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

Laboratory Report: Human-supervised and fully-automatic formant-trajectory measurement for forensic voice comparison – Female voices[*]

Cuiling Zhang[a,b], *Geoffrey Stewart Morrison*[b,**], *Ewald Enzinger*[b], *Felipe Ochoa*[b]

[a]Department of Forensic Science & Technology, China Criminal Police University, Shenyang, China

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

27 September 2012

Abstract

Acoustic-phonetic approaches to forensic voice comparison often include analysis of vowel formants. Such methods typically depend on human-supervised formant measurement, which is often assumed to be relatively reliable and relatively robust to telephone-transmission-channel effects, but which requires substantial investment of human labor. Fully-automatic formant trackers require minimal human labor but are usually not considered reliable. This study assesses the effect of variability within three sets of formant-trajectory measurements made by four human supervisors on the validity and reliability of forensic-voice-comparison systems in a high-quality v high-quality recording condition. Measurements were made of the formant trajectories of /iau/ tokens in a database of recordings of 60 female speakers of Chinese. The study also assesses the validity of forensic-voice-comparison systems including a human-supervised and five fully-automatic formant trackers under landline-to-landline, mobile-to-mobile, and mobile-to-landline conditions, each of these matched with the same condition and mismatched with the high-

[*]This research received support from multiple sources, including the following: The Australian Research Council, Australian Federal Police, New South Wales Police, Queensland Police, National Institute of Forensic Science, Australasian Speech Science and Technology Association, and the Guardia Civil through Linkage Project LP100200142. The China Scholarship Council State-Sponsored Scholarship Program for Visiting Scholars. The Ministry of Education of the People's Republic of China "Program for New Century Excellent Talents in University" (NCET-11-0836). An International Association of Forensic Phonetics and Acoustics Research Grant. Unless otherwise explicitly attributed, the opinions expressed are those of the authors and do not necessarily represent the policies or opinions of any of the above mentioned organizations. Earlier versions of this paper were presented at the Special Session on Forensic Acoustics at the 162nd Meeting of the Acoustical Society of America, San Diego, November 2011 [J. Acoust. Soc. Am. 130, 2519. doi:10.1121/1.3655044], and at the 21st Annual Conference of the International Association for Forensic Phonetics and Acoustics, Santander, August 2012.

[**]Author to whom correspondence should be addressed. geoff-morrison@forensic-voice-comparison.net.

quality condition. In each case the formant-trajectory systems were fused with a baseline mel-frequency cepstral-coefficient (MFCC) system, and performance was assessed relative to the baseline system. The human-supervised systems always outperformed the fully-automatic formant-tracker systems, but in some conditions the improvement was marginal and the cost of human-supervised formant-trajectory measurement probably not warranted.

1 INTRODUCTION

Measurement of vowel formant frequencies is a popular technique in forensic voice comparison; in a survey of 34 forensic-voice-comparison practitioners conducted by Gold and French (2011) 30 reported making use of formant measurements. Recordings provided to forensic-voice-comparison experts for analysis are often of poor quality, a typical scenario being that the offender recording comes from a telephone intercept and the quality of the speech signal is degraded by the telephone-transmission system. The suspect recording may be of better quality, e.g., a direct-microphone audio recording of a police interview. There would appear to be a general assumption within the acoustic-phonetic forensic-voice-comparison community that the second formant (F2) at least is relatively robust to channel effects and thus usable for forensic voice comparison under channel-mismatch conditions. A number of studies have examined the effects of telephone-transmission systems on formant measurement and given warnings about the degradation caused by such transmission systems (see especially Byrne & Foulkes, 2004; and Künzel, 2001).

Another concern related to the use of formant measurements for forensic voice comparison is the degree of reliability of formant measurement, even for human-supervised measurement under good recording conditions (Duckworth et al., 2011; Jessen, 2010). That the performance of fully automatic formant trackers is relatively poor is widely recognized (Chen et al., 2009; Deng et al., 2006; Remez, et al., 2011; Vallabha & Tuller, 2002). Most phoneticians employ human-supervised methods usually involving having a human expert select parameter settings for signal-processing algorithms and then comparing the results with an alternative form of analysis, such as overlaying the measured formant tracks on a spectrogram – the formant measurements are typically based on linear predictive coding (LPC, see Vallabha & Tuller, 2002, for a review) and the spectrogram is typically based on Fourier analysis. If not happy with the initial results, the phonetician tries different parameter settings. The human supervision is intended to produce more valid and reliable results, but probably most phoneticians would agree that the process depends to some extent on experience-based subjective judgment and that repeated measurements will typically result in different values (hopefully only slightly different values). See Byrne & Foulkes (2004), Duckworth et al. (2011), Harrison (2004), Hillenbrand et al. (1995), Kirchübel (2010), and Künzel (2001) on difficulties in human-supervised measurement of formant values.

The study reported on the present paper examines the reliability of human-supervised formant-trajectory measurement on high-quality recordings. We also examine the validity of human-supervised and fully-automatic formant-trajectory measurements as part of a forensic-voice-comparison system,

using both high-quality recordings and degraded versions of the same recordings. Recordings were degraded by passing them through landline-telephone and mobile-telephone transmission systems. The trajectories of the first, second, and third formants (F1, F2, F3) of tokens of Chinese /iau/ were measured up to three times each by up to four human supervisors. At least in its canonical form, the formant trajectory of /iau/ covers a large part of the vowel space. The tokens were extracted from a database of natural speech (not read speech) produced by 60 female speakers of Standard Chinese (Zhang & Morrison, 2011).

The choice of female speakers was due to a suitable database of female but not male speakers being available when the study began. The present study could be replicated using a database of male speakers. Note that the high fundamental frequencies (f0) typical of female speakers result in widely-spaced harmonics leading to sparse sampling and difficulty in measuring the spectral envelope due to the resonance properties of the vocal tract (see Vallabha & Tuller, 2002, on quantization errors due to harmonic spacing, and Assmann & Nearey, 1987, on the relationship between harmonics, LPC analysis, and perception of F1).

This study is concerned with dynamic formant trajectories rather than so-called "steady-state" measurements. This forms part of a line of investigation assessing the effectiveness of formant trajectories for forensic voice comparison (e.g., McDougall, 2006; Morrison, 2009a, 2011a, 2012b; Zhang, Morrison, & Thiruvaran, 2011). A whole-trajectory approach obviates the problem of where in the vowel to measure the "steady state" and the effect of this decision on the reliability of formant measurements (as encountered in Duckworth, et al., 2011), although decisions instead need to be made as to where to begin and end measuring the formant trajectories. In phonetic research in general, there is a growing recognition of the importance of formant dynamics and the inadequacy of "steady-state" representations (Morrison & Assmann, 2012).

In the present study, as well as directly assessing the reliability of the actual formant measurements (both within-supervisor and between-supervisor variability), the effect of the variability in these measurements is assessed with respect to the validity and reliability of forensic-voice-comparison systems which use formant-trajectory measurements as input. It should be pointed out that the accuracy per se of the formant measurements is not particularly important for their use as part of a forensic-voice-comparison system as long as the formant-measurement procedure produces precise results. For example, a formant tracker which always gave measurements which were 100 Hz too high compared to the true vocal-tract resonances would work just as well as one which, all else being equal, gave accurate measurements. In contrast, a formant tracker which gave unreliable (imprecise, inconsistent) measurements would result in poor performance for the forensic-voice-comparison system.

The present study also compares the validity of forensic-voice-comparison systems based on human-supervised formant tracking with forensic-voice-comparison systems based on five different fully-automatic formant trackers (WAVESURFER; PRAAT; Nearey, Assmann, & Hillenbrand, 2002; Mustafa & Bruce, 2006; Rudoy, Spendley, & Wolfe, 2007). For both human-supervised and fully-

automatic systems, performance is assessed as the degree of improvement over a mel-frequency-cepstral-coefficient (MFCC) Gaussian-mixture-model–universal-background-model (GMM-UBM) system applied to the entire speech-active portion of each recording (e.g., Reynolds, Quatieri, & Dunn, 2000). There have been some previous attempts to compare the performance of human-supervised and automatic formant tracking as part of a forensic-voice-comparison system. A study by de Castro, Ramos, and González-Rodríguez (2009) found that before feature selection a forensic-voice-comparison system based on human-supervised formant tracking (Morrison, 2009a) outperformed one based on automatic tracking (Rudoy, Spendly, & Wolfe, 2007), but after automatic feature selection the automatic system could outperform the non-feature-selected human-supervised system; however, the size and content of the database tested was limited (human-marked tokens of diphthongs in read speech produced by 27 speakers), and feature selection and testing were conducted using the same database. Some studies (Becker, Jessen, & Grigoras, 2008, 2009; Hansen, Slyh, & Anderson, 2001; Moos, 2008; Nolan & Grigoras, 2005) have evaluated human-supervised or automatic formant tracking as part of a forensic-voice-comparison system or an automatic-speaker-recognition system but have measured formant values over the entire voiced portion of the speech recordings without regard for vowel identity. Such systems have not outperformed MFCC-based systems, either alone of after fusion with the latter (see Jessen, 2010).

Before presenting the detailed description of the present study, earlier studies on the effects of telephone-transmission systems on formant measurement are reviewed.

1.1 Effect of landline-telephone-transmission systems on formants

Landline-telephone systems have a bandpass of 300–3400 Hz (the exact values vary from country to country). Nominally within the specified range the amplitude passed by the system does not drop more than 3 dB compared to its maximum value. Frequencies far enough outside this band are lost, and frequencies near the edges of the band are distorted. Digital landline-telephone systems sample the signal at 8 kHz and apply lossless compression and decompression (Guillemin & Watson, 2008).

Künzel (2001) examined the effect of digital telephone and Integrated Services Digital Network (ISDN) lines on first and second formant values. He used speech produced by 10 male and 10 female German speakers, who read a written passage out loud. Formants were measured using human-supervised procedures. Formant frequencies were measured "at the centre of [the] steady-state portion" (p. 85) of each vowel token, both in speech recorded directly using a high-quality microphone and in recordings made simultaneously via a telephone connection. One speaker's vowels were measured a second time and no significant differences were found in F1 and F2 between the two sets of measurements (average deviation of 3.5 Hz for F1 and 7.0 Hz for F2). Mean differences in F2 between the direct-microphone and telephone conditions did not exceed 2% for any vowel phoneme, but for F1 the mean measured frequency in the telephone condition was more than 9% higher for a number of vowel phonemes with intrinsically low F1 (/iː/, /ɪ/, and /u/ for both male and female speakers, and /ʊ/

for male speakers) (see similar ISDN results in Trawińska & Kajstura, 2004). Smaller amounts of F1 raising were observed for most other vowel phonemes. The F1 measurements of the vowel phonemes with the highest intrinsic F1 (/a/ and /aː/) were least affected. Lawrence, Nolan, and McDougall (2008) conducted a study with 10 male English speakers which was similar in design to that of Künzel, except that the speech was elicited in a semi-spontaneous manner (e.g., map task with some of the words being read-out place names), and only tokens of /i/, /u/, and /æ/ were measured. Comparing the telephone-transmitted recordings with the direct-microphone recordings, across all speakers F1 for /i/ and /u/ were higher (23% higher on average for /i/ and 18% higher on average for /u/, the mean direct-microphone F1 measurements for these vowels were 308 and 322 Hz respectively), whereas there was no substantial consistent difference for /æ/. Significant differences were not found for F2 or F3 for any of the three vowel phonemes.

The results summarized above were generally as expected given the bandpass of a landline telephone system. They indicate that the use of F1 would probably be problematic in a high-quality versus landline mismatch system for all vowel phonemes with the possible exception of the vowels with the highest intrinsic F1. The use of F2, however, should not present difficulties (see also Nolan, 2002, and Künzel, 2002). Based on a study involving 3 male English speakers, Rose & Simmons (1996) suggested that the use of F3 will also not be problematic (see also Rose, 2003 §99.870–§99.920).

1.2 Effect of mobile-telephone-transmission systems on formants

Guillemin and Watson (2008) describe a typical mobile-telephone system (Groupe Spécial Mobile / Global System Mobile Communication network) and contrast it with landline-telephone systems. Mobile telephone systems apply substantial amounts of data compression to the speech signal, and the degree of data compression applied can change from moment to moment (theoretically as often as every 20 ms for the Adaptive Multi-Rate, AMR, codec, although in practice the compression level is held for at least 40 ms). Compression results in transmission rates which range from 4.75 to 12.20 kbits/s, in contrast to 64 kbits/s for digital landline systems. The system has a bandpass with a lower limit of ~100 Hz and an upper limit ranging, depended on the compression rate, from ~2.8 to ~3.6 kHz. Data are transmitted in sequential packets, and if a packet is too badly corrupted or lost the preceding packet may be repeated or extrapolated. Although the compression algorithms were designed to maximize speech intelligibility while minimizing data transmission rates, they may result in substantial changes to measured speech properties such as formant frequencies.

Byrne and Foulkes (2004) examined the effect of mobile-telephone transmission on first, second, and third formant values (F1, F2, F3). Although not made explicit in the paper, what Byrne and Foulkes actually tested were recordings made of speech transmitted from a mobile telephone to a landline (personal communication from Paul Foulkes, 5 December 2011). They used speech produced by 6 male and 6 female English speakers, who read a written passage out loud. Formants were measured

using human-supervised procedures. Formant frequencies were measured at "stable points close to the centre of each vowel" (p. 88), both in speech recorded directly using a high-quality microphone and in recordings made simultaneously via a telephone connection. Across all vowel phonemes and across male and female speakers the measured F1 values in the telephone condition averaged 29% higher than in the direct-microphone condition, ranging from a mean of 7% for female speakers' /ʊ/ to a mean of 60% for female speakers' /ɒ/ (for all phonemes except female speakers' /ʊ/ the mean difference exceeded 14%). In contrast, mean differences between measured F2 values in the direct-microphone versus telephone recordings never exceeded 10%. For F3 the difference did not exceed 7% for any phonemes except for /i/. The problem with /i/ appeared to be due to its intrinsically high F3 being around the upper limit of the passband. The mean measured frequency of /i/ in the direct-microphones was ~3400 Hz for females and ~3475 Hz for males and in the telephone condition it was 13% lower for females and 11% lower for males. Lowering of F3 was also generally greater for the speakers with the highest measured F3 values in the direct-microphone condition. The general pattern in the results was similar to that for the landline condition reported in Künzel (2001): The differences in F1 between the microphone and mobile-to-landline condition would generally preclude its use in such a channel mismatch condition. F2 and F3 were much less affected by the mobile-to-landline system than was F1. Although the variability in measured F2 resulting from transmission through the mobile-to-landline system was greater than the variability resulting from transmission through the landline-only system, F2 (and F3) measurements may still be usable for forensic voice comparison. Chen et al. (2009) used a fully automatic formant measurement system on a large number of landline and mobile recordings (1 224 hours of speech from 3 673 English speakers), albeit not recordings of the same utterances or even the same speakers in both recording conditions. Their results suggested that for a number of vowel phonemes for females and more so for males the mobile condition resulted in lower F2 measurements.

A number of other studies (Enzinger, 2011; Guillemin & Watson, 2008; Jímenez Gómez, 2011; Meinerz & Masthoff, 2011; Masthoff & Meinerz, 2012) have also looked at the effects of landline-telephone systems, mobile-telephone systems, or the AMR codec on formant measurement, and obtained patterns of results similar to those reported above.

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 (Zhang & Morrison, 2011). See Morrison, Rose, and Zhang (2012) for details of the data collection protocol. The speakers were all first-language speakers of Standard Chinese from northeastern China, and were aged from 23 to 45 (with most being between 24 and 26). The recordings used were from an information-exchange task conducted over the telephone: Each of a pair of speakers received a "badly transmitted fax" including some illegible information, and

had to ask the other speaker to provide them with the missing information. The original recordings were approximately 10 minutes long. The first and second recording sessions were separated by 2–3 weeks. High-quality recordings were made at 44 100 samples per second 16 bit quantization 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.

Stressed tokens of /iau/ on tone 1 were manually located and marked using SOUNDLABELLER (Morrison, 2010b). There were between 6 and 41 stressed tokens of /iau/ per speaker per recording, median 21.5. Unlike an earlier study (Zhang, Morrison, & Thiruvaran, 2011) in which all the /iau/ tokens were the realizations of a single word ("yao" *one*) those in the present study were taken from a wider range of contexts.

In the tests of forensic-voice-comparison systems described below, /iau/ tokens 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 Signal degradation

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: *landline to landline*, *mobile to mobile*, and *mobile to landline*. In each case the telephone used to transmit the signal was placed in a sound booth (IAC 250 Series Mini Sound Shelter) in the vicinity of a loud speaker (Roland® MA-7A) connected to a computer via an external sound card (Roland® UA-25 EX). The high-quality recordings were played through 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 telephone. The landline telephones used were a Leader 852 HS for transmitting and a Polaris NRX EVO 450 for receiving, and the mobile telephones were both Nokia 2730 classic. The receiving mobile telephone had a 3.5 mm audio-output jack and the signal from this output was fed into the USB sound card attached to the computer. The receiving landline telephone was connected to the external sound card via a Trillium Telephone Recording Adaptor Studio Interface (REC-ADPT-SI). For the landline to landline condition, the call was routed via the external telephone system not just within the university's internal telephone system.

The recordings in the database were played through the telephone system one at a time. Custom software was written which started recording, started playing the high-quality recording, then stopped recording 500 ms after the high-quality recording had finished playing. The degraded signal was recorded at the same sampling rate and quantization as the high-quality recording (44.1 kHz at 16 bits). The degraded recording was aligned with the high-quality recording by sliding the degraded signal past the high-quality signal in the time domain and calculating the correlation between the two signals at

each sample displacement. At the displacement with the highest correlation, the degraded signal was truncated to the same start and end points as the high-quality signal. Alignment allowed the use of the same /iau/ markers as were created using the high-quality recordings, the present study being concerned with variability due to formant measurement only (variability in phonetic-unit marking would also be expected to contribute to imprecision in the likelihood-ratio output of a forensic -voice-comparison system).

The pairs of channel conditions tested and reported in this paper are:

– high-quality v high-quality

– landline-to-landline v landline-to-landline

– high-quality v landline-to-landline

– mobile-to-mobile v mobile-to-mobile

– high-quality v mobile-to-mobile

– mobile-to-landline v mobile-to-landline

– high-quality v mobile-to-landline

In each case, the channel condition on the left was treated as the condition of the suspect (known identity) recording, and the channel condition on the right was treated as the condition for the offender (questioned identity) recording.

2.3 Formant measurement

Formant trajectories were measured using a human-supervised procedure and five fully-automatic procedures. The choice of automatic procedures was based on the ready availability of software implementations which could be operated in batch mode.

2.3.1 Human-supervised formant measurement (FORMANTMEASURER)

The trajectories of the first three formants (F1, F2, and F3) of each vowel token were measured using FORMANTMEASURER (Morrison & Nearey, 2011). This software is based on the formant tracking procedure outlined in Nearey, Assmann, and Hillenbrand (2002): The number of LPC coefficients was fixed at 9 to extract sets of 3 formant measurements, and at 11 to extract sets of 4 formant measurements. The sets of formant measurements were extracted below 8 different cutoff values in a specified range. The cutoff values were equally spaced on a logarithmic scale in the range 3 kHz to 4.5 kHz for the high-quality recording, a range a priori selected as likely to be appropriate for female speakers. For the telephone-channel degraded recordings, the range was set to be from 3 kHz to 3.75

kHz, the upper limit of the bandpass being ~3.4 kHz for the landline system and ~2.8 to ~3.6 kHz for mobile systems. Measurements were obtained every 2 ms using a 100 ms wide power-four-cosine window. Formants were tracked using the algorithm described in Markel and Gray (1976). Fundamental frequency tracks were also measured using the autocorrelation algorithm of Boersma (1993). Intensity was also measured. The formant-track sets were visually displayed overlain on a spectrogram. The measured intensity, fundamental frequency, and formant frequencies were used to synthesize a vowel. The human supervisor could listen to the original vowel and a synthesized vowel based on any desired selection of tracks (the human supervisors listened via a Roland® UA-25 EX external soundcard and AKG® K701 or K702 Reference Headphones). The software used a number of heuristics to suggest the best formant-track for each of F1, F2, and F3, and these were indicated on the visual display and used as the initial basis for vowel synthesis. On the basis of visual and auditory comparison, the human supervisor selected what he or she judged to be the best formant track for each of F1, F2, and F3. As a last resort the human supervisor also had the option of manually editing formant tracks. Use of this option was discourage and it was primarily used to correct tracking errors near the temporal edges of the vowel tokens.

For the high-quality recordings, each of four human supervisors (CZ, EE, FE, and GSM) measured the /iau/ tokens in both sessions of all 60 speakers three times. Tokens from all 60 speakers were measured once, then tokens from all 60 speakers measured a second time, then tokens from all 60 speakers measured a third time. For the telephone-channel-degraded recordings, CZ measured both sessions of all 60 speakers once.

2.3.2 Nearey, Assmann, and Hillenbrand (2002) tracker

The Nearey, Assmann, and Hillenbrand (2002) tracker (hereafter NAH2002) is at the core of the FORMANTMEASURER software described above, but is fully automatic. It does not include any of the human supervised elements, but uses a combination of heuristics to select the best trackset from the 8 different F1-F2-F3 tracksets (3 peaks extracted from a model using 9 LPC coefficients) obtained using the 8 difference cutoff values (the cutoff values were the same as those used in the human-supervised system described in section 2.3.1 above). The heuristics are:

1. Presence: To what extent are good candidates available to fill the time slots?

2. BwReason: Are the bandwidths of the peaks reasonable?

3. AmpReason: Is the amplitude reasonable?

4. ContReason: Is there reasonable continuity within each formant track?

5. DistReason: Are the F2-F1 and F3-F2 distances reasonable?

6. RangeReason: Are the formant ranges reasonable given the frequency cutoff?

7. RfStable: Are formant tracks relatively stable when the number of LPC coefficients is increased from 9 to 11?

8. Rabs: Correlation of resynthesized spectrogram with original.

For each of the heuristics an algorithm assigns a value between 0 and 1, and these values are then multiplied together to obtain an overall goodness score. The trackset with the best score is used.

2.3.3 WAVESURFER

WAVESURFER (Sjölander & Beskow, 2000, 2011) is software used by many phoneticians. It uses the SNACK SOUND TOOLKIT (Sjölander, 2004) for its basic functions. Its formant tracking algorithm is based on the dynamic programming approach of Talkin (1987) which obtains formant candidates from the roots of the LPC polynomials and subsequently chooses formant tracks based on (1) constraints on plausible ranges for each formant, and (2) the degree of continuity of the tracks measured using a Viterbi search. The signal is first resampled to 10 kHz and the number of linear prediction coefficients used is 12.

WAVESURFER has a parameter setting which is the expected F1 value given the vocal-tract length of the speaker, which is the basis for calculating the plausible ranges for each formant. For the experiments in the present study this was set to 567 Hz, on the basis of Eq. 1.

$$F1 = c / 4L \tag{1}$$

Where c is the speed of sound (set to 34 000 cm/s) and L is the length of the vocal tract (set to 14.98 cm, the average vocal-tract length for 20 adult female Chinese speakers' reported in Xue and Hao, 2006).

2.3.4 PRAAT

PRAAT (Boersma & Weenink, 2011) is probably the most widely used software among phoneticians. Formant tracking is performed using the Burg autocorrelation LPC algorithm (Anderson, 1978) and a Viterbi algorithm which penalizes deviation from reference formant center values and bandwidths, and jumps in those values.

For the experiments in the present study, the recommended parameter values for female speakers in the software documentation were adopted: Maximum frequency 5500 Hz, maximum number of formants 5 (i.e, 10 LPC coefficients),and formant reference values of F1 = 550 Hz, F2 = 1650 Hz, F3 = 2750 Hz, F4 = 3850 Hz, and F5 = 4950 Hz. Frequency deviation, bandwidth, and transition costs were all set to their default value of 1.0.

2.3.5 Rudoy, Spendley, and Wolf (2007) tracker

In the algorithm described in Rudoy, Spendley, and Wolf (2007) (hereafter RSW2007), first LPC cepstra are extracted from overlapping frames, then estimates of formant center frequencies and bandwidths are obtained from the LPC cepstra using a non-linear mapping function. These estimates are subsequently smoothed over time using a statistical model that constrains their temporal evolution (Rudoy, 2010). The RSW2007 tracker extends the procedure of Deng et al. (2007) by accounting for the uncertainty of the presence of speech as well as modeling cross-correlation of formants.

As in RSW2007, we use formant estimates obtained from WAVESURFER to empirically estimate model parameters such as formant cross-correlation.

2.3.6 Mustafa and Bruce (2006) tracker

The approach by Mustafa and Bruce (2006) (henceforth MB2006) is specifically designed for robust tracking of formants under adverse conditions such as noise. After preprocessing and Hilbert transformation of the signal, it is filtered by four adaptive filters (a combination of an all-zero filter and a single-pole dynamic tracking filter), separating the signal into four different bands. Within each band a formant is estimated from a first-order LPC analysis. The formant estimates are then used to adapt the poles and zeros of the band-pass filters for the next frame. This is repeated for every sample. Formant frequency estimates are conditioned on the result of a voicing and energy detector (these were deactivated in the present study because the /iau/ tokens had already been selected), and on constraints on the proximity of formants: F1 must be at least 150 Hz greater than the fundamental frequency, and F2, F3, and F4 must be more than 300, 400, and 500 Hz greater than F1, F2, and F3 respectively. (Reducing these parameter values to 50, 100, 100, and 300 did not result in substantial improvement in the performance of the forensic-voice-comparison system.)

2.4 Forensic-voice-comparison systems

2.4.1 MFCC baseline system

The baseline forensic-voice-comparison system extracted 16 mel-frequency-cepstral-coefficients (MFCCs) 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 and included in the subsequent statistical modeling (Furui, 1986). Feature warping (Pelecanos & Sridharan, 2001) was applied to the MFCCs and deltas before subsequent modeling. A Gaussian mixture model - universal background model (GMM-UBM, Reynolds, Quatieri, & Dunn, 2000) was built using the background data to train the background model. After tests on the development set using different numbers of Gaussians, the number of Gaussians used for testing was set to 1024.

2.4.2 Formant-trajectory systems

Discrete cosine transforms (DCTs) were fitted to the measured formant trajectories of all the /iau/ tokens – this method of information extraction for forensic voice comparison has previously been applied in a number of studies including Morrison (2009a, 2011a, 2012b) and Zhang, Morrison, and Thiruvaran (2011). On the basis of tests made on the development set in Zhang, Morrison, and Thiruvaran (2011), the zeroth through fourth DCT coefficient values from F2 and F3 were used as variables in the present study. Likelihood ratios were calculated using the multivariate kernel density (MVKD) formula (Aiken & Lucy, 2004a, 2004b) implemented in Morrison (2007).

A separate system was built for each set of measurements from each human-supervisor (first, second, and third sets in the case of high-quality recordings), and for each automatic formant tracker.

2.4.3 MFCC on **/iau/** *system*

A second MFCC system was constructed which was identical to the first except that MFCCs and deltas were only calculated for the portions of the recordings which fell within the /iau/ markers, and (because of the smaller amount of data) only 32 Gaussians were included in the mixture. This system uses the same portions of the recordings as the formant-trajectory systems and is thus a diagnostic as to whether it is the selection of the /iau/ tokens which is important or whether the formant-trajectory procedures themselves also contribute to system performance.

2.4.4 Use of background, development, and test sets

In both the development and test sets, every speaker's Session 2 recording (nominal offender recording) was compared with their own Session 1 recording (nominal suspect recording) for a same-speaker comparison and with every other speaker's Session 1 recording (nominal suspect recordings) as different-speaker comparisons. In the GMM-UBM systems the nominal suspect recordings were used to build models and the nominal offender recordings were used as probes (for the MVKD systems the use of the pair of test recordings is symmetrical). In the channel-mismatch conditions, the nominal offender recordings were telephone-channel degraded recordings, and the nominal suspect recordings and the background were high-quality recordings. Both Session 1 and Session 2 recordings were included in the background.

The development set was used to calculate scores which were then used to calculate weights for logistic-regression calibration (Brümmer & du Preez, 2006; van Leeuwen & Brümmer, 2007; Morrison, 2012a) which was applied to convert the scores from the test set to likelihood ratios (calculations were performed using Brümmer, 2005, and Morrison, 2009b). Logistic regression was also used to fuse the scores from the baseline system with scores from other systems and convert them to likelihood ratios (Pigeon, Druyts, & Verlinde, 2000; Morrison, 2012a).

3 RESULTS AND DISCUSSION

3.1 Reliability of human-supervised formant measurement – high-quality recordings

The within-supervisor maximum-likelihood standard deviation, σ_v, across the three measurement repetitions across both recording sessions was calculated for each of the four human supervisors. The standard deviation was calculated both in hertz (Eq. 2a,b,d) and as a proportion relative to the mean of each formant value across the three measurement repetitions (Eq. 2a,c,d).

$$\sigma_v = \sqrt{\frac{1}{S}\sum_{s}^{S}\frac{1}{K_s}\sum_{k}^{K_s}\frac{1}{T_k}\sum_{t}^{T_k}\frac{1}{M}\sum_{m}^{M}\frac{1}{R}\sum_{r}^{R} y_{v,s,k,t,m,r}^2} \tag{2a}$$

$$y_{v,s,k,t,m,r}^2 = \left(x_{v,s,k,t,m,r} - \bar{x}_{v,s,k,t,m}\right)^2 \tag{2b}$$

$$y_{v,s,k,t,m,r}^2 = \left(\frac{x_{v,s,t,k,m,r} - \bar{x}_{v,s,t,k,m}}{\bar{x}_{v,s,t,k,m}}\right)^2 \tag{2c}$$

$$\bar{x}_{v,s,k,t,m} = \frac{1}{R}\sum_{r}^{R} x_{v,s,k,t,m,r} \tag{2d}$$

Where $x_{v,s,t,m,r}$ is the rth formant measurement made by supervisor v of formant m at time t in vowel token k produced by speaker s. $\bar{x}_{v,s,t,k,m}$ is the mean value over the r formant measurements made by supervisor v of formant m at time t in vowel token k produced by speaker s (calculations performed on hertz values). Each supervisor v made $R = 3$ measurement (repetitions) of each of $M = 3$ formants (F1, F2, F3) of $S = 60$ speakers' vowel tokens. The number of measurements points T_k across time for token k depended on the idiosyncratic duration of the token. The number of tokens K_s for speaker s was also idiosyncratic (tokens were pooled across both recording sessions).

The results are given in Table 1. The within-supervisor standard deviations of 45–55 Hz (2.4–2.9%) can probably be considered an acceptable range for reliability.

TABLE 1. Within-supervisor standard deviations for formant measurements over the three formants and the three replicated measurements of each of the four human supervisors, σ_v, and the between-supervisor standard deviation, σ_b (high-quality recordings). Results reported in hertz and as the percentage of the mean of the values measured across the three replications.

	σ	
supervisor	Hz	%
CZ	45	2.4
EE	51	2.5
FO	55	2.9
GSM	49	2.3
between	68	3.5

The between-supervisor maximum-likelihood standard deviation, σ_b, was calculated both in hertz (Eq. 3a,b,d) and as a proportion relative to the mean of each formant value across the three measurement repetitions made by each supervisor (Eq. 3a,c,d).

$$\sigma_b = \sqrt{\frac{1}{S}\sum_{s}^{S}\frac{1}{K_s}\sum_{k}^{K_s}\frac{1}{T_k}\sum_{t}^{T_k}\frac{1}{M}\sum_{m}^{M}\frac{1}{V}\sum_{v}^{V} y_{v,s,k,t,m}^2} \tag{3a}$$

$$y_{v,s,k,t,m}^2 = \left(\bar{x}_{v,s,k,t,m} - \bar{x}_{s,k,t,m}\right)^2 \tag{3b}$$

$$y_{v,s,k,t,m}^2 = \left(\frac{\bar{x}_{v,s,k,t,m} - \bar{x}_{s,k,t,m}}{\bar{x}_{s,k,t,m}}\right)^2 \tag{3c}$$

$$\bar{x}_{s,k,t,m} = \frac{1}{V}\sum_{v}^{V}\bar{x}_{v,s,k,t,m} \tag{3d}$$

Where $\bar{x}_{s,t,k,m}$ is the mean value over the R = 3 formant measurements (repetitions) and the V = 4 supervisors of formant m at time t in vowel token k produced by speaker s (calculations performed on hertz values).

The results are given in Table 1. The between-supervisor reliability was about 36% poorer than the mean within-supervisor reliability.

The human supervisors' perception was that F3 was harder to measure than F1 and F2, and that some speakers were harder to measure than others. Eq. 4 and 5 were used to calculate the within- and between-supervisor standard deviations ($\sigma_{v,m}$ and $\sigma_{b,m}$ respectively) for each formant across all speakers.

$$\sigma_{v,m} = \sqrt{\frac{1}{S}\sum_{s}^{S}\frac{1}{K_s}\sum_{k}^{K_s}\frac{1}{T_k}\sum_{t}^{T_k}\frac{1}{R}\sum_{r}^{R} y_{v,s,k,t,m,r}^2} \tag{4}$$

$$\sigma_{b,m} = \sqrt{\frac{1}{S}\sum_{s}^{S}\frac{1}{K_s}\sum_{k}^{K_s}\frac{1}{T_k}\sum_{t}^{T_k}\frac{1}{V}\sum_{v}^{V} y_{v,s,k,t,m}^2} \tag{5}$$

The results are given in Table 2. Although in hertz measurements the reliability of F3 measurements was worse than that of F2, which in turn was worse than that of F1, in proportional measurements, F3 measurements were actually more reliable than F1 and F2 measurements which had about the same degree of reliability as each other.

TABLE 2. Within-supervisor standard deviations for formant measurements per formant over the three replicated measurements of each of the four human supervisors, and the between-supervisor standard deviation (high-quality recordings). Results reported in hertz and as the percentage of the mean of the values measured across the three replications.

	σ					
	Hz			%		
superviso	F1	F2	F3	F1	F2	F3
CZ	19	48	58	2.8	2.7	1.5
EE	19	55	67	2.6	3.0	1.9
FO	22	50	80	3.2	3.1	2.1
GSM	16	47	68	2.3	2.6	1.7
between	25	68	93	3.8	3.9	2.6

Eq. 6 was used to calculate the within-supervisor standard deviations for each formant for each speaker, and the across-supervisor means of these standard-deviation values are given in Fig. 1. The formants of speakers 65 and 66 were particularly difficult to measure, and the supervisors frequently had to resort to hand tracking. If such difficulty were found with known- or questioned-voice recordings in casework, then one would probably decide not to use formant-trajectory measurements as a component of the system.

$$\sigma_{v,s,m} = \sqrt{\frac{1}{K_s}\sum_{k}^{K_s}\frac{1}{T_k}\sum_{t}^{T_k}\frac{1}{R}\sum_{r}^{R} y_{v,s,k,t,m,r}^2} \tag{6}$$

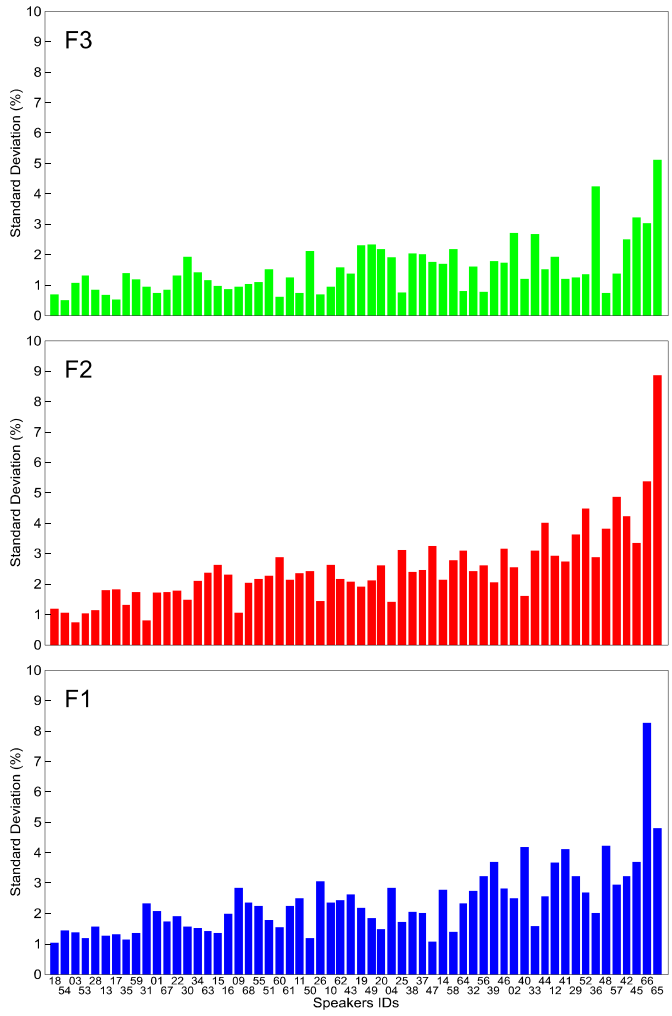

FIG. 1. Across-supervisor mean of the within-supervisor standard deviations for each formant for each speaker (high-quality recordings). Results reported as the percentage of the mean of the values measured across the three replications. Speakers are ranked according to the across-supervisor across-formant mean of their within-supervisor standard deviations.

3.2 Validity (and reliability) of the forensic-voice-comparison systems

3.2.1 High-quality v high-quality results

3.2.1.1 Validity

For each supervisor each of their three sets of formant-trajectory measurements, and for each automatic tracker their single set of formant-trajectory measurements, were used to build and test a forensic-voice-comparison system, and the C_{llr} of the post-calibrated test-set results were calculated as measures of the performance of the system (see Brümmer & du Preez, 2006; van Leeuwen & Brümmer, 2007). Each of the systems was also fused with the MFCC system and the C_{llr} of the fused systems

calculated. C_{llr} can be considered a measure of the validity of a forensic-comparison system (Morrison, 2011b), lower C_{llr} values indicate better validity. The results for the high-quality v high-quality recordings are shown in Fig. 2.

The baseline MFCC system had a C_{llr} value of 0.026. All of the systems which were fusions of the human-supervised formant-trajectory systems with the baseline MFCC system resulted in substantial improvements in performance over the baseline system alone. C_{llr} values were in the range 0.003 to 0.016, a 38% to 88% reduction relative to the baseline system. Of the systems which were fusions of the automatic formant-trajectory systems with the baseline system, only WAVESURFER gave an improvement which was within the range of the human-supervised systems, C_{llr} of 0.012, a 54% reduction relative to the baseline system.

Although a practice unlikely to be adopted for casework given the huge investment in human labor, a potential procedure could be to measure the formants of all vowels three times and then use the central value measured for each formant. The following procedure was adopted for the present study: For each formant of each vowel the mean vector was calculated for the three sets of DCT coefficient values from the three measurement repetitions. The squared Euclidian distance from the mean vector to each of the three sets of DCT coefficient values was then calculated. The set of DCT coefficient values closest to the mean vector was then used as input to the MVKD formula. This resulted in C_{llr} values in the range 0.005 to 0.007, a 73% to 81% reduction relative to the baseline system (see Fig. 2). If human labor were not an issue, on the basis of these results this would be the preferred procedure.

Fusion of the MFCC-on-/iau/ system with the baseline system gave a C_{llr} of 0.012, a 54% reduction relative to the baseline system. This performance was approximately the same as for the WAVESURFER system. Thus it appears that for fully-automatic systems it is primarily the selection of the /iau/ tokens which is the source of the improvement rather than the formant-trajectory procedures, bearing in mind that the quality of automatic formant tracking is unlikely to be as good as human-supervised formant tracking. In terms of C_{llr}, all but one of the human-supervised systems performed better than the MFCC-on-/iau/ and WAVESURFER systems. This suggests that for the human-supervised systems improvements in performance were not just due to the selection of the /iau/ tokens, but likely due to the use of the formant-trajectory procedures including good-quality formant tracking.

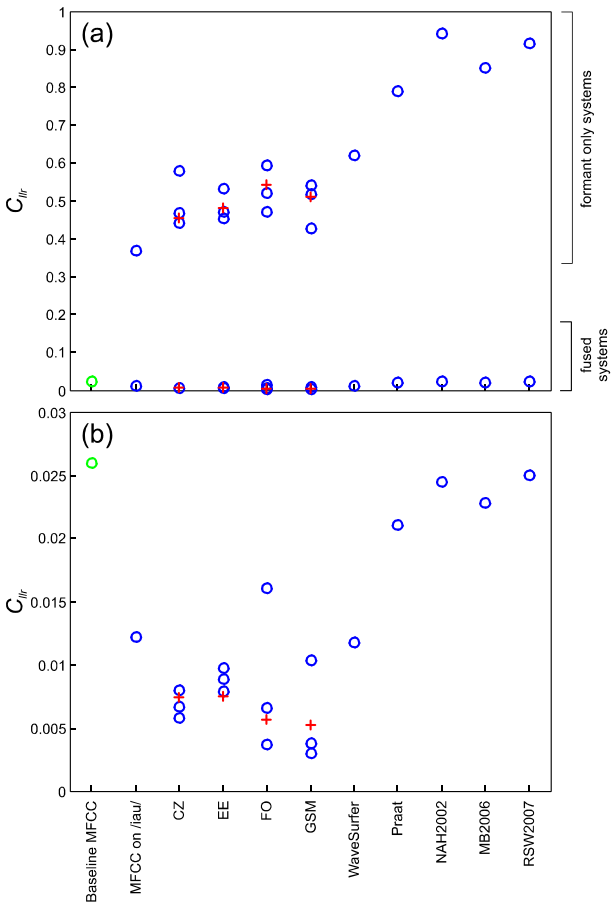

FIG. 2. C_{llr} calculated for the test set on each of the forensic-voice-comparison systems based on formant-trajectory measurements alone, for the MFCC-on-/iau/ system, for the baseline MFCC system, and for each of the former systems fused with the baseline system (high-quality v high-quality recordings). For human-supervised systems, circles represent C_{llr} based on a single set of formant-trajectory measurements, and crosses represent C_{llr} values based on each supervisor's central set of DCT coefficient values from each formant from each vowel token. (a) Results for formant-only systems and for fused systems. (b) Results for fused systems on a magnified scale.

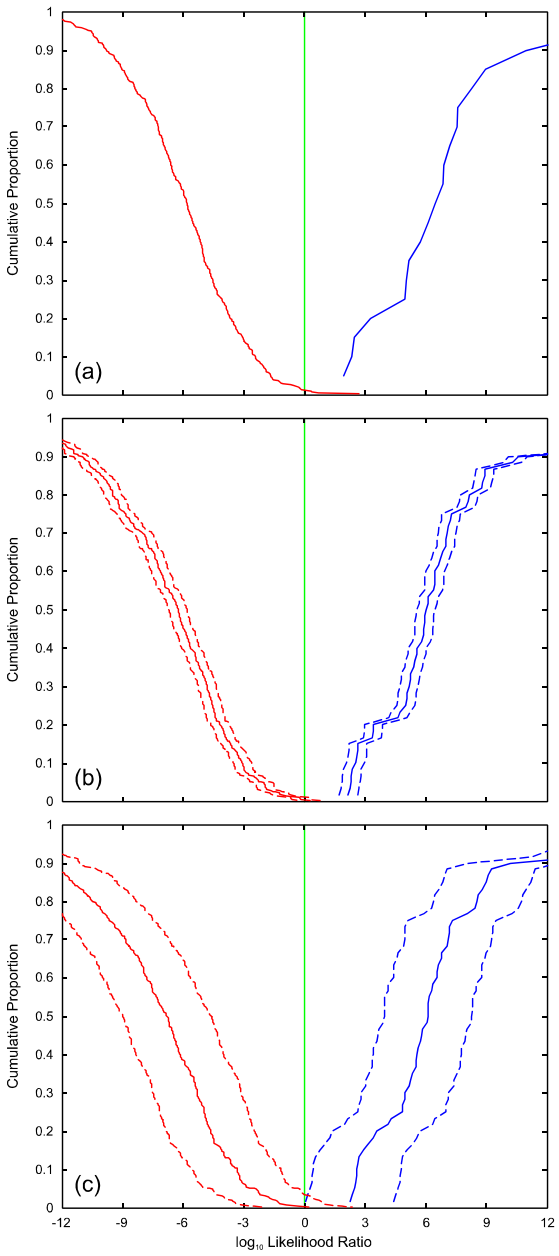

FIG. 3. Tippett plots showing system performance for (a) the baseline-MFCC system, and the fusion of the baseline system with human-supervised formant-trajectory systems: (b) supervisor CZ, (c) supervisor GSM (high-quality v high-quality recordings).

3.2.1.2 Reliability

This section describes the assessment of the reliability (precision) of the performance of human-supervised systems given the reliability of the human-supervised formant measurements. In all of the human-supervised systems considered so far there was a single likelihood-ratio estimate for each same-speaker and each different-speaker comparison. In order to differentiate validity and reliability (accuracy and precision) of system performance the three likelihood-ratio estimates for each same-

speaker and different-speaker comparison resulting from the three sets of formant-trajectory measurements are exploited (the three sets of formant measurements were those already calculated using the procedures described above). As a measure of validity the C_{llr} value from the means of the three likelihood-ratio estimates for each comparison was calculated, and as a measure of reliability the 95% credible interval (95% CI) was calculated using the parametric procedure described in Morrison (2011b, see also Morrison, Thiruvaran, & Epps, 2010). The results are given in Table 3. For the formant-only systems for all supervisors the 95% CI was 0.8 to 1.6 orders of magnitude. When fused with the baseline MFCC system the 95% CIs ranged from 0.45 to 2.35 orders of magnitude. Fig. 3 provides Tippett plots of the performance of the baseline system and of the baseline system fused with two of the human-supervised systems, that of CZ which had the best reliability and average validity, and that of GSM which had the best validity but one of the poorest reliabilities (for an introduction to Tippett plots see Morrison, 2010a §99.330, or Morrison, 2011a Appendix A). The 95% credible intervals are indicated on the Tippett plots as the dashed lines to the left and the right of the solid lines which represent the group-mean values.

TABLE 3. Validity and reliability (accuracy and precision) measures (C_{llr} on group means, and 95% credible interval expressed in log base ten, respectively) for forensic voice comparison systems based on human-supervised formant trajectory measurement (high-quality recordings). Imprecision is due to imprecision in formant measurement.

	formant only systems		fused systems	
supervisor	C_{llr}	95% CI	C_{llr}	95% CI
CZ	0.490	0.82	0.007	0.45
EE	0.477	0.84	0.009	1.08
FO	0.513	1.57	0.007	2.35
GSM	0.491	1.60	0.004	2.18

The results indicate very good performance for the baseline system, and complete separation when the baseline system was fused with any of the human-supervised formant trajectory systems. Some caution should be exercised in generalizing these results because speakers in the database were not selected to be particularly similar sounding and with high-quality recordings the task may be too easy and not representative of casework conditions. The results also indicate that there can be large differences in system reliability depending on which human supervisor makes the formant measurements, and the reliability of any system used for casework should therefore be assessed including the particular human-supervisor as a component of the system. Results reported in Duckworth et al. (2011) suggest that between-supervisor variability in formant measurement can be reduced with training; however, in the present study there is no clear pattern relating the reliability of individual human supervisors' formant measurements and the validity and reliability of the individual forensic-voice-comparison systems based on those measurements (a pattern may exist but be difficult to discern given only four supervisors).

A Tippett plot for the WAVESURFER systems (not shown) gave results which appeared to be similar to those obtained for human-supervised systems, but a Tippett plot for the MFCC-on-/iau/ system (not shown) did not have the level of improvement seen in human-supervised systems in terms of a reduction in the magnitude of log likelihood ratios from different-speaker comparisons which contrary-to-fact gave greater support to the same-speaker hypothesis than the different-speaker hypothesis. It therefore appears that for this high-quality v high-quality condition WAVESURFER could be substituted as a cheaper alternative to human-supervised formant-trajectory measurement without too deleterious an effect of system validity. The fully-automatic WAVESURFER system would presumably give the same formant measurements every time and hence variability in formant measurement would not contribute to imprecision in the output of the forensic-voice-comparison system.

3.2.2 Landline-to-landline v landline-to-landline results

Only one human supervisor (CZ) measured the telephone-channel degraded recordings, and she measured them only once, therefore only validity measures are reported for these conditions. Also, for simplicity, only results for fused systems are reported.

This section provides results of landline-to-landline v landline-to-landline comparisons. C_{llr} values are shown in Fig. 4. The baseline-MFCC system had a C_{llr} of 0.073, substantially worse than for the high-quality v high-quality condition. The system which was a fusion of the human-supervised formant-trajectory system with the baseline-MFCC system had a substantial improvement in performance over the baseline system, the C_{llr} value was 0.047, a 36% reduction relative to the baseline system. Of the fused systems including automatic formant trackers, the PRAAT and M&B2006 systems had similar improvements over the baseline system, C_{llr} of 0.046 and 0.050 respectively, 37% and 31% reduction relative to the baseline system (but see discussion of Tippett plots below). The other fused systems including automatic formant trackers did not perform as well, and the fusion including the RSW2007 tracker actually had worse performance than the baseline system.

Fusion of the MFCC-on-/iau/ system with the baseline system gave a C_{llr} of 0.038, a 48% reduction relative to the baseline system. This was better performance than any of the formant-trajectory systems including the human-supervised formant-trajectory system. The results suggest that under these conditions it may be /iau/ selection itself which is the primary cause of performance improvement and that the extra cost of formant tracking is not warranted; however, C_{llr} provides a single value summary of system performance and is a many-to-one mapping, and examination of the Tippett plots in Fig. 5 suggests a different interpretation of the results: It appears that the performance improvement for the MFCC-on-/iau/ system was due to a greater extent to large positive log likelihood ratios from same-speaker comparisons getting even larger, whereas for the human-supervised formant-trajectory system it was due to a greater extent to small positive likelihood ratios from same-speaker comparisons getting larger and positive likelihood ratios from different-speaker comparisons getting

smaller. Arguably, already good results from same-speaker comparisons getting even better is less important than weak results from same-speaker comparisons getting better and misleading results from different-speaker comparisons (those which contrary-to-fact provided greater support to the same-speaker hypothesis) getting better.

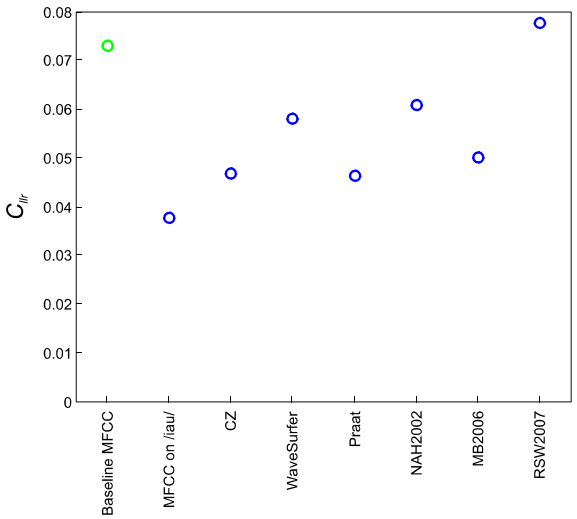

FIG. 4. C_{llr} calculated for the baseline MFCC system, and for the MFCC-on-/iau/ system and for each of the formant-trajectory systems fused with the baseline system (landline-to-landline v landline-to-landline recordings). Note: The *y*-axis scale is magnified and not necessarily the same as on other figures.

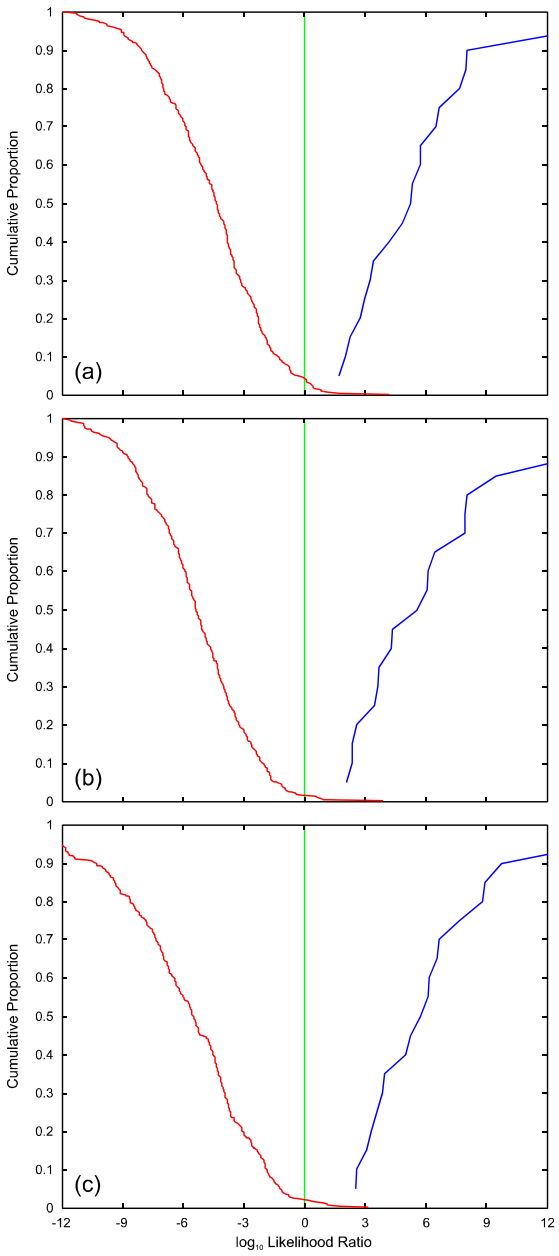

FIG. 5. Tippett plots showing system performance for (a) the baseline-MFCC system and the fusion of the baseline system with (b) the MFCC-on-/iau/ system and (c) the human-supervised formant-trajectory system (landline-to-landline v landline-to-landline recordings).

Tippett plots for the PRAAT and M&B2006 systems (not shown) were more similar to the Tippett plot for the MFCC-on-/iau/ system than to the Tippett plot for the human-supervised system; hence, despite the similarity in improvement in C_{llr}, the human-supervised formant-trajectory system can be said to have also outperformed the fully-automatic formant-trajectory systems.

3.2.3 High-quality v landline-to-landline results

This section provides results of high-quality v landline-to-landline comparisons. C_{llr} values are shown in Fig. 6. The baseline-MFCC system had a C_{llr} of 0.047, intermediate between those of the high-quality v high-quality and landline-to-landline v landline-to-landline conditions. The system which was a fusion of the human-supervised formant-trajectory system with the baseline-MFCC system had a substantial improvement in performance over the baseline system, the C_{llr} value was 0.029, a 39% reduction relative to the baseline system. The Tippett plots of the baseline and human-supervised systems in Fig. 7 indicate improvement for likelihood ratios from different-speaker comparisons which contrary-to-fact gave greater support to the same-speaker hypothesis. Of the fused systems including automatic formant trackers, only the NAH2002 systems gave an improvement over the baseline system, C_{llr} of 0.037, a 22% reduction relative to the baseline system. In a Tippett plot of the NAH2002 system (not shown) performance appeared to be intermediate between the baseline and human-supervised systems. The other fused systems including automatic formant trackers had worse performance than the baseline system. Fusion of the MFCC-on-/iau/ system with the baseline system gave a C_{llr} of 0.047, a 1% reduction relative to the baseline system. Under these channel-mismatch conditions the human-supervised formant-trajectory system clearly outperformed all other systems, and one can therefore conclude this was due to formant-trajectory measurement not just selection of /iau/ tokens.

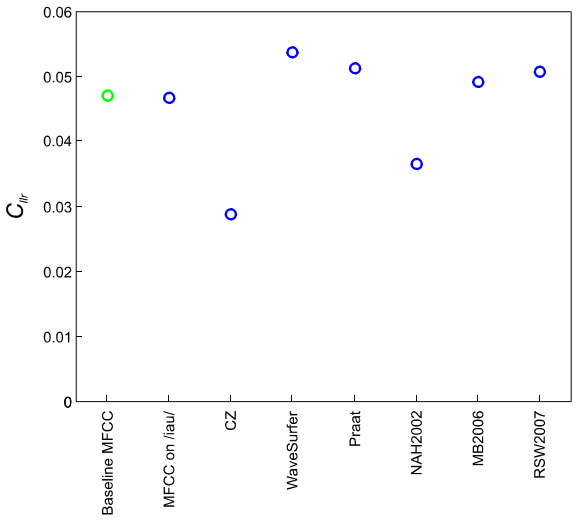

FIG. 6. C_{llr} calculated for the baseline MFCC system, and for the MFCC-on-/iau/ system and for each of the formant-trajectory systems fused with the baseline system (high-quality v landline-to-landline recordings). Note: The *y*-axis scale is magnified and not necessarily the same as on other figures.

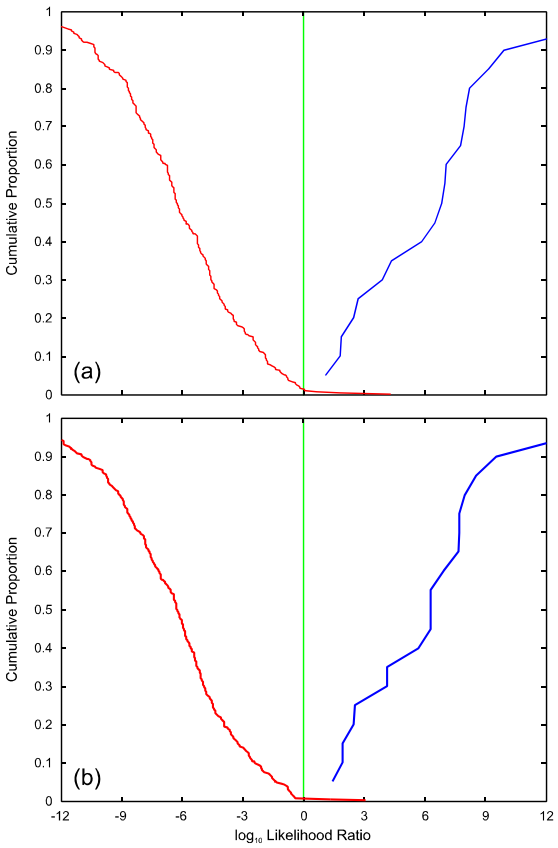

FIG. 7. Tippett plots showing system performance for (a) the baseline-MFCC system and (b) the fusion of the baseline system with the human-supervised formant-trajectory system (high-quality v landline-to-landline recordings).

3.2.4 Mobile-to-mobile v mobile-to-mobile results

This section provides results of mobile-to-mobile v mobile-to-mobile comparisons. C_{llr} values are shown in Fig. 8. The baseline-MFCC system had a C_{llr} of 0.111, the worst performance of any baseline system reported so far. The system which was a fusion of the human-supervised formant-trajectory system with the baseline-MFCC system had a substantial improvement in performance over the baseline system, the C_{llr} value was 0.083, a 25% reduction relative to the baseline system. Of the fused systems including automatic formant trackers, only the WAVESURFER system gave a substantial improvement over the baseline system, although this was only half as good as the human-supervised system: C_{llr} of 0.097, a 12% reduction relative to the baseline system. The other fused systems including automatic formant trackers gave less than 5% improvement over the baseline system.

Fusion of the MFCC-on-/iau/ system with the baseline system gave a C_{llr} of 0.077, a 30% reduction relative to the baseline system. This was better performance than any of the formant-trajectory systems including the human-supervised formant-trajectory system. The Tippett plots in Fig. 9 indicate that in this instance the human-supervised formant-trajectory system did not lead to the sort of improvement noted for the results of the landline-to-landline v landline-to-landline comparisons

(section 3.2.2). Under the mobile-to-mobile v mobile-to-mobile condition, it therefore appears that the cost of measuring formant-trajectories is not warranted.

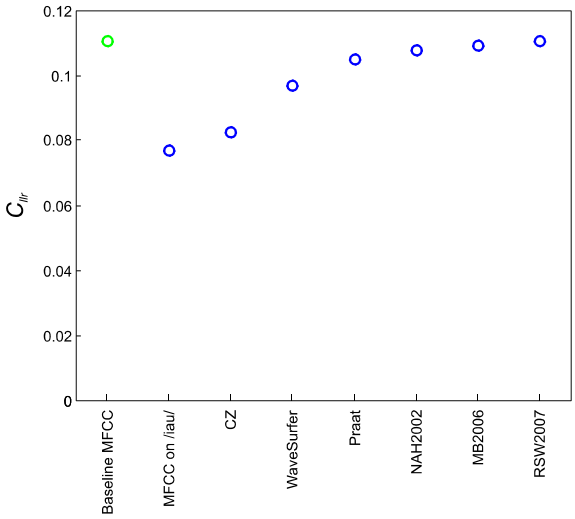

FIG. 8. C_{llr} calculated for the baseline MFCC system, and for the MFCC-on-/iau/ system and for each of the formant-trajectory systems fused with the baseline system (mobile-to-mobile v mobile-to-mobile recordings). Note: The *y*-axis scale is magnified and not necessarily the same as on other figures.

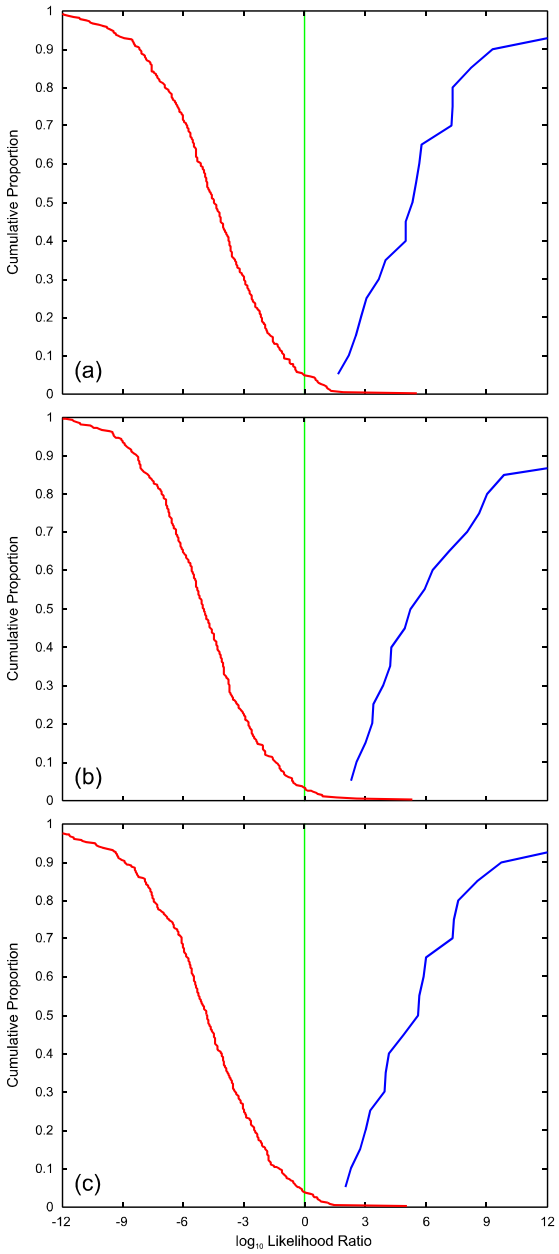

FIG. 9. Tippett plots showing system performance for (a) the baseline-MFCC system and the fusion of the baseline system with (b) the MFCC-on-/iau/ system and (c) the human-supervised formant-trajectory system (mobile-to-mobile v mobile-to-mobile recordings).

3.2.5 High-quality v mobile-to-mobile results

This section provides results of high-quality v mobile-to-mobile comparisons. C_{llr} values are shown in Fig. 10. The baseline-MFCC system had a C_{llr} of 0.121, slightly worse than for the mobile-to-mobile v mobile-to-mobile condition. The system which was a fusion of the human-supervised formant-trajectory system with the baseline-MFCC system had a substantial improvement in performance over the baseline system, the C_{llr} value was 0.085, a 30% reduction relative to the baseline system. Of the fused systems including automatic formant trackers, none resulted in more than a 7% reduction in C_{llr} relative to the baseline system, and PRAAT gave results which were actually 13% higher. Fusion of the MFCC-on-/iau/ system with the baseline system gave a C_{llr} of 0.134, an 11% increase (worse performance) relative to the baseline system. Under these channel-mismatch conditions the C_{llr} results suggest that the human-supervised formant-trajectory system outperformed all other systems; however, examination of the Tippett plots of the baseline and human-supervised systems in Fig. 11 indicate that improvement was primarily due to large magnitude log likelihood ratios supporting consistent-with-fact hypotheses getting even larger, with little improvement for problematic positive log likelihood ratios from different-speaker comparisons which contrary-to-fact gave greater support to the same-speaker hypothesis.

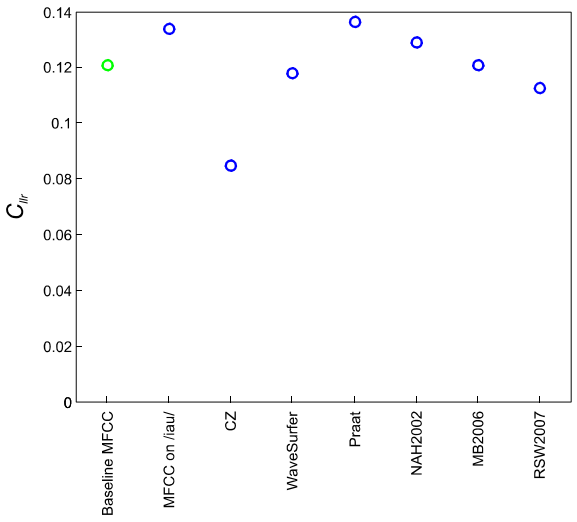

FIG. 10. C_{llr} calculated for the baseline MFCC system, and for the MFCC-on-/iau/ system and for each of the formant-trajectory systems fused with the baseline system (high-quality v mobile-to-mobile recordings). Note: The *y*-axis scale is magnified and not necessarily the same as on other figures.

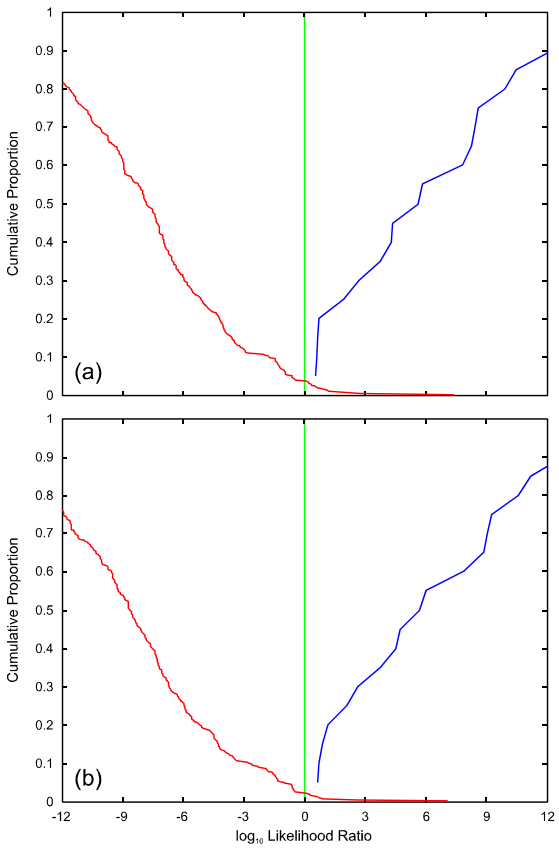

Fig. 11. Tippett plots showing system performance for (a) the baseline-MFCC system and (b) the fusion of the baseline system with a human-supervised formant-trajectory system (high-quality v mobile-to-mobile recordings).

3.2.6 Mobile-to-landline v mobile-to-landline results

This section provides results of mobile-to-landline v mobile-to-landline comparisons. C_{llr} values are shown in Fig. 12. The baseline-MFCC system had a C_{llr} of 0.226, the worst performance of any baseline system reported so far. The system which was a fusion of the human-supervised formant-trajectory system with the baseline-MFCC system had a substantial improvement in performance over the baseline system, the C_{llr} value was 0.107, a 53% reduction relative to the baseline system. Of the fused systems including automatic formant trackers, the WAVESURFER and PRAAT systems had some improvement over the baseline system, C_{llr} of 0.185 and 0.195 respectively, 18% and 13% reduction relative to the baseline system, but this was much less than for the system including human-supervised formant tracking.

Fusion of the MFCC-on-/iau/ system with the baseline system gave a C_{llr} of 0.102, a 55% reduction relative to the baseline system. This was better performance than any of the formant-trajectory systems including the human-supervised formant-trajectory system; however, partially similar to in the landline-to-landline v landline-to-landline condition, examination of the Tippett plots in Fig. 13 indicate that the C_{llr} performance improvement for the MFCC-on-/iau/ system was due to a greater extent to large positive log likelihood ratios from same-speaker comparisons getting even larger, whereas for the human-supervised formant-trajectory system it was due to a greater extent to positive likelihood ratios from different-speaker comparisons getting smaller. Arguably, already good results from same-speaker comparisons getting even better is less important than misleading results from different-speaker comparisons (those which contrary-to-fact provided greater support to the same-speaker hypothesis) getting better.

3.2.7 High-quality v mobile-to-landline results

This section provides results of high-quality v mobile-to-landline comparisons. C_{llr} values are shown in Fig. 14. The baseline-MFCC system had a C_{llr} of 0.320, the worst performance of any baseline system reported. The system which was a fusion of the human-supervised formant-trajectory system with the baseline-MFCC system resulted in a only a small improvement in performance over the baseline system, C_{llr} of 0.287, an 11% reduction relative to the baseline system. Of the systems which were fusions of the automatic formant-trajectory systems with the baseline system, only M&B2006 gave any improvement, C_{llr} of 0.296, a 7% reduction relative to the baseline system. WAVESURFER and NAH2002 resulted in performance which was 25% and 23 % worse than the baseline system respectively.

Fusion of the MFCC-on-/iau/ system with the baseline system gave a C_{llr} of 0.216, a 33% reduction relative to the baseline system, better than any of the formant-trajectory systems including the human-supervised system. The Tippett plots in Fig. 15 indicate that in this instance the human-supervised formant-trajectory system did not lead to the sort of improvement noted for the results of

the landline-to-landline v landline-to-landline comparisons (section 3.2.2), but rather to already large magnitude negative log likelihood ratios from different-speaker comparisons getting even larger. Under the high-quality v mobile-to-landline condition, it therefore appears that the cost of measuring formant-trajectories is not warranted.

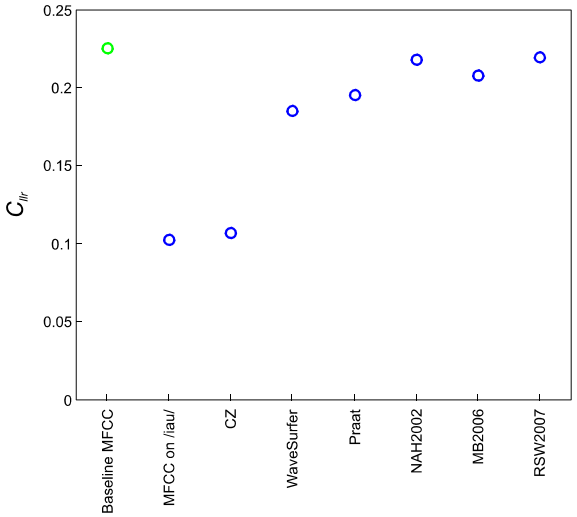

FIG. 12. C_{llr} calculated for the baseline MFCC system, and for the MFCC-on-/iau/ system and for each of the formant-trajectory systems fused with the baseline system (mobile-to-landline v mobile-to-landline recordings). Note: The *y*-axis scale is magnified and not necessarily the same as on other figures.

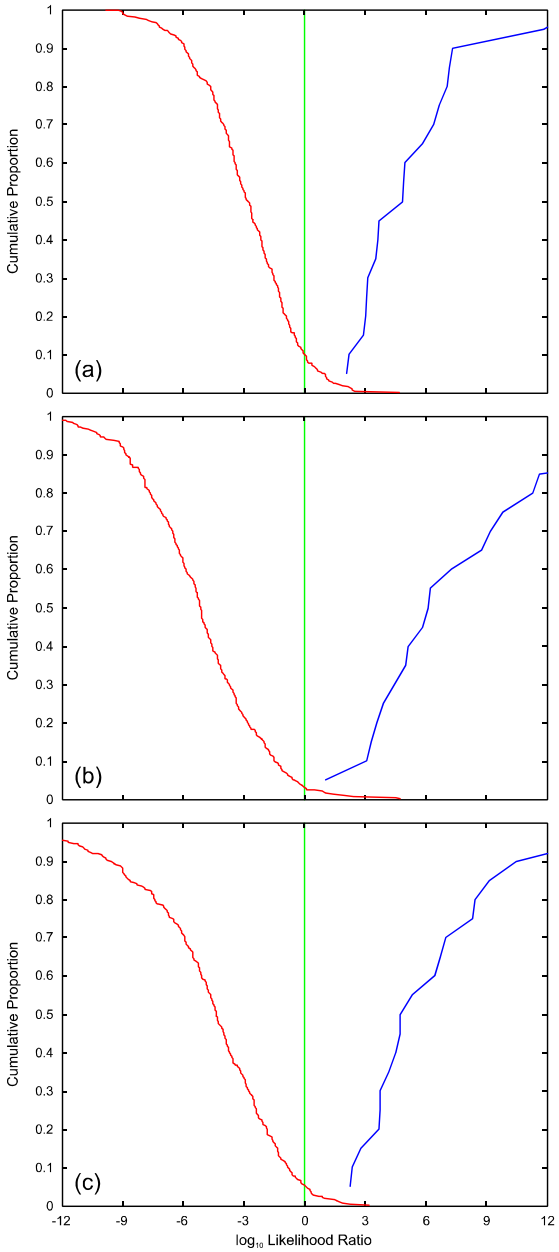

FIG. 13. Tippett plots showing system performance for (a) the baseline-MFCC system and the fusion of the baseline system with (b) the MFCC-on-/iau/ system and (c) the human-supervised formant-trajectory system (mobile-to-landline v mobile-to-landline recordings).

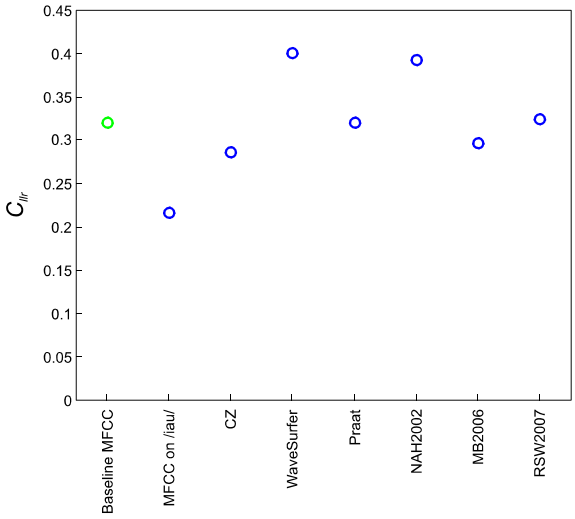

FIG. 14. C_{llr} calculated for the baseline MFCC system, and for the MFCC-on-/iau/ system and for each of the formant-trajectory systems fused with the baseline system (high-quality v mobile-to-landline recordings). Note: The *y*-axis scale is magnified and not necessarily the same as on other figures.

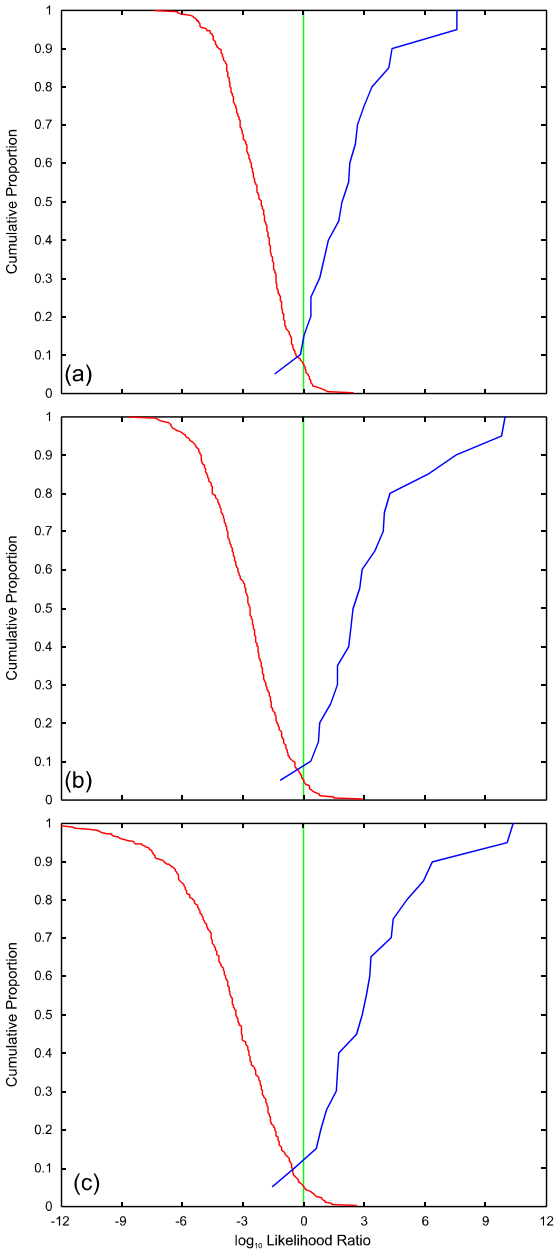

FIG. 15. Tippett plots showing system performance for (a) the baseline-MFCC system and the fusion of the baseline system with (b) the MFCC-on-/iau/ system and (c) the human-supervised formant-trajectory system (high-quality v mobile-to-landline recordings).

4 GENERAL DISCUSSION AND CONCLUSION

Direct assessment of within-supervisor reliability on formant-trajectory measurement resulted in standard deviations on the order of 2.5% of the absolute values of the formants measured, and little spread in the within-supervisor standard-deviation values across different supervisors. When these measurements were used as part of a forensic-voice-comparison system, however, there were large between-supervisor differences in the reliability of system performance, with 95% credible intervals for likelihood ratios ranging from less than half an order of magnitude to more than two orders of magnitude. There was also some variability in between-supervisor validity. The validity and reliability of any system used for casework which incorporates human-supervised formant-trajectory measurement should therefore be assessed not only under conditions reflecting those of the case under investigation, but also including the particular supervisor as a component of the system.

Human-supervised and fully-automatic formant-trajectory measurements were assessed as components of forensic-voice-comparison systems under several telephone-transmission conditions including mismatches with high-quality recordings. Fusion of the human-supervised system with the baseline system always led to improvement over the baseline system. Unless otherwise indicated, all discussion below refers to the performance of each system after fusion with the baseline system. Considering both C_{llr} and Tippett plots, human-supervised systems always clearly outperformed fully-automatic formant-trajectory systems, apart from the WAVESURFER system in the high-quality v high-quality condition. No single fully-automatic system consistently outperformed the others, and in some conditions after fusion with the baseline system some fully-automatic systems performed worse than the baseline system. In the following conditions the following fully-automatic formant trackers could be considered as cheaper alternatives to human-supervised formant tracking, obtaining substantial improvements over the baseline system, although in the latter two cases noticeably poorer performance than that of the human-supervised system:

– high-quality v high-quality: WAVESURFER

– landline-to-landline v landline-to-landline: PRAAT, MB2006

– high-quality v landline-to-landline: NAH2002

It is not apparent why one fully-automatic tracker should work better in one condition and another in a different condition. It should be noted that any condition involving mobile-telephone recordings was particularly problematic for fully-automatic formant trackers, and these also gave poorer results for the baseline system and for human-supervised systems.

Overall the different conditions could be ranked in the following order in terms of best to worst validity (according to the C_{llr} from the best-performing system on each condition):

– high-quality v high-quality

– high-quality v landline-to-landline

– landline-to-landline v landline-to-landline

– mobile-to-mobile v mobile-to-mobile

– high-quality v mobile-to-mobile

– mobile-to-landline v mobile-to-landline

– high-quality v mobile-to-landline

However, on examination of Tippett plots as well as C_{llr}, the amount of improvement due to inclusion of human-supervised formant-trajectory measurements for the two mismatch conditions including mobile telephones (high-quality v mobile-to-mobile and high-quality v mobile-to-landline) could be considered of marginal value.

To assess whether improvements over the baseline system were due to formant tracking, or only due to the selection of /iau/ tokens in and of itself, an MFCC-on-/iau/ system was also fused with the baseline system. In the high-quality v landline-to-landline and high-quality v mobile-to-mobile conditions this lead to negligible improvement and worse performance respectively compared to the baseline system. In the following conditions, however, it lead to better performance, in terms of C_{llr}, than the human-supervised formant-trajectory system:

– landline-to-landline v landline-to-landline

– mobile-to-mobile v mobile-to-mobile

– mobile-to-landline v mobile-to-landline

– high-quality v mobile-to-landline

Examination of Tippett plots, however, indicated that in the two same-channel conditions involving landline telephones (landline-to-landline v landline-to-landline and mobile-to-landline v mobile-to-landline) the sort of improvement resulting from the human-supervised formant-trajectory system (small magnitude positive log likelihood values from same-speaker comparisons getting larger, and positive log likelihood ratios from different-speaker comparisons getting smaller) was arguably more important than that due to the MFCC-on-/iau/ system (already large magnitude likelihood ratios giving more support to consistent-with-fact hypotheses getting even larger).

In terms of improvement in system performance human-supervised formant-trajectory measurement would therefore appear to be justified in the following conditions:

– high-quality v high-quality

– landline-to-landline v landline-to-landline

– high-quality v landline-to-landline

– mobile-to-landline v mobile-to-landline

and not in the following conditions:

– high-quality v mobile-to-mobile

– mobile-to-mobile v mobile-to-mobile

– high-quality v mobile-to-landline

The latter can be summarized as “mobile only or mismatches involving mobile”. Note also that for high-quality v high-quality, WAVESURFER could be an acceptable cheaper alternative.

One should, however, be cautious about generalizing these results to other phonemes, to other languages, and to male speakers, and always test the degree of validity and reliability of any system applied to casework under conditions reflecting those of the case at trial.

It should be remembered that apart from feature warping on MFCCs, no attempt was made in the present study to apply statistical modeling techniques to attempt to compensate for channel mismatches. A potential area of future research could be to developing channel compensation techniques for formant trajectories and for the relatively small amounts of suitable data available for forensic-voice-comparison casework compared to the amount typically available for automatic-speaker-recognition research and applications.

Finally, the question remains as to whether the degrees of improvement in system performance obtained by using human-supervised formant-trajectory measurement are justified given the cost in skilled human labor. Not including the initial marking of the /iau/ start and end boundaries, the average time for CZ to make a set of measurements on one session of one speaker was around 15 minutes, and a full set of measurements on two sessions of recordings from 60 speakers would take approximately 30 hours.

REFERENCES

Aitken, C. G. G., and Lucy, D. (2004a). “Evaluation of trace evidence in the form of multivariate data,” App. Stat. 53, 109–122. doi:10.1046/j.0035-9254.2003.05271.x

Aitken, C. G. G., and Lucy, D. (2004b). “Corrigendum: Evaluation of trace evidence in the form of multivariate data,” Appl. Stat. 53, 665–666. doi:10.1111/j.1467-9876.2004.02031.x

Anderson, N. (1978). “On the calculation of filter coefficients for maximum entropy spectral analysis,” *Modern Spectrum Analysis* edited by D. G. Childers (IEEE Press), pp. 252–255.

Assmann, P. F., and Nearey, T. M. (1987). “Perception of front vowels: The role of harmonics in the first formant region,” J. Acoust. Soc. Am. 81, 520–534. doi:10.1121/1.394918

Becker, T., Jessen, M., and Grigoras, C. (2008). "Forensic speaker verification using formant features and Gaussian mixture models," *Proceedings of Interspeech 2008, Brisbane, Australia* (International Speech Communication Association), pp. 1505–1508.

Becker, T., Jessen, M., and Grigoras, C. (2009). "Speaker verification based on formants using Gaussian mixture models," *Proceedings of NAG/DAGA International Conference on Acoustics, Rotterdam, The Netherlands* (German Acoustical Society DEGA, Berlin), pp. 1640–1643.

Boersma, P. (1993). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proceedings of the Institute of Phonetic Sciences 17* (Institute of Phonetic Sciences, Amsterdam), pp. 97–110.

Boersma, P., and Weenink, D. (2011). *Praat: doing phonetics by computer* (Version 5.2.26). http://praat.org/

Brümmer, N. (2005). *Tools for fusion and calibration of automatic speaker detection systems*. http://niko.brummer.googlepages.com/focal/

Brümmer, N. and du Preez, J. (2006). "Application independent evaluation of speaker detection," Comput. Speech Lang. 20, 230–275. doi:10.1016/j.csl.2005.08.001

Byrne, C., and Foulkes, P. (2004). "The 'mobile phone effect' on vowel formants," Int. J. of Speech, Lang. and the Law 11, 83–102.

Chen, N. F., Shen, W., Campbell, J. & Schwartz, R. (2009). "Large-Scale Analysis of Formant Frequency Estimation Variability in Conversational Telephone Speech," *Proceedings of Interspeech 2009, Brighton, UK*, (International Speech Communication Association), pp. 2203–2206.

de Castro, A., Ramos, D., & González-Rodríguez, J. (2009). "Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking," *Proceedings of Interspeech 2009, Brighton, UK* (International Speech Communication Association), pp. 2343–2346.

Deng, L., Acero, A., and Bazzi, I. (2006). "Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint," IEEE Trans. Audio Speech Lang. Process. 14, 425–434. doi:10.1109/TSA.2005.855841

Deng, L., Lee, L.J., Attias, H., and Acero, A. (2007). "Adaptive Kalman Filtering and Smoothing for Tracking Vocal Tract Resonances Using a Continuous-Valued Hidden Dynamic Model," IEEE Trans. Audio Speech Lang. 15, 13–23. doi:10.1109/TASL.2006.876724

Duckworth, M., McDougall, K., de Jong, G., and Shockey, L. (2011). "Improving the consistency of formant measurement," Int. J. of Speech, Lang. and the Law 18, 35–51. doi:10.1558/ijsll.v18i1.35

Enzinger, E. (2011). "Auswirkungen von Sprachcodecs auf Formantmessungen für Sprechervergleiche," *Proceedings of the 37th annual conference of the German Acoustical Society, DAGA, Düsseldorf*, pp. 877–878.

Furui, S. (1986). "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust., Speech and Sig. Proc. 34, 52–59. doi:10.1109/TASSP.1986.1164788

Gold, E., and French, J. P. (2011). "An international investigation of forensic speaker comparison practices," *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China*, edited by W.-S. Lee & E. Zee (Hong Kong: Organizers of ICPhS XVII at the Department of Chinese, Translation and Linguistics, City University of Hong Kong), pp. 1254–1257.

Guillemin, B. J., and Watson, C. (2008). "Impact of the GSM mobile phone network on the speech signal: Some preliminary findings," Int. J. of Speech, Lang. and the Law 15, 193–218. doi:10.1558/ijsll.v15i2.193

Hansen, E. G., Slyh, R. E., & Anderson, T. R. (2001). "Formant and F0 features for speaker recognition," *Proceedings of 2001: A Speaker Odyssey, The Speaker Recognition Workshop* (International Speech Communication Association).

Harrison, P. (2004). *Variability of formant measurements*. Master's Thesis, University of York, York, England, UK.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. 97, 3099–3111. doi:10.1121/1.411872

Jessen, M., and Becker, T. (2010). "Long-term formant distribution as a forensic-phonetic feature (A)," J. Acoust. Soc. Am. 128, 2378. doi:10.1121/1.3508452. Presentation slides available: http://cancun2010.forensic-voice-comparison.net/

Jiménez Gómez, J. J. (2011). "Estructura formántica y campo de dispersión de las vocales del español en telefonía móvil" ("The formant structure and vowel space of mobile-telephone transmitted Spanish vowel"), Estudios Fónicos / Cuadernos de Trabajo, 1, 39–58.

Kirchhübel, C. (2009). *The effects of Lombard speech on vowel formant measurements*. Master's Thesis, University of York, York, England, UK.

Künzel, H. J. (2001). "Beware of the 'telephone effect': The influence of telephone transmission on the measurement of formant frequencies," Forensic Ling. 8, 80–99.

Künzel, H. J. (2002). "Rejoinder to Francis Nolan's 'The 'telephone effect' on formants: A response'," Forensic Ling. 9, 83–86.

Lawrence, S., Nolan, F., and McDougall. K. (2008). Acoustic and perceptual effects of telephone transmission on vowel quality. Int. J. of Speech, Lang. and the Law 15, 161–192. doi:10.1558/ijsll.v15i2.161

Markel, J. D., and Gray, A. H. (1976). *Linear Prediction of Speech* (Springer-Verlag, Berlin).

Masthoff, H., and Meinerz, C. (2012). "Effect of telephone-line transmission and digital audio format on formant tracking measurements - revisted," *Paper presented at the 21st Annual Conference of the International Association for Forensic Phonetics and Acoustics, IAFPA, Santander*. Abstract: http://www.iafpa2012.com/AbstracsPDF/MEINERZ_CHRISTOPH.pdf

McDougall, K. (2006). "Dynamic features of speech and the characterisation of speakers," Int. J. of Speech, Lang. and the Law 13, 89–126.

Meinerz, C., and Masthoff, H. (2011). "Effect of telephone-line transmission and digital audio format on formant tracking measurements," *Proceedings of the 20th Annual Conference of the International Association for Forensic Phonetics and Acoustics, IAFPA, Vienna*. http://www.kfs.oeaw.ac.at/content/view/543/541/lang,8859-1/

Moos, A. (2008). Forensische Sprechererkennung mit der Messmethode LTF (long-term formant distribution) (Forensic speaker recognition using the method of long-term formant distribution measurements). Master's Thesis, University of the Saarland, Saarbrücken, Germany.

Morrison, G. S. (2007). *multivar_kernel_LR: Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation*. http://geoff-morrison.net/

Morrison, G. S. (2009a). "Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs," J. Acoust. Soc. Am. 125, 2387–2397. doi:10.1121/1.3081384

Morrison, G. S. (2009b). *Robust version of train_llr_fusion.m from Niko Brümmer's FoCal Toolbox* (release 2009-07-02). http://geoff-morrison.net/

Morrison, G. S. (2010a). "Forensic voice comparison," *Expert Evidence*, edited by I. Freckelton and H. Selby (Thomson Reuters, Sydney, Australia), ch. 99.

Morrison, G. S. (2010b). *SoundLabeller: Ergonomically designed software for marking and labelling portions of sound files* (Release 2010-11-18). http://geoff-morrison.net/

Morrison, G. S. (2011a). "A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multvariate kernel density (MVKD) versus Gaussian mixture model – universal background model (GMM-UBM)," Speech Comm. 53, 242–256. doi:10.1016/j.specom.2010.09.005

Morrison, G. S. (2011b). "Measuring the validity and reliability of forensic likelihood-ratio systems," Sci. and Just. 51, 91–98. doi:10.1016/j.scijus.2011.03.002

Morrison, G. S. (2012a in press). Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*.

Morrison, G. S. (2012b in press). "Vowel inherent spectral change in forensic voice comparison," *Vowel inherent spectral change*, edited by G. S. Morrison and P. F. Assmann (Springer-Verlag, Heidelberg, Germany), ch. 11.

Morrison, G. S., and Assmann, P. F. (Eds.) (2012 in press). *Vowel inherent spectral change* (Springer-Verlag, Heidelberg, Germany).

Morrison, G. S., and Nearey, T. M. (2011). *FormantMeasurer: Software for efficient human-supervised measurement of format trajectories* (Release 2011-05-26). http://geoff-morrison.net/

Morrison, G. S., Rose, P., and Zhang, C. (2012). "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice," Aus. J. of Forensic Sci. doi:10.1080/00450618.2011.630412

Morrison, G. S., Thiruvaran, T., and Epps, J. (2010). "Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system," in *Proceedings of Odyssey 2010: The Language and Speaker Recognition Workshop, Brno, Czech Republic*, edited by H. Cernocký and L. Burget (International Speech Communication Association), pp. 63–70.

Mustafa, K., and Bruce, I. C. (2006). "Robust formant tracking for continuous speech with speaker variability," IEEE Trans. Audio, Speech Lang. Process., 14, 435–444. doi:10.1109/TSA.2005.855840

Nearey, T. M., Assmann, P. F., and Hillenbrand J. M. (2002). "Evaluation of a strategy for automatic formant tracking," J. Acoust. Soc. Am. 112, 2323 (A). Presentation slides available from: http://www.ualberta.ca/~tnearey/ResearchLinks.html

Nolan, F. J. (2002). "The 'telephone effect' on formants: a response," Forensic Ling. 9, 74–82.

Nolan, F. J., and Grigoras, C. (2005). "A case for formant analysis in forensic speaker identification," J. of Speech, Lang. and the Law 12, 143–173.

Pelecanos, J., and Sridharan, S. (2001). "Feature warping for robust speaker verification," *Proceedings of the Odyssey Speaker Recognition Workshop* (International Speech Communication Association), pp. 213–218.

Pigeon, S., Druyts, P., and Verlinde. P. (2000). "Applying logistic regression to the fusion of the NIST'99 1-speaker submissions," Digit. Sig. Proc. 10, 237–248. doi:10.1006/dspr.1999.0358

Remez, R. E., Dubowski, K. R., Davids, M. L., Thomas, E. F., Paddu, N. U., Grossman, Y. S., and Moskalenko, M. (2011). "Estimating speech spectra for copy synthesis by linear prediction and by hand," J. Acoust. Soc. Am. 130, 2173–2178. doi:10.1121/1.3631667

Reynolds, D. A., Quatieri, T. F., Dunn, R. B. (2000). "Speaker verification using adapted Gaussian mixture models," Digit. Signal Process. 10, 19–41. doi:10.1006/dspr.1999.0361

Rose, P. (2003). "The technical comparison of forensic voice samples," edited by I. Freckelton and H. Selby (Thomson Lawbook, Sydney, Australia), ch. 99.

Rose, P., and Simmons, A. (1996). "F-pattern variability in disguise and over the telephone comparisons for forensic speaker identification," *Proceedings of the 6th Australian International Conference on Speech Science and Technology*, edited by in P. McCormack and A. Russell (Australian Speech Science and Technology Association), pp. 121–126.

Rudoy, D. (2010). Nonstationary Time Series Modeling with Application to Speech Signal Processing. PhD dissertation, School of Engineering and Applied Sciences, Harvard University, Cambridge, MA.

Rudoy, D., Spendley, D. N., and Wolfe, P. J. (2007). "Conditionally linear Gaussian models for estimating vocal tract resonances," *Proceedings of Interspeech 2007, Antwerp, Belgium* (International Speech Communication Association), pp. 526–529.

Sjölander, K.. (2004). *Snack Sound Toolkit* (Version 2.2.10). http://www.speech.kth.se/snack/

Sjölander, K., and Beskow, J. (2000). "WaveSurfer - an open source speech tool," *Proceedings of the 6th International Conference on Speech and Language Processing*, edited by B. Yuan, T. Huang, and X. Tang, pp. 464–467.

Sjölander, K., and Beskow, J. (2011). *Wavesurfer* (Version 1.8.8). http://www.speech.kth.se/wavesurfer/

Talkin, D. (1987). "Speech formant trajectory estimation using dynamic programming with modulated transition costs," J. Acoust. Soc. Am. 82, S55. doi:10.1121/1.2024869

Trawińska, A., and Kajstura, M. (2004). "The inbuilt recorder of mobile phones - Possibilities of forensic speaker identification," Problems of Forensic Sci. 57, 51–80.

Vallabha, G., and Tuller, B. (2002). "Systematic errors in formant analysis of steady-state vowels," Speech Comm. 38, 141–160. doi:10.1016/S0167-6393(01)00049-8

van Leeuwen, D. A., and Brümmer, N. (2007). "An introduction to application-independent evaluation of speaker recognition systems," *Speaker Classification I: Selected Projects*, edited by C. Müller (Springer, Heidelberg, Germany), pp. 330–353. doi:10.1007/978-3-540-74200-5_19

Xue, S. A. N., Hao, J. G. (2006). "Normative standards for vocal tract dimensions by race as measured by acoustic pharyngometry," J. of Voice 20, 391–400. doi:10.1016/j.jvoice.2005.05.001

Zhang, C., and Morrison, G.S. (2011). *Forensic database of audio recordings of 68 female speakers of Standard Chinese*. http://databases.forensic-voice-comparison.net/

Zhang, C., Morrison, G. S., and Thiruvaran, T. (2011). "Forensic voice comparison using Chinese /iau/," *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China*, edited by W.-S. Lee & E. Zee (Organizers of ICPhS XVII at the Department of Chinese, Translation and Linguistics, City University of Hong Kong), pp. 2280–2283.