A CAUTIONARY TALE FOR PHONETIC ANALYSIS: THE VARIABILITY OF SPEECH BETWEEN AND WITHIN RECORDING SESSIONS

Sula Ross, Kate Earnshaw, Erica Gold

University of Huddersfield
[s.m.ross, k.earnshaw, e.gold @ hud.ac.uk]

ABSTRACT

This paper investigates within and between session variability using a subset of 60 British English male speakers from the WYRED project. Three separate speaking tasks were compared using extracted i-vector PLDA scores within iVOCALISE. Different speaker pairs from contemporaneous (within-session) recordings and non-contemporaneous (between-session) recordings were tested. A within-session, between-task comparison was also performed in order to consider variation in speech style in addition to non-contemporaneity. EER and C_{llr} values indicate that non-contemporaneity is not the only factor which needs to be taken into account when conducting phonetic analysis or evaluating speaker comparison systems, as speech style also seems to play an important role. Further analysis supports the requirement for (forensic/socio-) phoneticians to sample data from the entirety of a recording, especially if the nature of the speech elicitation may change during the task, as the degree of variability is dependent on which portion of the sound file is sampled.

Keywords: Non-contemporaneous, forensic speaker comparison, phonetic analysis, i-vector, variation

1. INTRODUCTION

It is well known that voices are highly plastic, and variability in a person's speech can be caused by a number of factors (e.g. interlocutor, style, topic, health, time of day, etc.). For this reason, the best way to capture variability in speech is often to collect data from multiple sessions via tasks that elicit a range of speech styles. Despite this, phonetic, sociolinguistic, and forensic speech science research often involves analysis using data from only a single speaking task. Furthermore, the data used for analysis is sometimes only extracted from a sub-section of a recording (e.g. the beginning or middle). Consequently, the extent to which speakers vary within tasks, within recording sessions, and across recording sessions is not fully understood. This motivates the need for empirical testing of the levels of intra-speaker variability that exist within and between tasks, made over different recording sessions.

A recent study compared a range of data collection methodologies with sociolinguistic variation in mind, and reflected on the usefulness of analysing large volumes of data containing stylistic variability when using controlled and replicable laboratory recordings [3]. If laboratory recordings included data from the same speaker recorded over separate sessions, stylistic variability between and within sessions could be examined.

The importance of using between-session recordings has been highlighted as a factor that can hinder speaker recognition ability [15] and therefore provides a significant concern to forensic speech scientists, due to the inherent variability present [8, 13, 14]. Automatic speaker recognition (ASR) systems are being used more and more frequently in casework around the world [9]. For this reason, it is crucial that the performance of these systems are tested under experimental conditions, using forensically relevant data. A variety of both acoustic-phonetic and ASR systems were found to overestimate the validity and reliability of results when only within-session data was considered [8] and results degraded in ASR systems when using between-session recordings [17]. The ASR studies involved the use of Gaussian Mixture Model - Universal Background Model (GMM-UBM) approach and Mel Frequency Cepstral Coefficients (MFCCs). One recent study focussing on the effect of speaking style on between-session recordings, compared MFCCs using state-of-the-art i-vector probabilistic linear discriminant analysis (PLDA) [12]. The study modified recordings to make them more forensically realistic. Results indicated that mismatched data had a negligible effect on system performance. However, this contrasts with the findings of another recent study [18] that tested how well LTF0 performs, using a likelihood ratio framework, under both matched and mismatched conditions in terms of speech style. They found that the strength of forensic speaker recognition evidence was weaker under mismatched conditions than the matched conditions.

The focus of this paper is to review the relative speaker discriminatory performance of contemporaneous (within-session) and non-contemporaneous (between-session) comparisons using data from the West Yorkshire Regional English

Database (WYRED) [10]. Using iVOCALISE [1], levels of intra- and inter-speaker variation are extracted in order to quantify the strength of the evidence with respect to the competing same-speaker and different-speaker hypotheses. When levels of intra-speaker variation are low, and inter-speaker variability is high, the system is expected to perform well. When the system performs perfectly, all same speaker and different speaker pairs are correctly identified. By using contemporaneous and non-contemporaneous data, it is possible to evaluate to what degree non-contemporaneity causes intra-speaker variation to increase.

The inclusion of data from the WYRED project allows for the analysis of a large volume of studio quality recordings, from a carefully stratified population, over three distinct spontaneous speaking tasks. This enables three within-task comparisons, two within-session between-task comparisons (recorded on the same day), and four between-session comparisons recorded at least six days apart. This multi-file analysis allows for exploration of variability in stylistic differences that may be present within the separate laboratory recorded tasks.

2. METHODOLOGY

2.1. Data

WYRED is the largest forensically-relevant database of British English speech. In total, 180 participants were recorded undertaking four style-controlled tasks. The current study includes a subset of 60 speakers equally divided across three boroughs within West Yorkshire (Northern England): Bradford, Kirklees, and Wakefield. All speakers are native British monolingual males, aged 18-30, who grew up and went to school in West Yorkshire.

2.2. Recording sessions

Recordings were carried out over two separate sessions. Participants recorded the first two tasks on their initial visit, and recorded the final two tasks in their second visit. Session 1 and 2 were recorded a minimum of six days apart for all participants, but due to limitations in recruitment and participant availability some participants attended their second session up to 104 days later. The average length between Session 1 and Session 2 was 17.3 days. The tasks within Session 1 were adaptations of the first two tasks used in the DyViS database [16]. The first task in Session 2 was a paired conversation using adapted topic prompt cards [19] and the second was an experimental task where the participant left an answerphone message. Further details on the tasks are provided in [10].

2.3. Recording set-up

The database was recorded in a purpose-built sound booth. High quality recordings were made using a Sennheiser HSP 4 omnidirectional headband microphone and recorded onto a Marantz PMD661 MKII Handheld Solid State Recorder in PCM WAV format (44.1kHz, 16 bit). Only studio quality recordings were used in this investigation.

It is important to note that similar studies examining the effect of speaker style on non-contemporaneous recordings have attempted to replicate a realistic forensic case comparison by simulating extrinsic conditions that may be found [12]. However, the current study seeks to examine results from a controlled baseline to review non-contemporaneity and mismatched speech styles using high quality studio data.

2.4. Preparation of files

Prior to analysis, the original sound files for all 60 speakers were manually edited to remove any interlocutor speech and background noises (e.g. coughs, sneezes, fidgeting noises) within Praat [4]. This reduced the length of the files that were used for this investigation. Each task was subsequently divided into two halves. The two halves were necessary to carry out within-task comparisons, as well as increasing the overall number of comparisons within and between session recordings. We split task files in half to capture the level of variation present within each recording.

A decision was made to exclude the WYRED Task 4 recordings in this study as the resulting files were considered to have an insufficient amount of net speech (min length: 31s, max length: 93s, average length: 63s). As speaker recognition performance can be highly influenced by file length [11], the durational difference between Task 4 and the other three tasks was considered too great. Table 1 provides a summary of the minimum, maximum and average file length of recordings used in this study, after being edited and divided in half.

Table 1: Amount of net speech per task

Task	Min (sec)	Max (sec)	Avg (sec)
Session 1:			
1. Mock Police Interview	137	668	301
2. Accomplice Call	277	499	373
Session 2:			
3. Paired Conversation	62	470	235

2.5. Forensic speaker comparison system

Forensic speaker comparisons were performed using iVOCALISE [1]. Using the classifier framework of i-vector – PLDA [6, 7], MFCCs were obtained for all speaker files. Default settings were used: 13 MFCCs extracted, Delta features selected, 24 Filter banks, Channel Normalisation: Mean Subtraction, 1024 Gaussians, and 10 Train Cycles.

Nine multi-file comparisons were performed using pre-trained models within iVOCALISE as a reference sample and results were calibrated using a reference normalisation subset of 60 different (non-test), studio quality WYRED speakers. This created matrices of i-vector PLDA scores which were then cross-validated using Bio-Metrics [2], in order to calculate Equal Error Rate (EER) and Log-Likelihood Ratio Cost (C_{llr}) results [5].

3. RESULTS

Nine comparisons were performed: three within-task, two within-session across separate tasks, and four between-sessions. Figure 1 shows the system performance in each iteration, evaluated in terms of validity using C_{llr} and EER (the higher the values the poorer the system performance). The axes have been foreshortened in order to visualise the small scores.

3.1. Within-task results

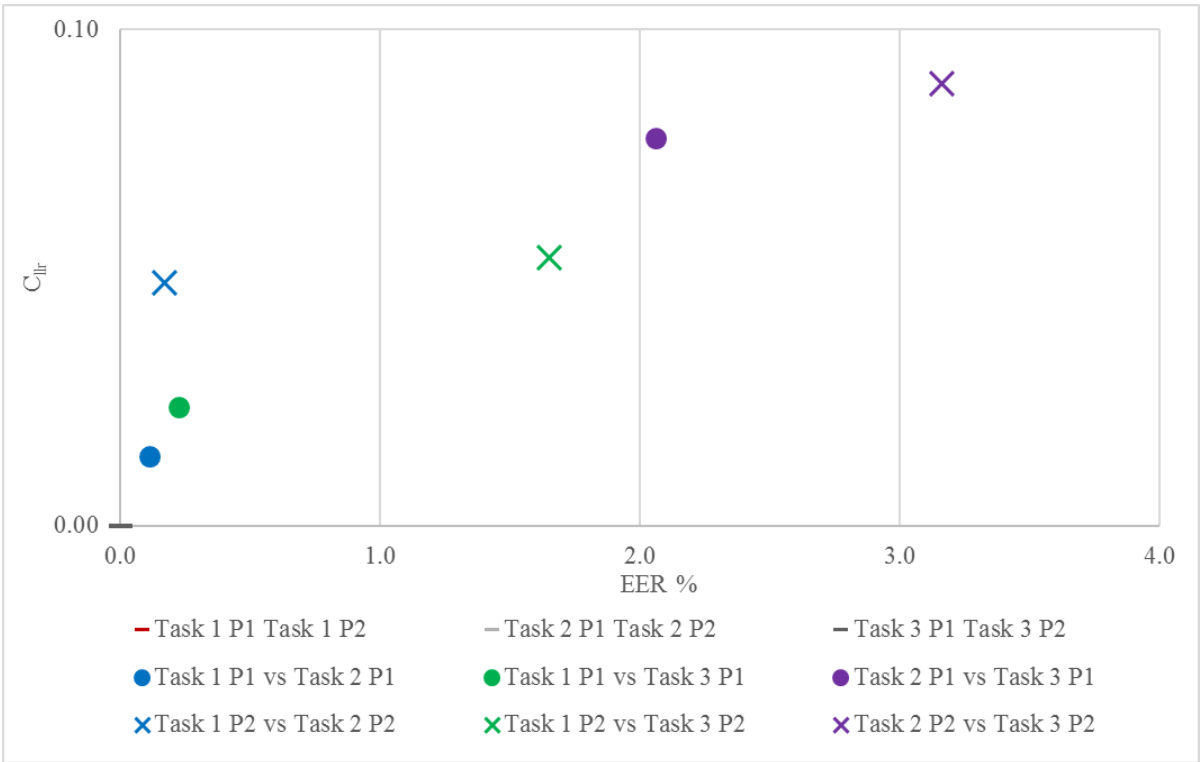
Figure 1 shows that all three of the contemporaneous comparisons, comparing the two halves (P1, P2) of Tasks 1, 2 and 3, respectively, resulted in a EER of 0% and a C_{llr} value of <0.001. This means that all same speaker pairs and different speaker pairs were correctly identified, signalling that there were greater levels of inter-speaker variation than intra-speaker variation within the individual tasks. These results are not considered to be surprising when taking into account the fact that they were obtained using samples that matched in terms of technical quality (they were all high quality studio recordings containing little to no background noise), and were obtained using a state-of-the-art i-vector framework [6, 7].

It should be noted here that the specific EER and C_{llr} values obtained are not the focus, as we are not testing how well the i-vector framework works in general; rather, it is the relative system performance in the subsequent comparisons that we are interested in, as these will demonstrate the effect of using mismatched data in terms of task and non-contemporaneity.

3.2. Within-session, between-task results

Within-session, between-task results were initially obtained by comparing the first half of Task 1 with the first half of Task 2. The recording tasks in the first WYRED recording session were completed one after the other, with a gap of approximately five minutes between them while instructions were provided for Task 2. The set-up remained the same, in that the participant remained in the same seat wearing a

Figure 1. Multi-file comparisons of performance of all iterations based on EER and C_{llr} values



headset, but a wireless telephone was introduced and the research assistant left the booth before Task 2 began. Task 1 and Task 2 involved different female interlocutors. Although these tasks could be considered to be "contemporaneous", in Figure 1 it can be seen that system performance reduces compared to the within-task comparisons, as the EER and C_{llr} values increase slightly. This may be the result of differences in speech style elicited between Tasks 1 and 2. If this is the case, these differences may have been more extreme by the second half of the two tasks, as system performance decreases further when we compare the second halves of Task 1 and 2. However, the EER and C_{llr} values remained relatively low. Levels of intra-speaker variation between-tasks are greater than those within-tasks, however, levels of inter-speaker variation still allow for the correct identification of same and different speaker pairs in the most comparisons.

3.3. Between-session results

Four between-session comparisons were made, comparing the halves of Task 1 and Task 3, and then Task 2 and Task 3. The comparisons resulted in the first half of the tasks performing better than the second half, but Task 2 compared with Task 3 produced the highest values of both EER and C_{llr}.

3.3.1. Task 1 vs. Task 3

Task 1 and Task 3 were recorded at least six days apart, but were recorded in the same sound booth, using the same equipment, and involved face-to-face interactions. The style of recording varied as Task 1 involved a high number of closed questions that could be answered by referring to a map task in front of them. Task 3 began using prompt cards with open ended questions. Participants in Task 3 were all male, from the same boroughs, and often similar postcodes. The speech style was informal, often contained laughter, and sometimes included mimicked speech.

The C_{llr} and EER values of the first half of Task 1 and Task 3 are relatively comparable with the comparison of first half of Task 1 and Task 2 with EER <0.05% and C_{llr} <0.05. However, when the second half of Task 1 and Task 3 were compared, the C_{llr} value remained stable but the EER value notably increased. This may reflect the unscripted nature and inherent variation present in Task 3, as participants relaxed and used prompt cards less frequently in the latter half of the recording. Furthermore, the length and content of speech in Task 3 per participant was less controlled than Task 1, lending more potential variation between speakers. The levels of intra-speaker variation increased between-sessions to a greater extent in the second half of the tasks.

3.3.2. Task 2 vs. Task 3

Both comparisons of Task 2 and Task 3 yielded higher EER and C_{llr} values than any other comparison, indicating the greatest values of intra-speaker variation. The higher overall values could reflect that the variation in between-session recordings are emphasised when speech style differs. In Task 2 the speaker is in isolation using a telephone rather than participating in a face-to-face interaction.

The comparison of the second half of the recordings follow a similar pattern to Task 1 versus Task 3, as the EER degrades in the latter part of both tasks. This emphasises the importance of sampling throughout the entirety of a recording to reflect the variation found within a single speech sample.

4. DISCUSSION

It should be noted that there are limitations within the dataset used for this investigation. The length of recordings was not controlled, and this could have had an impact on the performance of the system [11]. Although Task 4 was omitted from the study due to its extremely short length and limited data, the other tasks varied in length, per speaker, and between tasks. The focus of this study was in the relative performance of the high-quality studio recordings between and within sessions.

5. CONCLUSION

This paper has highlighted the variation that occurs within a recording session, between two separate tasks, and across recording sessions that were recorded on different days. Furthermore, the extent to which the comparison results vary is dependent upon which portion of the recording has been chosen for analysis. The results have been found to degrade when the second half of the recordings were compared. This is especially apparent when the nature of the task elicits stylistically varied speech due to its unscripted nature. As speaking styles differ, the level of variation in speech appears to increase. The influence of speech style is emphasised further when comparing non-contemporaneous recordings.

In order to more accurately reflect forensic speaker comparison cases, future research would benefit by controlling for file lengths, the amount of data present, and introducing extrinsic factors such as channel mismatch. However, it is important that speaking style is not overlooked as a factor. It is also important for phonetic and sociophonetic research that the data analysed from a task or across tasks is sampled over the entirety of a recording, and not limited to a specific selection of the task, in order to get a better overall picture of a speaker's voice.

6. REFERENCES

[1] Alexander, A., Forth, O., Aryal, A., Kelly, F. 2016. VOCALISE: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features. *Proc. Odyssey*, Bilbao.

[2] Bio-Metrics 1.8 performance metrics software, Oxford Wave Research Ltd., http://www.oxfordwaveresearch.com/products/bio-metrics

[3] Boyd, Z., Elliott, S., Fruehwald, J., Hall-Lew, L. Lawrence, D. 2015. An Evaluation of Sociolinguistic Elicitation Methods. *Proc 18th ICPhS*, Glasgow.

[4] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer [Computer program]. Version 6.0.43, retrieved 8 September 2018 from http://www.praat.org/

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

[6] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P. 2011. Front-end factor analysis for speaker verification, *IEEE Transactions on Audio, Speech & Language Processing*, 19, 4, 788–798.

[7] Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J., Niemi, T. 2015. *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition*, Frankfurt: Verlag für Polizeiwissenschaft.

[8] Enzinger, E., Morrison, G. S. 2012. The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems. *Proc. 14th Australian International Conference on Speech Science and Technology*, Sydney, 137-140.

[9] Gold, E., French, P. Submitted. International Practices in Forensic Speaker Comparisons: Second Survey. *International Journal of Speech, Language and the Law.*

[10] Gold, E., Ross, S., Earnshaw, K. 2018. The 'West Yorkshire Regional English Database': Investigations into the Generalizability of Reference Populations for Forensic Speaker Comparison Casework. *Proc. Interspeech, Sep 2-6 2018*, Hyderabad, 2748-2752.

[11] Hasan, T., Saeidi, R., Hansen, J. H., van Leeuwen, D. A. 2013. Duration mismatch compensation for ivector based speaker recognition systems. *ICASSP*,Vancouver, 7663-7667.

[12] Koschwitz, J. 2018. *The effect of speaking style on the performance of a forensic voice comparison system*. Master's thesis, Uppsala University.

[13] Morrison, G. S., Ochoa, F., Thiruvaran, T. 2012. Database selection for forensic voice comparison. *Proc. Odyssey. The Language and Speaker Recognition Workshop*, Singapore, 74-77.

[14] Morrison, G. S., Rose, P., Zhang, C. 2012. Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Australian Journal of Forensic Sciences*, 44, X, 155-167.

[15] Nolan, F. 1983. *The Phonetic Bases of Speaker Recognition.* Cambridge: Cambridge University Press.

[16] Nolan, F., McDougall, K. de Jong, G., Hudson, T. 2010. The DyViS database: style-controlled recordings of 100 homogenous speakers for forensic phonetic research, *International Journal of Speech, Language and the Law*, 17, 1, 143-152.

[17] Rhodes, R. 2012. *Assessing the strength of non-contemporaneous forensic speech evidence.* PhD thesis, University of York.

[18] Rose P., Zhang, C. 2018. Conversational Style Mismatch: its Effect on the Evidential Strength of Longterm F0 in Forensic Voice Comparison. *Proc ASSTA*, Sydney, 157-160.

[19] Wormald, J. 2016. *Regional Variation in Panjabi-English.* PhD thesis, University of York.