* being the smallest change in weight of evidence we can conceptually process [13, p.90; 20, p.92]. For example, the maximum same-speaker Bayes Factor of ca. 3 for the conversation-interview comparison in figure 2 has a weight of about [$10*\log_{10}3$ =] 4.7 *db* in favour of it being the

same speaker. Given flat priors (i.e. 0 *db*), a LR of 3 from a comparison of LTF0 from suspect's conversations and offender's Police interview should rationally tip the scales a little in favour of your belief that the same speaker was involved. But the average LR has a hardly noticeable weight of only ($10*\log_{10}1.7 =$) 2.3 db: you would hardly notice the scale moving!

As with C_{llr}, an *overall* evaluation of weight is perhaps more informative than evaluation for individual Bayes' Factors. This can be captured with an *expected weight of evidence* in favour of the same-speaker hypothesis ($EWoE_{ss}$). This can be conventionally estimated as at (2) from the product of the probability of the two mutually exclusive outcomes and their respective weights [13, p.91].

$$EWoE_{ss} = P(\mathrm{LR} > 1|H_{ss}) * 10 * log_{10}\left[\frac{P(LR > 1|H_{ss})}{P(LR > 1|H_{ds})}\right] - P(\mathrm{LR} < 1|H_{ss}) * 10 * log_{10}\left[\frac{P(LR < 1|H_{ss})}{P(LR < 1|H_{ds})}\right] \quad (2)$$

The expected weights of evidence from the various comparisons were given in table 1. With the exception of the fax-fax comparison they are decidedly light.

5. Summary and conclusion

This paper has shown that strength of forensic speaker recognition evidence from LTF0, already weak when conditions are matched, becomes even weaker under mismatched conditions. The expected weight of evidence of systems seeking to extract forensically useful information from comparisons of LTF0 was shown to be very light indeed. LTF0's demonstrated weakness as an acoustic phonetic parameter – at least when quantified as in this paper – does not appear to warrant its popularity, at least when considered on its own.

The poor strength of evidence from LTF0 – and it should be remembered that the parameters were obtained from quite a lot of speech – indicates that it can at the moment at best only be used in conjunction with other features. Perhaps its strength can be improved by better modeling of the LTF0 distributions. One might for example separate the creaky values from the modal, and model the latter with an appropriate non-normal (log-normal? Gamma?) distribution; certainly a means should at least be found for compensating for the mismatch.

6. Acknowledgements

This paper was conceived in 2018 when the first author was visiting professor at the *School of Criminal Investigation* at the *Southwest University of Political Science and Law* in Chongqing, China. The visit was funded from a grant under the *Chongqing Municipality Attracting Overseas Expertise Scheme* 巴渝海外引智计划, for which we would like to express our gratitude. The research was also supported by the *National Social Science Foundation of China Key Program* (Grant No. 16AYY015), *Southwest University of Political Science and Law Research Funding* (2015-XZRCXM003), and *Chongqing Social Enterprise and People's Livelihood Guarantee Scientific and Technological Innovation Special Research and Development Key Project* (cstc2017shms-zdyfX0060).

7. References

[1] Rose, P., "Likelihood ratio-based forensic voice comparison with higher level features: research and reality", in E. Lleida and L. J. Rodriguez-Fuentes [Eds], Recent Advances in Speaker and Language Recognition and Characterisation, Computer Speech and Language *Special Issue,* 476-502, 2017.

[2] Holdren, J.P., Lander, E.S. et.al. "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods," Science and technology advisory body to the President of the United States, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf, 2017.

[3] González-Rodriguez, J., Rose, P., Ramos, D., Torre, D. and Ortega-García, J., "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition", IEEE Trans. on Audio Speech and Language Proc., 15(7):2104-2115, 2007.

[4] Enzinger, E., Morrison, G.S. and Ochoa, F., "A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case", Science and Justice 56:42-57, 2016.

[5] Rose, P., "Where the science ends and the law begins – likelihood ratio-based forensic voice comparison in a $150 million telephone fraud", Int'l J. Speech Language and the Law 20(2):277-324, 2013.

[6] Lander, E.S., "Response to the ANZFSS council statement on the President's Council of Advisors on Science and Technology Report", Australian J. Forensic Sci., 2017.

[7] Gold, E. and French, J. P., "International practices in forensic speaker comparison", Int'l J. of Speech, Language and the Law 18(2):293-307, 2011.

[8] da Silva, R.R., da Costa, J.P.C.L., Miranda, R. K. and Del Galdo, G., "Applying base value of fundamental frequency via the multivariate Kernel-Density in Forensic Speaker Comparison", IEEE 10th Int'l conf. on Signal Processing and Communication Systems, 2016.

[9] Kinoshita, Y., Ishihara, S. and Rose, P**.,** **"**Exploring the Discriminatory Potential of F0 Distribution Parameters in Traditional Forensic Speaker Recognition", Intl. J. of Speech Language and the Law, 16(1): 91-111, 2009.

[10] Aitken, C.G.G. and Taroni, F., Statistics and the Evaluation of Evidence for Forensic Scientists, Wiley, 2004.

[11] Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J. and Niemi, T., Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition, Verlag für Polizeiwissenschaft, 2016.

[12] Ramos, D. and Gonzalez-Rodriguez, J., "Reliable Support: Measuring Calibration of Likelihood Ratios", Forensic Science International 230(1-3):156-169, 2013.

[13] Good, I.J., "Weight of Evidence and the Bayesian Likelihood Ratio", in C.G.G. Aitken and D.A. Stoney [Eds], The Use of Statistics in Forensic Science, 85-106, Ellis Horwood, 1991.

[14] Morrison, G. S., Rose, P. and Zhang, C. **"**Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice", Australian J. of Forensic Sci., 44(2):155-167, 2012.

[15] Morrison, G.S., Sound file cutter-upper matlab code, http://geoff-morrison.net/documents/Sound%20File%20Cutter%20Upper%20documentation.pdf

[16] Aitken, C.G.G. and D. Lucy, D., "Evaluation of trace evidence in the form of multivariate data", Applied Statistics 53(4):109-122, 2004.

[17] Morrison, G.S., 2012. "Tutorial on logistic regression calibration and fusion: converting a score to a likelihood ratio", Australian J. Forensic Sci., 1-25, 2012.

[18] Brümmer, N., Focal Toolkit http://www.dsp.sun.ac.za/nbrummer/focal

[19] Brümmer, N. and du Preez, J., "Application independent evaluation of speaker detection", Computer Speech and Language IEEE Odyssey 2004 Issue 20(2-3):230-275, 2006.

[20] Jaynes, E.T., Probability Theory - The Logic of Science, Cambridge University Press, 2003.