A Likelihood ratio-based forensic voice comparison in microphone vs. mobile mismatched conditions using Japanese /ai/

Michael J. Carne

Introduction

Forensic voice comparison (FVC) involves the analysis of one or more unknown speech samples (the offender) with known samples (the suspect) to determine how much more likely the observed differences/similarities between the two samples are to arise under the hypothesis they are of the same origin versus the hypothesis they are of different origins. A common approach in FVC is to compare speakers with respect to the F-pattern, which encodes information about a speaker’s physiological and articularatory settings during the production of speech sounds. Forensic analyses based on formants are, however, not without complications. A critical issue is the susceptibility of formants to transmission effects and in particular mobile telephony. In modern global telecommunications mobile telephony is almost universally preferred over landline systems, not only for legitimate purposes but also to conduct criminal activities. The operation of mobile networks differs substantially from landline systems in transmission of the speech signal, as do the impacts on speech parameters important to forensic voice comparison (FVC). This presents problems when comparing samples in channel mismatch conditions (e.g. mobile phone versus microphone recordings). This occurs frequently in FVC: typically the offender sample(s) is obtained via a phone intercept, while the suspect sample(s) is recorded via microphone in a police interview.

The acoustic differences between channels can be rather large and statistically significant enough to impede the accuracy of speaker discrimination systems [1]. These acoustic differences appear greatly exacerbated in mobile systems [2,3,4]. In addition, heuristics relating to the selection of reliable features in formant-based analyses in landline mismatches [1,5,6] may be difficult to directly apply to mobile conditions, given a host of distinctive factors introduced by mobile transmission. The principal distinction lies in the operation of codecs. The effect of codecs is to implement varying bit rates during transmission of the speech signal. Codecs also implement algorithms to handle background noise, lost or corrupted frames and silence [7]. In concert, these factors mean not only are components of the speech signal containing speaker-specific information lost, but importantly that the reproduction of individual acoustic parameters is constantly changing [3]. Despite the impact mobile effects may have on the reliability and validity of FVC analyses, relatively little work assessing its impacts on formant-based analyses exist within the Likelihood Ratio-based framework (described in these key publications [8, 9, 10]). There are however a few exceptions [11,12,13]. To date, almost all of the very few studies have involved testing English speech parameters — the author is aware of only one exception for Chinese [13]. This paper examines how the validity of a formant-based analysis is affected under mobile transmission conditions for the first time in Japanese.

This was the principal aim of this paper using the vowel combination /ai/ from (‘hai’, ‘yes’). Since we do not know a priori the performance of /ai/ in the selected population, it is impossible to attribute system performance (good or bad) to the effects of mobile telephony without a control to determine the base-line performance of the parameter. Fortunately, Japan’s National Research Institute of Police Science (NRIPS) speech database provides just that. The database consists of a large number of male speakers recorded simultaneously on mobile phone and microphone. Because of this, the ratio of within-to-between speaker variation is constant and the effect of the transmission assessed directly. System performance is therefore tested separately on both a matched group (microphone vs. microphone) and a mismatched group (microphone vs. mobile) allowing the differences in validity to be assessed. It is important to highlight at this point that the mobile phone calls were made to a landline telephone and recorded (mobile-to-landline), so the mismatch is between microphone versus mobile-to-landline speech data. Mobile-to-mobile phone calls would produce different effects on performance, as experimental work has shown [13].
The multivariate kernel density likelihood ratio formula [14] was employed to perform 30 same-speaker and 435 different-speaker trials from non-contemporaneous speech samples of /ai/. The likelihood-ratio cost function \( C_{ij} \) and Tippet plots were used to compare the performance of the FVC system under the matched and mismatched conditions. Two- and three-formant combinations of the F-pattern (F1,F2) and F1,F2,F3) were also tested separately to test the hypothesis that lower frequency formants may in fact be better preserved in mobile conditions [2,4], due to a lower band pass cut-off in mobile systems (~100Hz to ~2800-3600Hz [15], versus ~300Hz to ~3500 Hz in landline systems [1]).

2. Method

2.1. Reference and test data

The data used in this experiment was obtained from the National Research Institute of Police Science (NRIPS) speech database. The NRIPS mobile recordings (10 kHz with 12 bit quantisation) were made on a FOMA N902i (NEC) to a landline and recorded through a phone adapter (MRH-441W/N). Two replicates of /ai/, from two non-contemporaneous sessions from 30 speakers were measured for each of the recording conditions (i.e microphone and mobile-to-landline). The matched group consisted of a reference population of the microphone recordings, which was in turn partitioned into speakers session 1 (S1) recordings (representing the suspect recordings) and session 2 (S2) recordings (offender recordings). The channel mismatch group replicated this, except speaker’s S2 recordings were replaced with the corresponding mobile recording to simulate the scenario where the offender sample consists of a mobile recording and the suspect a microphone recorded police interview.

2.2. Formant measurements

Often acoustic differences in formant measurements obtained from different channels stem from formant tracking errors as a result of weak or distorted signal energy in transmission [1]. Therefore it is worthwhile highlighting the measurement procedure in some detail. Speakers’ formant trajectories where extracted using Praat over the vocalic segment of /hai/. The segmentation procedure for sampling defined vocalic onset as where periodically excited formant structure became apparent following a burst of high-frequency turbulent noise associated with /h/, the offset was defined as the absence of periodicity. Praat formant settings were adjusted by the experimenter in the mobile-to-landline recordings, as it was apparent that this resulted in a closer alignment with the true location of F2 in many cases. This is demonstrated in Figure 1, where we can see a reduced amount of spurious formant tracks in the three versus four formant setting. The corresponding microphone spectrogram is included as well (top panel, Figure 1), so readers can appreciate the differences in formant tracking and spectral features between channels. Better alignment of F2 is at the cost of F3 tracking though, which is tracked slightly higher than the centre of the high amplitude region at ~2800Hz which (very faintly) indicates the location of F3. Weak signal energy above ~2250 Hz in the mobile-to-landline recordings rendered measurement of F3 extremely difficult. Noteworthy too is the absence of energy above 3750 Hz in the mobile-to-landline recordings, meaning higher frequency formants such as F4 are, and in particular the extent to which they impede the accuracy of the forensic voice comparison, will become apparent following multivariate kernel density likelihood ratio testing.

Processing involved four stages: feature extraction (described above), parameterisation, multivariate kernel density likelihood ratio (MVKD-LR) estimation and calibration. Parameterisation involved modeling speakers time-normalised formant trajectories to cubic polynomial functions, the coefficients from which can readily be used input parameters for MVKD-LR estimation as described in [16]. The MVKD formula assesses the difference between suspect and offender samples with respect to their typicality in reference to a background reference population. Within-speaker variance is modelled via a normal distribution, while a non-Gaussian distribution (kernel density) models between speaker variance. MVKD-LR formula was then implemented to generate Log10likelihood ratios for same- and different-speaker trials. MVKD-LR tests were performed and Log10 LRs derived from each of the groups outlined in Section 2.1 and separately again for the two- and three-formant analyses. The fourth stage of processing calibrated the raw LRs using a logistic regression fusion technique [17]. Leave-one-out cross-validation (LOOCV) was applied, though the results of an intrinsic comparison (non-cross-validated comparisons)
where no speakers were removed are also included. The likelihood-ratio cost function ($C_{llr}$) given in (1), was used as a measure of system performance in terms of its accuracy (validity).

$$C_{llr} = \frac{1}{2} \left( \frac{1}{N_s} \sum_{i=1}^{N_s} \log_{2}(1 + \frac{1}{LR_{di}}) - \frac{1}{N_d} \sum_{i=1}^{N_d} \log_{2}(1 + LR_{di}) \right)$$

(1)

Where: $N_s$ and $N_d$ are the number of same-speaker (SS) and different-speaker (DS) comparisons; and $LR_{di}$. $LR_{di}$ are the derived likelihood ratios. $C_{llr}$ provides a metric assessing information loss in bits. $C_{llr}$ is seen to penalise systems for high contrary to fact LRs (i.e systems with poor validity). Optimum validity is achieved when $C_{llr} = 0$ and decreases as $C_{llr}$ approaches and exceeds 1. Tippet plots were also produced as a measure of performance and show the cumulative proportion of SS and DS trials correctly classified, as well as the magnitude (strength of evidence) of LRs in support of SS and DS hypotheses.

### 3. Results and discussion

$C_{llr}$ values for cross-validated (CV) and non-cross-validated (non-CV) LRs are both presented due to there being relatively large differences after the cross-validation procedure (Table 1). Non-CV tests in general showed far better performance, evident from the Tippet plots as well as superior $C_{llr}$ values.

<table>
<thead>
<tr>
<th>group</th>
<th>CV</th>
<th>Non-CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1-F2MATCHED(CV)</td>
<td>0.77</td>
<td>0.28</td>
</tr>
<tr>
<td>F1-F2MISMATCH(CV)</td>
<td>0.85</td>
<td>0.48</td>
</tr>
<tr>
<td>F1-F2-F3MATCHED(CV)</td>
<td>0.79</td>
<td>0.19</td>
</tr>
<tr>
<td>F1-F2-F3MISMATCH(CV)</td>
<td>0.93</td>
<td>0.34</td>
</tr>
</tbody>
</table>

This is not unusual in cross-validation of the MVKD-LR where relatively small test and reference populations are used. It is almost certainly due to a lower ratio of between-to-within speaker variance as a result of the reduced population size after CV removes speakers from the reference population.

There are two main effects introduced by channel mismatch relative to the performance of the matched microphone data. These are: (1) the magnitude of Log$_{10}$LR is reduced in support of both hypotheses, with a corresponding increase in magnitude of counter-factual LRs and; (2) $C_{llr}$ values in turn point to a decline in system validity (i.e. accuracy). The first effect can be judged by observing the cumulative proportion of LR trials from the Tippet plots in Figure 2 (top and bottom panel show F1-F2 and F1-F2-F3 analyses respectively). The magnitude of the Log$_{10}$LR is proportional to the strength of evidence, with values < 0 favouring the DS hypothesis and those > 0 the SS hypothesis. The closer $\pm$Log$_{10}$LR values are to zero the more likely they are under both hypotheses; and thus the weaker their evidentiary value. From the Tippet plots we observe that LRs in the mismatched trials (red-lines) saturate much faster towards zero compared to the matched group (blue-lines), indicating a weakening of evidence in favour of both hypotheses. Indeed for the two-formant analysis (top panel Figure 2), the smallest microphone matched DSLR was Log$_{10} = -4.4$, this reduced to -1.8 in the mismatched conditions. For the SSLRs the control best matched comparison belonged to Speaker 17 ($Log_{10}$LR = 0.77), in the mismatch condition this is again reduced ($Log_{10}$LR = 0.54). The addition of F3 appears to be even more injurious to strength of evidence, an effect witnessed by a further narrowing of Log$_{10}$LRs towards zero (see bottom panel Figure 2).

Figure 2: Tippet plots. Top panel = F1-F2 analysis. Bottom panel = F1-F2-F3 analysis. x-axis = Log$_{10}$LRs, y-axis = 1 – the cumulative proportion of same and different speaker trials. Solid blue lines = cross-validated matched trials, solid red lines = cross-validated channel mismatch trials. Corresponding dotted lines show non-cross-validated results.

The main term governing the magnitude of the LR is the ratio of variance between the suspect and offender samples relative to the distributional characteristics of the reference population. The larger this is, the larger the magnitude of the resultant LR [8]. The mismatched conditions would therefore appear to reduce this ratio. This is because the distance between the sample means of the suspect (microphone) and offender (mobile) samples is likely greater in the mismatched group due to differences introduced by the transmission. As this distance increases, so too does the variance between the samples. If the variance in the reference population remains constant and the variance between the samples increase, the ratio of variance (reference population/suspect-offender variance) will be lower – and hence why the magnitude of LRs are diminshed. This
affects system validity by the reducing strength of evidence in the mismatched trials and, more importantly, increasing the magnitude of counter factual LRs. Consequently, this increases the penalty $C_{\text{in}}$ applies to the system, indicating reduced validity. For the two-formant and three-formant analyses the $C_{\text{in}}$ difference equates to a reduction in system validity of 10\% and 18\% respectively. This result suggests that F3 may be more affected by the transmission; hence a larger impact on $C_{\text{in}}$ is seen. This is likely related to the measurement difficulties described in Section 2.2, where weak signal energy in regions above ~2250 Hz made F3 very difficult to locate and measure precisely in many instances. There are almost certainly non-linear distortion effects resulting from the operation of the mobile codec at play as well, which are impacting all three formants. Such effects we would expect to have a greater impact than the band-pass cut-off effects observed in the spectral analysis earlier. Codec mechanisms for handling frame loss and corruption, noise suppression, silence frames, background noise and variations in the dynamic coding rate are more than likely introducing additional acoustic variation between channels [7]. Overall though, the relative differences observed between the matched and mismatched conditions appear rather minor – a 0.08 bit $C_{\text{in}}$ increase in for F1-F2 combination and a 0.14 bit increase with the inclusion of all three formants.

The reductions in validity (i.e. increases in $C_{\text{in}}$) seen in the cross-validated results are relatively small compared to the findings of other studies of mobile transmission effects on validity in FVC systems. In a study which involved testing the performance of formant trajectories from 60 Chinese female speakers in various combinations of mobile mismatched conditions [13] the microphone versus mobile-to-landline results yielded a $C_{\text{in}}$ of 0.216 bits, compared to a $C_{\text{in}}$ value of 0.008 in a group high-quality studio recordings (i.e. a 0.208 bit difference). More recent work examining the effects of various mobile phone codecs [12] saw reductions in validity of up to 170\% relative to a control group of studio recordings. The non-cross-validated results in the current study are, however, much more closely aligned with these previous findings. Non-cross-validated results evince far larger reductions in validity in the channel mismatch relative to the matched group. These were 0.28 vs. 0.48 bits for F1-F2 (71\% reduction) and 0.19 vs. 0.34 bits (79\% reduction) for the F1-F2-F3 analysis. This suggests that the limited sample size in the current study may be responsible for the relative small percentage reductions in validity seen for the cross-validated results.

4. Summary, conclusions and limitations

The aim of this paper was to investigate the performance of the F-pattern of /ai/ in mobile-to-landline vs. microphone channel mismatch in a forensic voice comparison system. The results should be taken as preliminary due a number of experimental limitations that will be discussed shortly. Despite these, it is possible to make some conclusions about the parameter investigated. In optimal conditions (i.e. the matched group) the parameter /ai/ yielded cross-validated $C_{\text{in}}$ values of 0.77 and 0.79 for the two- and three- formant analyses respectively. These values are much larger than to be hoped for, indicating that /ai/ is a reasonably poor discriminator in terms of its accuracy – although the $C_{\text{in}}$ values are still <1, meaning the system is still retaining some information. However, our interest here was not in the performance of /ai/ per se, but rather how the parameter performs in mobile-to-landline mismatched conditions. Thus, the baseline performance in the optimal condition was compared with that of mismatched recordings. Cross-validated results from the mismatched group show validity is indeed reduced, though not substantially. These were 10\% (two-formants) and 18\% (three-formants). Based on this we might conclude the parameter is relatively robust to channel differences. Although this is perhaps no great gain to FVC given reasonably poor performance in optimal conditions.

The accuracy of the likelihood ratio depends a great deal on the accuracy of measurements obtained from the speech signal. Spectral analysis of F3 for /ai/ indicated the formant should probably be excluded from the analysis. Measurement errors resulting from poor formant tracking of F3 were seen to reduce the accuracy of LR estimates, manifest in a higher log-likelihood ratio cost relative to the two-formant analysis. This provides some evidence that lower frequency formants may indeed be better preserved for FVC analysis in mobile conditions. A more representative sample though, investigating parameters other than /ai/, is needed to test this hypothesis more comprehensively. This is the first limitation. A second relates to the suggestion that non-linear distortions are impacting system validity as well, perhaps more so than band-pass filter effects. There is a wide range of possible mobile transmission conditions that may be present at any one time or location that affect the level of background noise and interference. These in turn effect how the codec impacts the speech signal. Network type also plays a role, contributing to this variety and manifest in small but important differences in the codec’s speech handling mechanisms [18]. As such, this experiment only represents a study of a subset of all the possible mobile transmissions effects we might expect in real-world mismatch scenarios. A future task is to better enumerate these effects on formant-based analyses using methodologies such as those described in [18]. A final limitation is the size of the reference population. The fact that removal of a few speakers, as part of the LOOCV procedure, altered the magnitude of LRs considerably, points to a highly variable ratio of between-to-within speaker variation depending on the speakers included in the reference population. The sample size of the reference population is crucial to accurate modeling of the probability densities of within- and between-speaker variation and deriving accurate likelihood ratios. It has been suggested that at least 70 speakers and greater than two replicates per speaker are required in this regard [19,20]. It is possible, therefore, that the limited sample size is leading to an inaccurate estimation of system validity. This might explain why much smaller impacts associated with the mismatch were seen in the cross-validated results than perhaps expected. A study involving a larger reference population will therefore also be required to ensure accurate modeling of probability densities and better validate the effects of the mismatch on performance.

5. Acknowledgements

This work would not have been possible without Japan’s National Research Institute of Police Science (NRIPS) database, nor Dr. Shunichi Ishihara who has hosted the author at the ANU and provided comments on an earlier incarnation of this paper. Thanks are owed as well to Dr. Phil Rose whose R code was used for the experiment.
6. References


