The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems

Ewald Enzinger, Geoffrey Stewart Morrison

Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, Australia

e.enzinger@student.unsw.edu.au, geoff-morrison@forensic-voice-comparison.net

Abstract

In this paper we report on a study which demonstrates the importance of using non-contemporaneous test data in evaluating the validity and reliability in forensic-voice-comparison systems. We test four different systems: one MFCC GMM–UBM, one vowel formant-trajectory based, one nasal spectra based, and the fusion of the three systems. Each system is tested on the same set of test recordings, including same-speaker and different-speaker pairs. In one condition, the same-speaker pairs are from contemporaneous (within-session) recordings and in the other they are from non-contemporaneous (between-session) recordings. Within-session testing always overestimated the performance of the systems compared to between-session testing.

Index Terms: session variability, likelihood ratio, validity, reliability, forensic voice comparison

1. Introduction

In forensic casework there is always a difference in time between when the recording of the offender is made and when a recording of a suspect is made. Thus in cases where the suspect and offender recordings are produced by the same speaker, they are produced by the same speaker speaking on different occasions. The characteristics of a person's voice are expected to vary more from occasion to occasion than on a particular occasion (see [1], p. 235; [2], p. 12). Factors increasing variability include, amongst others, differences due to relaxation versus stress on vocal folds (e.g. when speaking the first time in the morning versus speaking for a long time), state of health (e.g. nasal congestion, 'cold speech' [3], laryngitis), speaking style, emotional state, as well as other random or pseudorandom variation from occasion to occasion. Given these obvious influences, an appropriate default assumption would be that between-session variability does matter for forensic voice comparison. If one were to test the performance of a forensic voice comparison system using same-session data, one would be assuming that between-session variability does not matter and have to be able to present evidence to justify this assumption.

The existence of between-session variability affecting distributions of features extracted from speech samples of a speaker has been acknowledged in both forensic-phonetic (e.g. [2], p. 106; [4]) and automatic-speaker-recognition communities (e.g. [5, 6]). However, empirical studies on forensic voice comparison have often not accounted for this, and have tested using data obtained from a single recording session (for example, formants [7], formant trajectories [8, 9, 10, 11], voice source [12], and automatic systems [13]).

Authors often acknowledge the need for between-session testing, but perform within-session testing because they are using convenient databases which do not include multiple non-contemporaneous recordings of each speaker; however, [11] (p. 33) claims that "the importance of non-contemporaneity and the issue of whether it furnishes greater within-speaker variation than in a single natural recording remains an empirical question."

In this paper we investigate the validity and reliability (accuracy and precision) of four different systems: one mel-frequency-cepstral-coefficient Gaussian-mixture model (MFCC GMM–UBM) based, one vowel formant-trajectory based, one nasal-spectra based, and the fusion of the three aforementioned systems. Each system was trained and tested using the same database of voice recordings. In this investigation we are not concerned with between-session effects due to recording- and transmission-channel differences or speaking-style differences and do not aim to systematically investigate differences due to state of health or fatigue. Rather, we focus on naturally occurring occasion-to-occasion variability.

Data obtained from each of two recording sessions was divided into two non-overlapping parts to allow for within- and between-session same-speaker comparisons. Speaking-style as well as the amount of data used for offender and suspect samples were controlled. In the analysis and presentation of results we focus on same-speaker comparisons, since different-speaker comparisons are by definition between-session. Since the amount of data employed is limited, the absolute performance of the system is of less interest than the relative differences in performance between the two conditions.

To emulate a set of conditions which may be representative of forensic casework, all comparisons involve a mismatch in transmission channel, a common feature of forensic speech material. Data used as nominal offender samples are taken from a recording of a mobile-to-landline transmission and data used as nominal suspect samples are taken from a high-quality recording. This mismatch is accounted for in the modeling of the background data, data used for calibration training, as well as data used for testing.

2. Methodology

2.1. Data

The data were extracted from a database of two non-contemporaneous voice recordings of each of 60 female speakers of Standard Chinese (Mandarin, Putonghua). See [14] for details of the data collection protocol. The first and second recording sessions were separated by 2-3 weeks. High-quality recordings were made at 44.1 kHz 16 bit using flat-frequency response lapel microphones (Sennheiser MKE 2 P-C) and an

external soundcard (Roland®UA-25 EX), with one speaker on each of the two recording channels. In addition to the original high-quality recordings, degraded sets of recordings were created by passing the high-quality set of recordings through transmission channels. In this study we use a mobile-to-landline condition to investigate the effects under a mismatch condition. The mobile telephone (Nokia 2730 classic) used to transmit the signal was placed in a sound booth (IAC 250 Series Mini Sound Shelter) in the vicinity of a loudspeaker (Roland MA-7A) connected to a computer via an external sound card (Roland®UA-25 EX). The high-quality recordings were played though the loudspeaker and the acoustic signal picked up by the in-built microphone of the transmitting telephone through which a call was established to the receiving landline telephone (Polaris NRX EVO 450), which was connected to the external sound card via a Trillium Telephone Recording Adapter Studio Interface (REC-ADPT-SI). (No attempt was made to simulate potential variability due to changes in the position of the telephone relative to the speaker's mouth).

In the tests of forensic-voice-comparison systems below, recordings from the first 20 speakers (identification numbers: 01–04, 09–20, 22, 25, 26, 28) were used as background data, data from the next 20 speakers (29–48) were used as development data, and data from the last 20 speakers (49–68) were used as test data.

2.2. Forensic-voice-comparison systems

2.2.1. Automatic MFCC GMM–UBM system

The automatic MFCC GMM–UBM system was of generic design. 16 Mel-frequency cepstral-coefficient (MFCC) values were extracted every 10 ms over the entire speech-active portion of each recording using a 20 ms wide hamming window. Delta coefficient values were also calculated [15]. Feature warping [16] was applied to the MFCCs and deltas before subsequent modelling. A Gaussian mixture model - universal background model (GMM–UBM, [17]) was built using the background data to train the universal background model. After tests on the development set in order to optimize the number of Gaussians in the model, the number of Gaussians used for testing was 1024.

2.2.2. Formant-trajectory system

Manual formant measurements were made of the formant trajectories of stressed /iau/ triphthong tokens on tone 1 using the FORMANTMEASURER software [18]. Discrete cosine transforms (DCTs) were fitted to time-normalized trajectories of the second and third formant. The 0th through 3rd coefficient values used as input to Aitken & Lucy's multivariate kernel density formula (MVKD) [19]. See [20] for details on the procedure.

2.2.3. Nasal system

Spectral characteristics of syllable-initial bilabial nasal stop (/m/) tokens were modelled by pole-zero model estimates obtained from the middle 70% of the segments [21]. The order of the denominator and numerator polynomials was set to 13 and 7, respectively. Cepstral coefficients were computed from the pole-zero model envelopes and used as input to the MVKD formula. After tests on the development set in order to optimize the number of cepstral coefficients, the number used for testing was 12.

2.2.4. Calibration and fusion

Individual systems were calibrated and fused using logistic-regression calibration and fusion [22, 23, 24, 25, 26, 27]. The weights for the linear transform were obtained using the pooled procedure [28] from scores calculated from the development set and then applied to the scores from the test set.

2.3. Evaluation procedure

In order to control for the amount of information used for within-session test comparisons and between-session test comparisons, each condition used the same amount of information. The amount of information was defined according to the number of tokens of a particular phoneme or the number of frames, depending on the procedure employed for extracting information from the acoustic signal. Each recording was split into two parts (part 1 and part 2), each containing the same amount of information (5 tokens of /iau/ for the formant-trajectory system, 10 tokens of /m/ for the nasal system, and 6016 frames for the MFCC GMM-UBM system).

Within-session same-speaker comparisons were made by comparing session 1 part 1 with session 1 part 2, and by comparing session 2 part 1 with session 2 part 2. Between-session same-speaker comparisons were made by comparing session 1 part 1 with session 2 part 1, and by comparing session 1 part 2 with session 2 part 2. All different-speaker comparisons were by-definition between-session and made using the same scheme as for same-speaker between-session comparisons. Background models were trained using between-session full-length recordings.

Weights for calibration and fusion were trained using scores obtained from the development set. Data used as nominal offender and suspect samples was constrained to the same amount of information as was used for testing, as outlined above (Pilot studies indicated that weights calculated from scores from comparisons using all of the data available in the original recordings resulted in poor calibration). In both same- and different-speaker comparisons, session 1 part 1 was compared with session 2 part 2. Scores obtained from both within-session and between-session condition tests were calibrated using the same set of weights.

Validity and reliability (accuracy and precision) were assessed on the results from the test set using procedures from [29, 28]. Validity was assessed using the log-likelihood ratio cost (C_{llr}, [22]). Reliability was assessed by calculating the 95% credible interval (95% CI) using the parametric method [29, 28].

Since the different-speaker test comparisons are identical in both conditions, we also present measures estimated using only same-speaker comparisons pairs, so as to more clearly illustrate the difference between the two conditions. Validity on same-speaker comparisons was assessed by the first term of C_{llr} associated with same-speaker comparisons,

$$C_{llr}^{ss} = \frac{1}{N_{ss}} \sum_{i=1}^{N_{ss}} \log_2 \left(1 + \frac{1}{LR_{ss_i}}\right), \qquad (1)$$

which was calculated from same-speaker comparison pairs, e.g. session 1 part 1 versus session 2 part 1, and session 1 part 2 versus session 2 part 2. Reliability was assessed by calculating the 95% credible interval (95% CI) on same-speaker pairs rather than all comparison pairs, since the difference in within-session and between-session same-speaker comparisons

is otherwise obscured by the much higher number of different-speaker comparisons.

3. Results and discussion

Figure 1a shows the performance in terms of C_{llr} and the 95% credible interval estimated from both same- and different-speaker comparisons, and Figure 1b shows the C_{llr}^{ss} and the 95% credible interval estimated only from same-speaker comparisons, respectively.

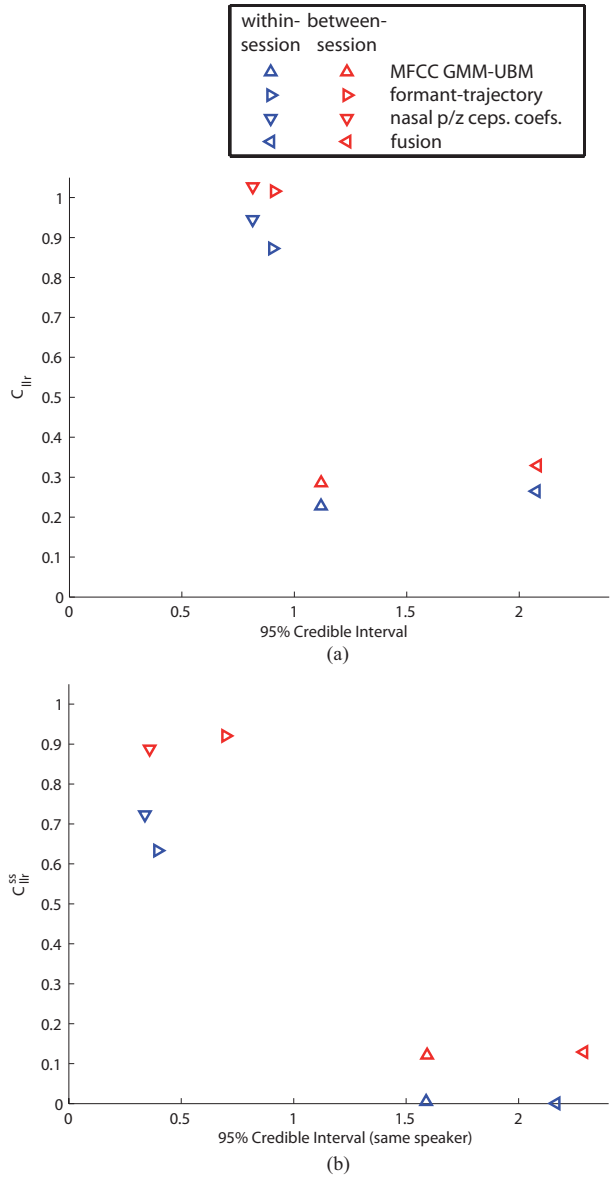

Figure 1: (a) Measures for validity (C_{llr}) and reliability ($log10$ 95% credible interval) for the different systems estimated from both same- and different-speaker comparisons from within-session (*blue*) and from between-session (*red*) data. (b) Measures for validity (C_{llr}^{ss}) and reliability ($log10$ 95% credible interval) for the different systems estimated only from same-speaker comparisons from within-session (*blue*) and from between-session (*red*) data (mobile-to-landline v high-quality recordings).

Within-session testing always overestimated the performance of the systems compared to between-session testing. In all instances it clearly overestimated the degree of validity. In some cases the reliability of the systems is also estimated to be higher, particularly for the formant-trajectory system (Figure 1b).

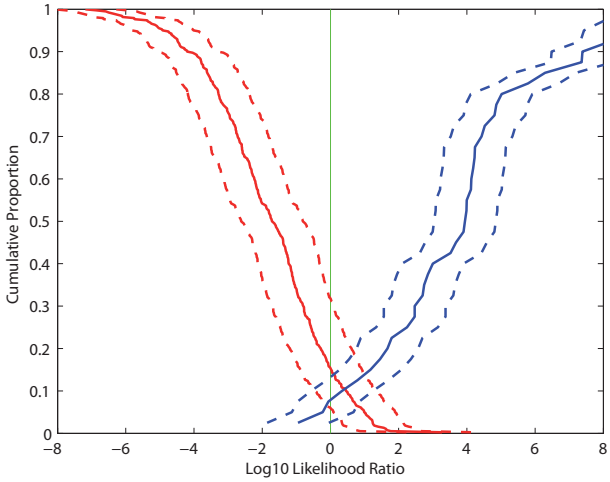

Figure 2: Tippett plot showing the performance of the fused system including the 95% credible interval estimated from both different-speaker and between-session same-speaker comparisons (mobile-to-landline v high-quality recordings).

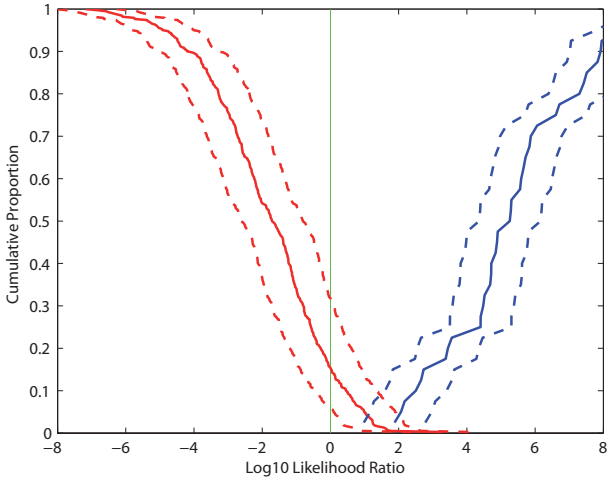

Figure 3: Tippett plot showing the performance of the fused system including the 95% credible interval estimated from both different-speaker and within-session same-speaker comparisons (mobile-to-landline v high-quality recordings).

Figures 2 and 3 show Tippett plots of the fused system on tests of between-session same-speaker comparisons and within-session same-speaker comparisons, respectively. The red curve showing likelihood ratios obtained from different-speaker comparisons is the same in both conditions. The blue curve representing the likelihood ratios obtained from same-speaker comparisons is further to the right for the within-session condition compared to the between-session condition, indicating generally higher likelihood ratios for same-speaker comparisons in

this condition.

4. Conclusion

The results presented in this study demonstrate a clear overestimation of validity and reliability when testing on within-session data rather than between-session data. The differences in performance presented here are due to factors other than mismatch in channel conditions and the amount of data available, as these have been controlled and accounted for in data used by the systems. Since forensic samples are always non-contemporaneous, between-session variability should be accounted for when testing the validity and reliability of forensic-voice-comparison systems.

5. References

[1] P. Rose, *Forensic Speaker Identification*. London: Taylor & Francis, 2002.

[2] F. Nolan, *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press, 1983.

[3] R. G. Tull, "Acoustic analysis of cold-speech: Implications for speaker recognition technology and the common cold," Ph.D. dissertation, Northwestern University, 1999.

[4] P. Rose, "Long- and short-term within-speaker differences in the formants of australian *hello*," *Journal of the International Phonetics Association*, vol. 29, no. 1, pp. 1–31, 1999.

[5] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "A comparison of session variability compensation techniques for svm-based speaker recognition," in *Proc. Interspeech*, Antwerp, Belgium, August 2007, pp. 790–793.

[6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in gmm-based speaker verification," *IEEE Trans. Audio, Speech Lang. Proc.*, vol. 15, no. 4, pp. 1448–1460, 2007.

[7] T. Becker, M. Jessen, and C. Grigoras, "Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models," in *Proc. Interspeech 2008 incorporating SST'08*. ISCA, September 2008, pp. 1505–1508.

[8] E. Enzinger, "Characterizing formant tracks in viennese diphthongs for forensic speaker comparison," in *Proceedings of the AES 39th International Conference – Audio Forensics*, 2010, pp. 47–52.

[9] P. Rose, "The intrinsic forensic discriminatory power of diphthongs," in *Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology*, 2006, pp. 64–69.

[10] K. McDougall and F. Nolan, "Discrimination of speakers using the formant dynamics of /uː/ in British English," in *Proceedings of the 16th International Congress of Phonetic Sciences*, J. Trouvain and W. Barry, Eds. Saarbrücken: ICPhS, August 2007, pp. 1825–1828.

[11] V. Hughes, "The effect of variability on the outcome of numerical likelihood ratios for forensic voice comparison," Master's thesis, University of York, 2012.

[12] P. Gómez-Vilda, A. Álvarez, L. Mazaira-Fernándeza, R. Fernández-Baillo, V. Nieto, R. Martínez, C. Muñoz, and V. Rodellar, "A hybrid parameterization technique for speaker identification," in *Proceedings of the 16th European Signal Processing Conference (EUSIPCO)*, Lausanne, Switzerland, 2008.

[13] H. Künzel, "Automatic speaker recognition of identical twins," *International Journal of Speech, Language, and the Law*, vol. 17, no. 2, pp. 251–277, 2010.

[14] G. S. Morrison, P. Rose, and C. Zhang, "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice," *Australian Journal of Forensic Sciences*, 2012.

[15] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech and Sig. Proc.*, vol. 34, pp. 52–59, 1986.

[16] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proceedings of the Odyssey Speaker Recognition Workshop*. International Speech Communication Association, 2001.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[18] G. S. Morrison and T. M. Nearey. (2011) FormantMeasurer: Software for efficient human-supervised measurement of formant trajectories. [Online]. Available: http://www.geoff-morrison.net/

[19] C. G. G. Aitken and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," *Applied Statistics*, vol. 53, no. 1, pp. 109–122, 2004.

[20] C. Zhang, G. S. Morrison, and T. Thiruvaran, "Forensic voice comparison using Chinese /iau/," in *Proceedings of the17th International Congress of Phonetic Sciences*, Hong Kong, China, 2011, pp. 2280–2283.

[21] D. Marelli and P. Balazs, "On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis," *IEEE Trans. Audio, Speech Lang. Proc.*, vol. 18, no. 2, pp. 237–248, Feb. 2010.

[22] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, pp. 230–275, 2006.

[23] D. A. van Leeuwen and N. Brümmer, *Speaker Classification I. Fundamentals, Features, and Methods*, ser. Lecture Notes in Artificial Intelligence (LNAI). Springer, 2007, vol. 4343, ch. An Introduction to Application-Independent Evaluation of Speaker Recognition Systems, pp. 330–353.

[24] N. Brümmer. (2005) Tools for fusion and calibration of automatic speaker detection systems. [Online]. Available: http://niko.brummer.googlepages.com/focal

[25] G. S. Morrison. (2009) Robust version of train_llr_fusion.m from Niko Brümmer's FoCaL Toolbox (release 2009-07-02). [Online]. Available: http://geoff-morrison.net/

[26] S. Pigeon, P. Druyts, and P. Verlinde, "Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions," *Digital Signal Process.*, vol. 10, pp. 237–248, 2000.

[27] N. Brümmer, L. Burget, J. Černocký, O. Glembek, F. Grézl, M. Karafiát, D. A. van Leeuwen, P. Matějka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Trans. Audio, Speech Lang. Proc.*, vol. 15, no. 7, pp. 2072–2084, 2007.

[28] G. Morrison, T. Thiruvaran, and J. Epps, "An issue in the calculation of logistic-regression calibration and fusion weights for forensic voice comparison," in *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*, Melbourne, 2010, pp. 74–77.

[29] G. S. Morrison, "Measuring the validity and reliability of forensic likelihood-ratio systems," *Science & Justice*, vol. 51, pp. 91–98, 2011.