EVALUATING DISCRETE FORENSIC VOICE EVIDENCE: A PRELIMINARY INVESTIGATION BASED ON FILLED PAUSE OCCURRENCE

Michael Carne

Speech and Language Laboratory, The Australian National University
michael.carne@anu.edu.au

ABSTRACT

There are a variety of statistical methods for generating likelihood ratios (LR) from continuous acoustic data for the purposes of forensic voice comparison (FVC). Few methods exist for evaluating voice evidence where the data are discrete and those that do, remain largely untested. The pattern of filled pause (FP) occurrence in speech is a behavioural characteristic that is potentially individuating and is an example of discrete voice evidence. This paper evaluates the evidential value of FP occurrence (*um* /ɐːm/, *uh* /ɐː/) in Australian English (AE) using a Poisson-Gamma model for LR estimation. Results indicate that while speakers do exhibit idiosyncratic patterns of FP occurrence, the discriminatory potential of the feature set is limited. The paper suggests that including a larger set of speech disfluency-based features may yield improvement, as well as fusion with acoustic measurements. Limitations associated with the statistical model are also canvassed.

Keywords: forensic voice comparison, likelihood ratios, filled pauses, speech disfluency.

1. INTRODUCTION

In a likelihood ratio-based (LR) approach to evidence evaluation [1] [2], forensic material (e.g DNA, fingerprints, voice recordings) from a known source and unknown source are compared with the goal of determining how much more likely the two are from the same-origin (H_p) *vs.* different origins (H_d). Given evidence (E) in the form of forensic sample(s), H_p and H_d are evaluated as ratio of conditional probabilities: $p(E|H_p)/p(E|H_d)$) (i.e. the LR term in Bayes Theorem). In FVC, E is usually quantified by set of measurements from continuously-valued acoustic features. Typically these are either measurements based on phonetically informed features (e.g. vowel formants) or automatic acoustic features derived from signal processing techniques (e.g. MFCC). Acoustic features capture speaker-specific characteristics associated with anatomical differences between vocal tracts. A variety of methods for LR estimation exist for continuous acoustic variables, including MVKD [3], GMM-UBM, PLDA *i*-vector and *x*-vector LR estimation [4].

Speaker-specific characteristics are also reflected in longer term patterns of language use and are quantifiable discretely; either in terms of their presence or absence in speech, or by the frequency of their occurrence. These characteristics include things like a speakers' habitual lexical and syntactic choices, discourse patterns, conversational style, patterns of speech disfluency and so forth. Auditory phonetic features that specify pronunciation differences between speakers (e.g. /sɪti/ *vs.* /sɪɾi/ in AE), delineate regional accents and non-obligatory phonological processes are other examples. Little attention has been given to evaluation of discrete voice data within the LR framework. In [5] the evidential value of lexical choice in AE was evaluated. [6] evaluates the occurrence of clicks (paralinguistic/discourse level usage) in British English. In this paper a Poisson-Gamma LR model described in [6] is used to evaluate the evidential value of AE FP occurrence for the first time. The aim is to determine how well can we distinguish one speaker from another based on the occurrence of *ums* and *uhs* in speech applying an LR-based approach to evidential evaluation. In doing so, it builds previous work by evaluating a new feature type for FVC, as well as on [7] by demonstrating a feature-based procedure for LR estimation for discrete data.

1.1. Speech disfluency and forensics

Disfluency is a natural part of spontaneous speech. Speaker's hesitate, repeat words, or parts of words, and use various other strategies to delay or prolong speech. FPs (e.g *ums* and *uhs*) are a category of this phenomenon for which various explanations exist. There is evidence FPs serve pragmatic and discourse functions [8][9]. For instance, FPs are

often used to indicate when one is about to start or continue speaking. Cognitive studies suggest FPs are a strategy that allow time for a speaker to synchronize cognitive and speaking processes [9], as well as potentially play a role in listener's language processing [10]. The occurrence of FPs also appear to be linked to indexical factors such as age, sex and education [11]. Forensic phonetician's interest in FPs stems from evidence that individual speakers vary in the choice of filler words, patterns of occurrences [12][13] as well as acoustically [11]. While there are several studies evaluating the acoustic properties of FPs within the LR framework [11][14][15], there is only one evaluating FP occurrence [7]. A shortcoming of [7] though is that it employs a score-based LR estimation method, which only accounts for the similarity, not the typicality of forensic samples [16][17][18]. An alternative is a feature-based approach which estimates LRs directly from the feature values and incorporates typicality, such as the Poisson-Gamma model evaluated here.

2. METHODS

2.1. Data, pre-processing & feature extraction

The data consists of 208 speech transcriptions from 104 normally fluent AE male speakers (2 x non-contemporaneous recording sessions per speaker). Transcriptions were made from unscripted conversational speech obtained from [19]. The recordings varying between 3-5 minutes for each conversation-side. For these experiments transcriptions of the first 2 minutes are used. Counts of *um*, *uh* and *total FPs* (i.e. the total count of both) were vectorised separately for each speaker and each transcription. Summary statistics of tokens elicited from each session are given in Table 1.

recording	feature	$\bar{x}$	sd	range
1	uh	4.95	4.45	0 – 22
1	um	6.63	4.85	0 – 26
2	uh	4.56	4.19	0 – 20
2	um	6.82	5.53	0 – 24

Table 1: FP occurrence statistics (N speakers = 104)

2.2. Poisson-Gamma LR model

The following description is adapted from [6]. Let $x = (x_1, ..., x_{k_x})$ be the count of *ums* or *uhs* from the suspect (or known) recording and $y = (y_1, ..., y_{k_y})$ those from the offender (or unknown) recording. Counts of each are taken from consecutive time periods (in minutes) from the recordings, with k_x and k_y representing the number of periods sampled in each. To estimate the LR we are concerned with two quantities from the voice evidence t_x and t_y, which are the total number of occurrences from the recordings and given by $t_x = \sum_{i=1}^{k_x} x_i$ and $t_y = \sum_{i=1}^{k_y} y_i$. Assuming the counts in consecutive periods are independent and follow Poisson distribution (the implications of this assumption are discussed in Section 4), evidential value can be evaluated using the Poisson-Gamma LR model given in 1.

(1)
$$LR = \frac{\Gamma(\alpha + t_x + t_y)\Gamma(\alpha)}{\Gamma(\alpha + t_x)\Gamma(\alpha + t_y)} * \frac{(\beta + k_x)^{\alpha + t_x}(\beta + k_y)^{\alpha + t_y}}{\beta^{\alpha}(\beta + k_x + k_y)^{\alpha + t_x + t_y}}$$

where, Γ is a gamma distribution defined by α (shape parameter) and β (rate parameter). The Γ distribution assesses the typically of samples under comparison and is estimated using reference data from the relevant speaker population. In these experiments α and β were estimated via maximum likelihood estimation.

2.3. Testing and validation

The following features are evaluated: (1) counts of *um*; (2) counts of *uh* (3) total FP count (*um* + *uh*) and; (4) fused *um* and *uh*. The scores estimated from the Poisson-Gamma model were calibrated using a logistic-regression procedure [20]. In the case of (4) the procedure simultaneously calibrates and fuses LRs. The data were randomly split into test (20%), training (40%) and reference data (40%) and 5-fold cross-validation applied. The test data was used to simulate suspect and offender comparisons, the reference data to assess the typicality of the samples being compared and the training data to calculate weights for the calibration-fusion procedure. All features were tested using 60 and 120 seconds of netspeech.

2.4. Evaluation

The log-likelihood ratio cost function (C_{llr}) [20] and the Equal Error Rate (EER) are used to evaluate performance. C_{llr} is an information-theoretic measure of the validity (accuracy) of voice comparison systems. C_{llr} can be decomposed into two additional measures C_{llr}^{min} and C_{llr}^{cal} ($C_{llr} = C_{llr}^{min} + C_{llr}^{cal}$) which tell us the contribution of discrimination and calibration loss to the overall performance. The

C_{llr} value is a gradient metric which is small when LRs support the correct hypothesis (e.g. same-speaker comparisons are assigned an LR greater than 1) and increases as the magnitude of counter factual LRs do. A $C_{llr} = 0$ indicates perfect accuracy. Accuracy is worse as C_{llr} approaches and exceeds 1. The EER is a statistic commonly used in the evaluation biometric systems and is the point at which the false acceptance (or false positive) and false rejection (false negative) rates are equal; the lower the EER, the better the accuracy.

3. RESULTS

Table 2 shows the C_{llr} and $\%EER$ values for 60 and 120 seconds of net speech for each of the feature parametrisations described in Section 2.3. Results are mean values following a 5-fold cross-validation. The C_{llr} values are rather high, ranging between 0.94 and 0.98 bits. C_{llr} values are less than one though, indicating that some speaker-specific is present in FP occurrence albeit very little given values are close to 1. Commensurate with this result, the high $\%EERs$ reflect rather poor discrimination accuracy; the average $\%EER$ is 36.76%. The best performance is achieved via fusing FP features (*um* and *uh*) (C_{llr} = 0.94). The same result is obtained for both 60 sec. and 120 sec. of net speech. However, a slightly lower $\%EER$ is achieved (33.98% *vs.* 34.16%) and discrimination performance is marginally better with 120 sec. ($C_{llr}^{min} = 0.8$ *vs.* 0.83).

Table 2: Performance metrics for FP features for 60 and 120 seconds of net speech.

	feature	C_{llr}	C_{llr}^{min}	C_{llr}^{cal}	$\%EER$
60*sec.*	fused	0.94	0.83	0.11	34.16
	uh	0.96	0.85	0.11	38.79
	um	0.98	0.89	0.09	38.93
	total	0.96	0.85	0.10	36.13
120*sec.*	fused	0.94	0.80	0.14	33.98
	uh	0.95	0.84	0.11	35.81
	um	0.98	0.87	0.12	37.49
	total	0.98	0.84	0.15	38.86

The amount of net speech used does not result in meaningful performance gains. The additional minute of speech data only yields improvement in C_{llr} for *uh* and it is very slight - only 0.01 bits (C_{llr} = 0.96 *vs.* 0.95). For *um* C_{llr} is the same (0.98) and for *total FPs* it is worse (0.96 *vs.* 0.98). We do though see an improvement in discrimination performance in the 120 sec. condition, where C_{llr}^{min} values overall are lower. However, this improvement comes at the cost of poorer calibration which is evident from the higher C_{llr}^{cal} values obtained. Overall, it appears the strength of evidence obtained from FP features is rather weak. This can be appreciated in more detail via the Tippett plot in figure 1 which shows the cumulative proportion of same- and different speaker trials for 120 secs. of speech data for the best performing settings (i.e. fused, 120 secs.).

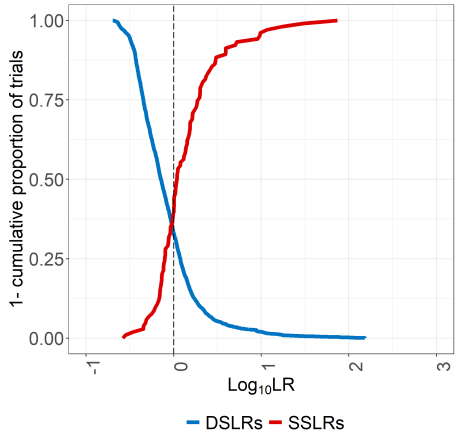

Figure 1: Tippett plot showing $Log_{10}LRs$ from SS (red line) and DS (blue line) comparison trials based on fused FP features (*uh* + *um*) from 120 seconds of net speech.

In figure 1 blue lines converging from the left are LRs from different-speaker trials (DSLRS) and red lines converging from the right are same-speaker trials (SSLRs). The *y*-axis shows the cumulative proportion of trials plotted as a function of $Log_{10}LR$ on the *x*-axis. The minimum $Log_{10}LR$ achieved for DS comparisons is -0.69 and overall a large proportion of $Log_{10}LR$ from the DS comparisons are close to the decision threshold (i.e. 0). This indicates most comparisons yield very weak support for the DS hypothesis. The magnitude of $Log_{10}LRs$ supporting the incorrect hypothesis (i.e. $Log_{10}DSLRs$ greater than zero) are also relatively large (the maximum $Log_{10}DSLR = 2.18$). The situation is slightly less severe for the SS comparisons in that the magnitude of counter factual SSLRs is less (minimum SSLR = -0.58). However, the strength of evidence for correctly assessed SS pairs is weak. The maximum $Log_{10}SSLR$ achieved is 1.87). Figure 2 shows Tippett plots from the voice comparisons for individual FP features. Figure 2A and figure 2B (*um*) are components of the fused system shown in figure 1 (*uh*). The results for *um* (figure 2A.) show a similar configuration to what we have just seen for the fused result. That is, $Log_{10}DSLRs$ yield relatively weak strength of

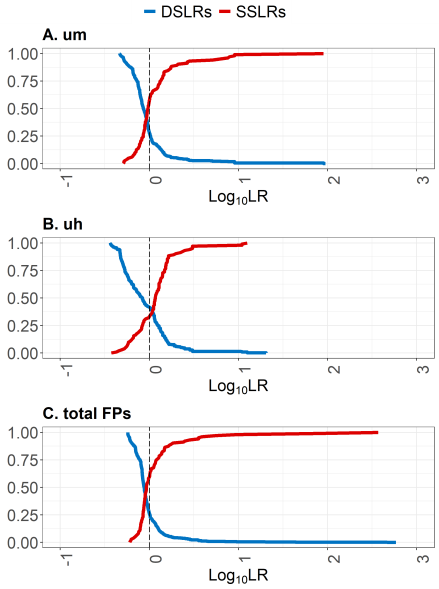

Figure 2: Tippett plots showing $Log_{10}LRs$ from SS (red line) and DS (blue line) comparison trials based for individual FP features (*uh* , *um* and *total FPs*) from 120 seconds of net speech, *y*-axis = 1 - cumulative proportion of trials.

evidence with a large proportion of trials supporting the contrary-to-fact hypothesis. The minimum LR is slightly weaker ($Log_{10}DSLRs$ = -0.33) than in the fused system, but this is to be expected. SS comparisons fair better and produce almost identical results to SS comparisons in the fused system in terms of LR magnitude. For *uh* (figure 2B), LRs are less biased towards the SS comparisons evident from the more symmetric appearance of the Tippett lines for SS and DSLRs. Here the magnitude of LRs for SS comparisons is slightly weaker than in figure 2A, but this is balanced by a reduction in the magnitude of counter-factual DSLRS. The magnitude of counter-factual DSLRs for *total FPs* (figure 2C) by contrast is substantial relative to the tippetts in 2A and B.

4. DISCUSSION & CONCLUSIONS

The results indicate AE FP occurrence is of relatively limited forensic use given the poor strength of evidence that is obtainable. This is reflected in high C_{llr} values (close to 1) and $\%EER$ observed. That said, these experiments have only considered a narrow set of speech disfluencies. Discriminant analysis in [13] demonstrated that speaker classification rates are improved substantially when several types of disfluencies are combined (e.g. word repetitions, silent and FPs, prolongations etc.). Similarly, in a previous LR-based study of FP occurrence in Japanese [7] substantial improvements are seen in C_{llr} values as the number of FPs used in LR estimation is increased. In fact, evidence of this can been seen in the results presented in this paper where logistic regression fusion of *um* and *uh* is shown to yield better performance. Also worth exploring is fusion of disfluency-based features with acoustic and temporal information. It was shown in [11] that formant and nasal duration information from *um* yields low C_{llr} and $\%EER$ values. It has also been shown that F_0 associated with FPs [12] varies between speakers. Incorporating this information into the voice comparison would also likely be beneficial. While acoustic features require continuous methods for LR estimation, LRs evaluating different kinds of evidence (from different models) can be combined to give a measure of the overall strength of evidence; either directly through multiplication of LRs or, where evidence is correlated, via a fusion procedure such as the logistic-regression procedure used here. In this way, FP occurrence and other disfluency features may prove useful in providing complementary information to acoustic-based systems; particularly in instances where data is limited.

Another set of limitations relate to the statistical model used to generate LRs. First, the Poisson-Gamma model assumes FPs occurrence in consecutive periods are independent. In reality, FP occurrence is very likely dependant on factors such as discourse structure. For example, where FP are used to signal turn-taking, the occurrence of FPs will be dependant the discourse structure between interlocutors - this is likely to vary minute-to-minute. Second, Poisson-Gamma is a univariate model and therefore LRs were calculated separately for *um* and *uh*. This ignores the potential for interactions between individual speakers' choice of FP. For example, it may be that speakers exhibit a preference for *uh* over *um* and do so in different contexts. A multivariate discrete LR model, such as a multinomial LR, may be more appropriate to better capture these kinds of interactions.

This paper has evaluated the evidential value of FP occurrence in AE using a discrete model for LR estimation. The feature set was shown to be a limited forensic value, but may complement acoustic-based systems where data is limited.

5. ACKNOWLEDGMENTS

The author's research is supported by an Australian Government Research Training Scholarship and

ANU Supplementary Scholarship. Dr. Shunichi Ishihara provided valuable comments on a draft version of this paper. I would also like to thank the reviewers for their constructive comments.

6. REFERENCES

[1] I. W. Evett, "Towards a uniform framework for reporting opinions in forensic science casework," *Science and Justice*, vol. 38, no. 3, pp. 198–202, 1998.

[2] I. W. Evett, J. A. Lambert, and J. S. Buckleton, "A bayesian approach to interpreting footwear marks in forensic casework," *Science and Justice*, vol. 38, no. 4, pp. 241–247, 1998.

[3] C. Aitken and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," *Journal of the Royal Statistical Society: Series C (Applied Statistics))*, vol. 53, pp. 109–122, 2004.

[4] G. Morrison, E. Enzinger, D. Ramos, J. GonzÃ¡lez-RodrÃguez, and A. Lozano-DÃez, *Statistical models in forensic voice comparison*. Boca Raton, FL: CRC, 2020, book section 21.

[5] M. Carne, Y. Kinoshita, and S. Ishihara, "High level feature fusion in forensic voice comparison," in *Proc. Interspeech 2022*, 2022, pp. 5293–5297.

[6] C. Aitken and E. Gold, "Evidence evaluation for discrete data," *Forensic Science International*, vol. 230, no. 1-3, pp. 147–155, 2013.

[7] S. Ishihara and Y. Kinoshita, "Filler words as a speaker classification feature," in *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*, M. Tabain, J. Fletcher, D. Grayden, J. Hajek, and A. Butcher, Eds., 2010, pp. 34–37.

[8] M. Swerts, "Filled pauses as markers of discourse structure," *Journal of Pragmatics*, vol. 30, no. 4, pp. 485–496, 1998.

[9] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–111, 2002.

[10] M. Corley, L. J. MacGregor, and D. I. Donaldson, "It's the way that you, er, say it: Hesitations in speech affect language comprehension," *Cognition*, vol. 105, no. 3, pp. 658–668, 2007.

[11] V. Hughes, S. Wood, and P. Foulkes, "Strength of forensic voice comparison evidence from the acoustics of filled pauses," *International Journal of Speech Language and the Law*, vol. 23, no. 1, pp. 99–132, 2016.

[12] A. Braun and A. Rosin, "On the speaker-specificity of hesitation markers," in *The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow., 2015, pp. 731–736.

[13] K. McDougall and M. Duckworth, "Profiling fluency: An analysis of individual variation in disfluencies in adult males," *Speech Communication*, vol. 95, pp. 16–27, 2017.

[14] X. Wang, V. Hughes, and P. Foulkes, "The effect of speaker sampling in likelihood ratio based forensic voice comparison," *International Journal of Speech, Language and the Law*, vol. 26, no. 1, pp. 97–120, 2019.

[15] B. X. Wang, V. Hughes, and P. Foulkes, "The effect of sampling variability on systems and individual speakers in likelihood ratio-based forensic voice comparison," *Speech Communication*, vol. 138, pp. 38–49, 2022.

[16] G. S. Morrison and E. Enzinger, "Score based procedures for the calculation of forensic likelihood ratios - scores should take account of both similarity and typicality," *Sci Justice*, vol. 58, no. 1, pp. 47–58, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29332694

[17] A. Bolck, H. Ni, and M. Lopatka, "Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic mdma comparison," *Law, Probability and Risk*, vol. 14, no. 3, pp. 243–266, 2015.

[18] S. Ishihara and M. Carne, "Likelihood ratio estimation for authorship text evidence: An empirical comparison of score- and feature-based methods," *Forensic Science International*, vol. 334, pp. 111–268, 2022.

[19] G. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B. Folkes, S. De Souza, N. Cummins, D. Chow, and A. Szczekulska, "Forensic database of voice recordings of 500+ australian english speakers (auseng 500+)," 2021. [Online]. Available: Available:http://databases.forensic-voice-comparison.net

[20] N. Brummer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2, pp. 230–275, 2006.