Voice source features for forensic voice comparison – an evaluation of the GLOTTEX® software package

Ewald Enzinger[1,2], *Cuiling Zhang*[1,3], *Geoffrey Stewart Morrison*[1]

[1]Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, Australia
[2]Acoustics Research Institute, Austrian Academy of Sciences, Austria
[3]Department of Forensic Science & Technology, China Criminal Police University, Shenyang, China

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

Abstract

GLOTTEX® is a software package which extracts information about voice source properties, including estimates of properties related to physical structures of the vocal folds. It has been proposed that the output of GLOTTEX® can be used as part of a forensic-voice-comparison system. We test this using manually labeled segments from a database of voice recordings of 60 female Chinese speakers. Performance was assessed relative to a baseline MFCC GMM-UBM system. GMM-UBM systems based on features extracted by GLOTTEX® were combined with the baseline system using logistic-regression fusion. System performance was assessed in three channel conditions: high-quality v high-quality, mobile-to-landline v mobile-to-landline, and mobile-to-landline v high-quality. Substantial improvements over the baseline system were not observed.

1. Introduction

Auditory-phonetic (as well as auditory-spectrographic) forensic voice comparison has often included experience-based subjective analysis of voice quality (e.g. [1, 2, 3]). "Voice quality" is normally understood to refer to laryngeal vocal-tract settings and to refer to physiologically constrained as well as voluntarily controllable aspects of speech. Objectively-measurable properties of laryngeal settings such as creaky voice, and pathological conditions, or less extreme idiosyncrasies in laryngeal physiology may be useful features to exploit in data- and statistical-model-based forensic voice comparison. The study reported here tests whether this is the case, via an evaluation of the effectiveness of features extracted by the GLOTTEX® software package [4, 5]. The software was originally developed for medical applications, including as a non-invasive diagnostic tool, e.g., modeling physical properties of the vocal folds, including pathologies, on the basis of the acoustic signal generated by the speaker's vocal tract. Gómez-Vilda et al. [6], however, propose that GLOTTEX® would also be effective in forensic voice comparison.

Voice-source features have previously been applied in several automatic-speaker-recognition studies. Farrús et al. [7] reported an increase in performance of an automatic speaker verification when a system based on jitter and shimmer measurements was fused with a baseline system based on spectral and prosodic features. Plumpe et al. [8] modeled the glottal flow derivative using the Liljencrants-Fant (LF) model for coarse structure, and energy and perturbation measures for fine structure. Addition of these features to a baseline mel-frequency-cepstral-coefficient (MFCC) system resulted in improvement in performance for tests on both high-quality and landline-telephone recordings. Gudnason and Brooks [9] calculated MFCCs on the linear-prediction (LP) spectra exacted from samples taken from the closed-glottis phase of voicing. The latter were subtracted from regular MFCCs and this was used as a parameterization of the voice-source spectrum. This lead to an improvement in performance on a speaker-verification task.

2. Methodology

2.1. Database

The data were extracted from a database of two non-contemporaneous voice recordings of each of 60 female speakers of Standard Chinese [10]. See Morrison et al. [11] for details of the data collection protocol. The speakers were all first-language speakers of Standard Chinese from north eastern China, and were aged from 23 to 45. 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, with the second recording of each speaker recorded 2–3 weeks after the first. High-quality recordings were made at 44.1 kHz 16 bit using flat-frequency response lapel microphones (Sennheiser MKE 2 P-C) and an external soundcard (Roland® UA-25 EX), with one speaker on each of the two recording channels.

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

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U.S. Government.

2.2. Channel degradation

In addition to the original high-quality recordings, the database also includes versions of the same recordings which have been degraded by passing them through transmission channels. In the present study we compare high-quality v high-quality recordings, mobile-to-landline v mobile-to-landline recordings, and mobile-to-landline v high-quality recordings, where on each trial the first member of each pair is the channel of the nominal offender recording and the second the channel of the nominal suspect recordings.

Mobile-to-landline recordings were created as follows: A mobile telephone (Nokia 2730 classic) was place in a sound booth (IAC 250 Series Mini Sound Shelter) in the vicinity of a Roland® MA-7A loudspeaker which was in turn connected to a computer via one output channel of a Roland® UA-25 EX sound card. A call was established between the mobile telephone and a landline telephone (Polaris NRX EVO 450). The high-quality recordings were played though the loudspeaker and the acoustic signal picked up by the built-in microphone of the mobile telephone through which a call was established to the receiving telephone. The landline telephone was connected to an input channel of the sound card (not the same channel as was being used to output the original recording) via a Trillium Telephone Recording Adapter Studio Interface (REC-ADPT-SI). Custom software started recording, started playing a high-quality recording, then stopped recording 500 ms after the latter recording had finished playing. The degraded signal was recorded at the same sampling rate as the high-quality recording (44.1 kHz at 16 bits). The degraded recording was time-aligned with the original recording by displacing the degraded recording relative to the original recording one frame at a time and calculating the correlation between the two signals. At the displacement with the highest correlation, the degraded recording was truncated to the same start and end points as the original recording. Because of this alignment, the same markers as had been used to extract the tokens for the high-quality recording (see §2.3) could than also be used for the degraded recording.

2.3. Segment selection

For the extraction of frequency-domain based voice source features in GLOTTEX® the filtering effects of the vocal tract have to be reversed [12]. Ideally, this is done using relatively long speech segments with stationary supralaryngeal vocal-tract configurations (short speech segments do not provide a sufficiently large sample and rapidly changing segments, such as diphthongs, do not provide a sample of a fixed vocal-tract configuration). In medical settings, this is achieved by having the patient produce sustained vowel sounds, particularly tokens of /a/, but it is not possible to obtain such speech in forensic contexts. In the spontaneous speech typical of forensic contexts, the most appropriate segments to analyze are typically pause fillers, such as “um” and “ah” in English. In the Chinese database, an number of pause fillers (/a/, /e/, /n/, /ŋ/) were manually located and their start and end times marked. Of these, only /n/ had a sufficiently large number of tokens per speaker per recording session for there to be an a priori expectation of reasonable estimates of parameters in the forensic-voice-comparison system’s statistical models. There were between 3 and 73 tokens of /n/ per recording per speaker.

2.4. Voice source features

A number of different measurements of voice-source properties made by GLOTTEX® were evaluated. These are listed below. Each major dot point in the list subsumes a number of features which were grouped together to form individual forensic-voice-comparison systems (the abbreviation for each group of features / system appears in italics within parenthesis). Features marked with an asterisk are described in thew sections indicated, detailed descriptions of the full set of features is provided in [4].

- Distortion features and fundamental frequency (*distortion*, 6 features total):
    1. Absolute normalized jitter, measured as ratio of the difference of $f0$ between neighboring phonation cycles normalized by the average value for the segment.
    2. Normalized amplitude shimmer, measured as the ratio of the difference between maximum peak amplitudes of neighboring phonation cycles, normalized by the average value for the segment.
    3. Slenderness shimmer. The negative spike of the glottal pulse forms an approximate triangle. The slenderness is defined as the height of the triangle divided by its width. Slenderness shimmer is the ratio of the difference between the slenderness of neighboring phonation cycles normalized by the average value for the segment.
    4. Area shimmer, measured as the ratio of the difference of the area under the curve of the glottal source signal of neighboring phonation cycles, normalized by the average value for the segment.
    5. Glottal-to-noise excitation ratio.
    6. Fundamental frequency ($f0$).
- First-through-14th cepstral coefficients obtained from the mucosal wave correlate power spectrum (*MWC cepstra*, 14 features total). *§ 2.4.3
- Singularities in the mucosal-wave correlate power spectrum (*PSD singularities*, 14 features total): *§ 2.4.4
    1. Amplitude and frequency of first maximum.
    2. Normalized amplitude and frequency of second and third maxima, and of the first and second minima.
    3. Slenderness of the first two minima.
- Relative times of singularities extracted from the decomposed glottal source signal, MWC signal, and MWC time-derivative over a phonation cycle (*time based*, 9 features total). See Figure 5.

2.4.1. Separation of glottal source and vocal tract

The voice source parameterization depends on a separation of the effects of the glottal source and the vocal tract systems using iterative inverse filtering (Figure 1). After compensating for the high-pass effect of radiation from the lips ($R_l^{-1}(z)$), the voiced signal $s_l(n)$ is filtered by a *Glottal Pulse Inverse Model* $H_g(z)$, a k-th order prediction error adaptive lattice filter, to remove the strong glottal formant spectral envelope [4]. From the residual, the de-glottalized signal $s_v(n)$, parameters of a *Vocal Tract Model* $F_{VT}(z)$ are obtained. These estimates are in turn used in the *Vocal Tract Inverse Model* $H_{VT}(z)$ on the radiation-compensated speech signal to cancel the filtering effect of the vocal tract. From the residual, the parameters of the *Glottal Pulse Model* $F_g(z)$ are updated, which in the next iteration are used as the *Glottal Pulse Inverse Model* $H_g(z)$. This procedure is repeated 2–3 times to obtain the glottal source signal $s_g(n)$ [4].

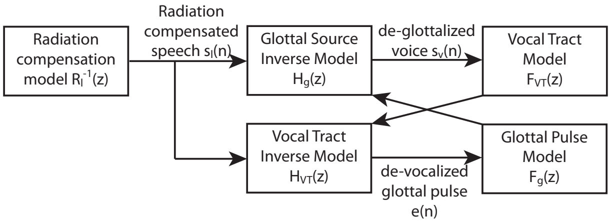
Radiation compensation model $R_l^{-1}(z)$
Radiation compensated speech $s_l(n)$
Glottal Source Inverse Model $H_g(z)$
de-glottalized voice $s_v(n)$
Vocal Tract Model $F_{VT}(z)$
Vocal Tract Inverse Model $H_{VT}(z)$
de-vocalized glottal pulse e(n)
Glottal Pulse Model $F_g(z)$

Figure 1: Procedure for separating the effects of the glottal source and the vocal tract (from Gómez-Vilda et al. [4]).

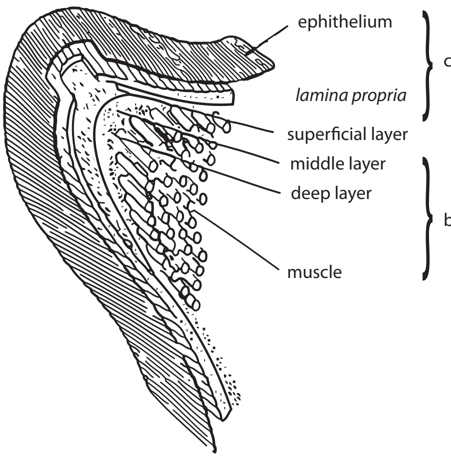

Figure 2: The layered structure of the vocal folds (from Hirose [13], p. 140).

2.4.2. Voice source parameterization

The vocal folds have a layered structure consisting of the mucosa epithelium, the lamina propria mucosa, and the vocalis muscle (see Figure 2). Hirano [14] proposed a *body-cover model* of the vocal folds, in which the "body" surrounds the vocalis muscle and the "cover" is formed by the epithelium and the superficial layer of the lamina propria. Based on this, mechanical k-mass models have subsequently been developed to mathematically describe the dynamic properties of the vocal folds. The body is characterized by a large mass component and $k-1$ cover masses linked by springs between themselves and the body mass. Figure 3 shows a two-mass vocal fold model which captures lateral vocal fold as well as mucosal wave motion [15].

The vibration of the cover tissue during phonation is commonly called the *mucosal wave*. The upper part of the vocal folds follows the lower part, forming a wave-like motion. Figure 4 shows the cycle of vocal fold vibration. In k-mass models, these movements are described by the masses and springs/dampers.

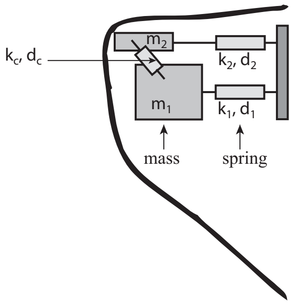

Figure 3: Two-mass vocal fold model (from Story [15]).

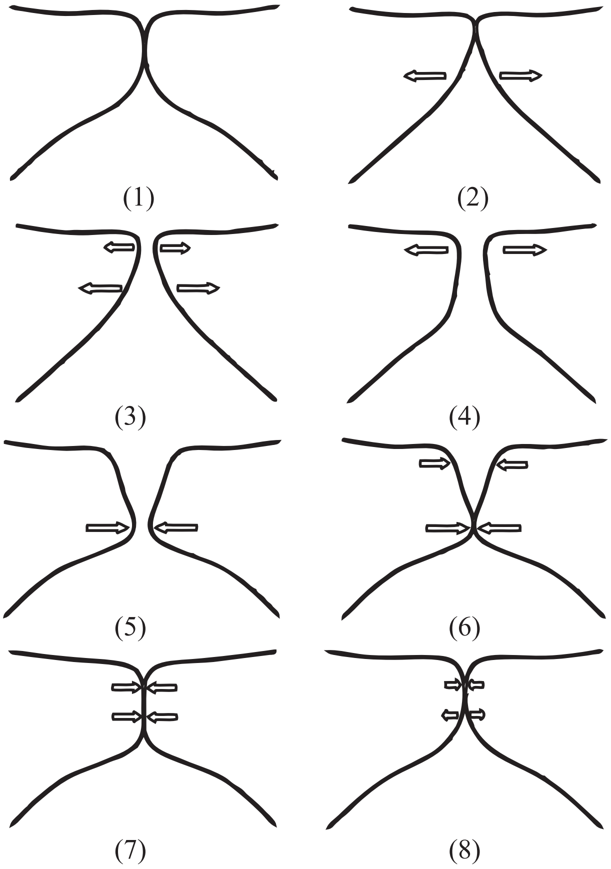

Figure 4: Idealized cycle of vocal fold vibration (from Story [15], p. 197).

2.4.3. Mucosal-wave correlate power spectrum

The glottal source signal obtained by separating the effects of the glottal source and the vocal tract (§2.4.1) includes biomechanical effects of both the vocal fold body and the cover. To characterize the latter, the mucosal wave, GLOTTEX® decomposes it in two parts, the *Average Acoustic Wave* (AAW) representing low-order vibration of the vocal folds, i.e., the vocal fold body, and the *Mucosal Wave Correlate* (MWC) capturing the higher-order vibrations of the cover. The AAW represents a second order system response (one-mass model) [4] and is defined as a sinusoid,

$$s_{sk}(n) = y_{0k} \sin(\omega_k n \tau), n \in N_k. \tag{1}$$

N_k represents the samples in the k-th phonation cycle, $\omega_k = \pi/T_k$ is the angular frequency corresponding to double the cycle period, τ is the sampling period, and y_{0k} is the optimal amplitude which is adaptively minimized to the difference between the AAW and glottal source signal. The details of the algorithm are described in Gómez-Vilda et al. [4]. The mucosal wave correlate is then obtained as the difference of the glottal source and the average acoustic wave as

$$s_{mk}(n) = s_{gk}(n) - s_{sk}(n). \tag{2}$$

The power spectrum calculated from the mucosal wave correlate is then characterized by obtaining first-to-14th order cepstral coefficients.

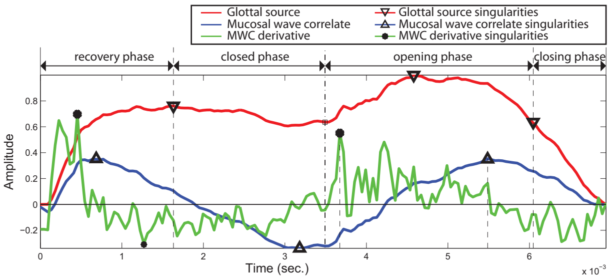

Figure 5: Estimating times of instances of maxima and times of phonation phases from the glottal source, mucosal wave correlate, and derivative signals.

2.4.4. *MWC power spectrum singularities*

Figure 6 shows an example of the mucosal wave correlate power spectrum. There are a series of local maxima and minima (collectively designated as *singularities*) which are related to the vocal-fold biomechanics [4]. GLOTTEX® measure the frequencies and amplitudes of the singularities indicated in Figure 6. The slenderness of the first two minima in the spectrum $(\sigma_{m1}, \sigma_{m2})$ is measured as:

$$\sigma_{mq} = \frac{f_{Mq}(2T_{mq} - T_{Mq+1} - T_{Mq})}{2(f_{Mq+1} - f_{Mq})}; q \in \{1, 2\}. \tag{3}$$

where T and f indicate absolute amplitudes and frequencies, and m indicates a minimum and M indicates a maximum.

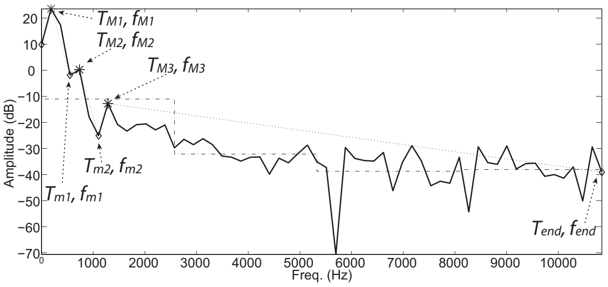

Figure 6: Estimating singularities of the mucosal wave correlate power spectral density.

2.5. Likelihood ratio calculation

2.5.1. *Voice-source-feature systems*

Likelihood ratios were calculated by using the Gaussian mixture model - universal background model (GMM-UBM) approach [16]. A UBM was trained using the background data pooled across both recording sessions. Suspect GMMs were trained via maximum a-posteriori (MAP) adaptation from the UBM. Full covariance matrices and a relatively small numbers of Gaussians were used, since in preliminary work this was found to give better performance than diagonal covariance matrices and larger numbers of Gaussians. For each group of features, the optimal number of Gaussians in the mixture and the optimal number of adaptation iterations were empirically determined via tests using development data. Table 1 shows the number of Gaussians and number of MAP adaptation iterations applied to each group of features in each channel condition.

Table 1: Number of Gaussians in the mixture and number of MAP adaptation iterations for each group of features.

high-quality v high-quality recordings:

feature group	num Gaussians	num iterations
Distortion	12	1
MWC Cepstra	32	3
PSD Singularities	8	1
Time-based	32	1

mobile-to-landline v mobile-to-landline recordings:

feature group	num Gaussians	num iterations
Distortion	16	3
MWC Cepstra	16	1
PSD Singularities	8	1
Time-based	16	1

mobile-to-landline v high-quality recordings:

feature group	num Gaussians	num iterations
Distortion	12	1
MWC Cepstra	16	5
PSD Singularities	8	2
Time-based	8	1

For each comparison trial, a score was obtained as in Equation 4:

$$score = \frac{1}{k}\sum_{i=1}^{k}(\log p(X_i|\lambda_{\text{speaker}}) - \log p(X_i|\lambda_{\text{UBM}})). \tag{4}$$

where X_i are the feature vectors from the offender recording, and $\lambda_{speaker}$ and λ_{UBM} are the GMM for the speaker and the background, respectively.

Scores calculated on the development set were used to calculate weights for logistic-regression calibration and fusion [17, 18, 19, 20] which was subsequently applied to convert the scores from the test set to likelihood ratios (calculations were performed using [21] and [22]).

In both the development and test sets, every speaker's Session 1 recording (nominal offender recording) was compared with their own Session 2 recording (nominal suspect recording), and also with every other speaker's Session 2 and Session 1 recordings separately (nominal suspect recordings). This resulted in 20 scores from same-speaker comparisons and 760 pairs of scores from different-speaker comparisons.

In the channel-mismatch condition, the nominal offender recordings were mobile-to-landline, and the nominal suspect recordings and the background recordings were high-quality recordings.

2.5.2. Baseline system

The voice-source-feature-based systems were fused with a baseline MFCC GMM-UBM system: 16 MFCC values were extracted every 10 ms over the entire speech-active portion of every recording using a 20 ms wide hamming window. Delta coefficient values were also calculated and included in the subsequent statistical modeling [23]. Feature warping [24] was applied to the MFCCs and deltas before subsequent modeling. On the basis of tests on the development set the number of Gaussians in the GMM-UBM was set to 1024.

3. Results

In the results below we focus on the performance of systems which are fusions of voice-source-feature systems with the baseline system and the performance of these systems relative to the baseline system.

3.1. Evaluation metrics

The validity and reliability of the systems was evaluated using the log likelihood-ratio cost (C_{llr}) as a metric of validity (accuracy), and an estimate of the 95% credible interval (95% CI) as a metric of reliability (precision) [25, 26] (C_{llr} was calculated using the mean procedure and the 95%CI using the parametric procedure). Readers familiar with automatic speaker recognition but not forensic voice comparison should note that metrics such as equal error rate (EER) and plots such as detection error trade-off (DET) [27] are not presented here since they are based on imposing hard thresholds on posterior probabilities and are therefore incompatible with the likelihood-ratio framework for the evaluation of forensic evidence [25, 28].

Since validity and reliability results must be simultaneously considered, these are plotted in two dimensions with the 95% CI on the x axis and C_{llr} on the y axis. For both metrics, smaller values indicate better performance, hence results closer to the origin are better (both are constrained to be greater than zero and, if the system is appropriately calibrated, C_{llr} is not expected to be greater than one). Tippett plots of selected systems are also provided (see [29] §99.330 for an introduction to the interpretation of Tippett plots).

3.2. High-quality v high-quality recordings

Figure 7 shows the results for the high-quality v high-quality tests. The baseline system had a C_{llr} of 0.021 and a $\log_{10}$ 95% CI of 1.42 (a Tippett plot is provided in Figure 8). Of the fusions of individual voice-source-feature systems with the baseline system, no fused system clearly outperformed the baseline system.

3.3. Mobile-to-landline v mobile-to-landline recordings

Figure 11 shows the results for the mobile-to-landline v mobile-to-landline tests. The baseline system had a C_{llr} of 0.099 and a $\log_{10}$ 95% CI of 2.11 (a Tippett plot is provided in Figure 10). Of the fusions of individual voice-source-feature systems with the baseline system, no fused system clearly outperformed the baseline system.

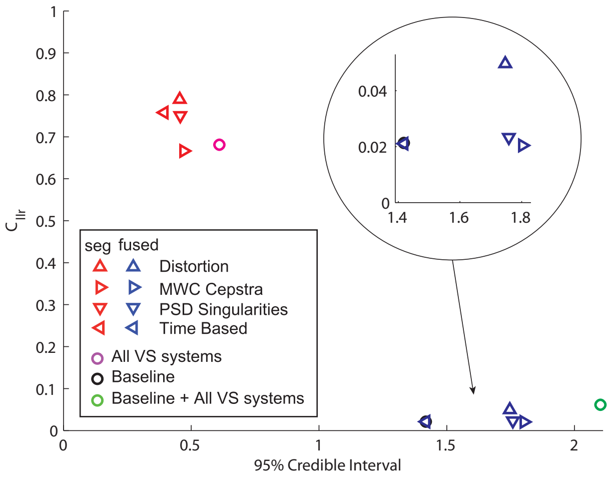

Figure 7: Measures for validity (C_{llr}) and reliability ($log10$ 95% credible interval) for the voice source feature systems individually (*red*) as well as after fusion with the generic fully-automatic baseline system (*blue*) (high-quality v high-quality recordings).

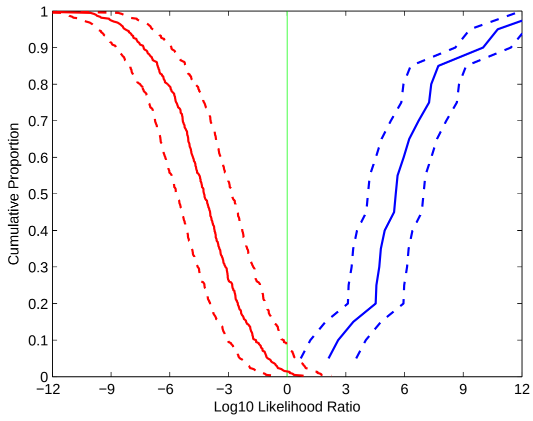

Figure 8: Tippett plot of the baseline fully-automatic MFCC-based system (high-quality v high-quality recordings).

3.4. Mobile-to-landline v high-quality recordings

Figure 9 shows the results for the mobile-to-landline v high-quality tests. The baseline system had a C_{llr} of 0.064 and a $\log_{10}$ 95% CI of 3.11. The best two fused systems were *PSD singularities* with a C_{llr} of 0.064 and a $\log_{10}$ 95% CI of 3.07, and *MWC cepstra* with a C_{llr} of 0.069 and a $\log_{10}$ 95% CI of 2.98. These showed small improvements in reliability with no degradation or only slight degradation in validity. This suggests that the use of these voice-source features may help in the most-challenging channel-mismatch condition, but given the small differences in performance observed, one would want to attempt to replicate the results on other databases before making any stronger claims. Tippett plots of the baseline system and the *MWC cepstra* system are provided in Figure 12.

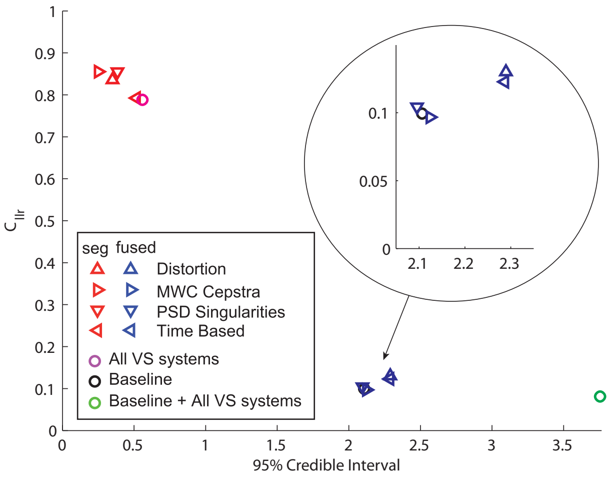

Figure 9: Measures for validity (C_{llr}) and reliability ($log10$ 95% credible interval) for the voice source feature systems individually (*red*) as well as after fusion with the generic fully-automatic baseline system (*blue*) (mobile-to-landline v mobile-to-landline recordings).

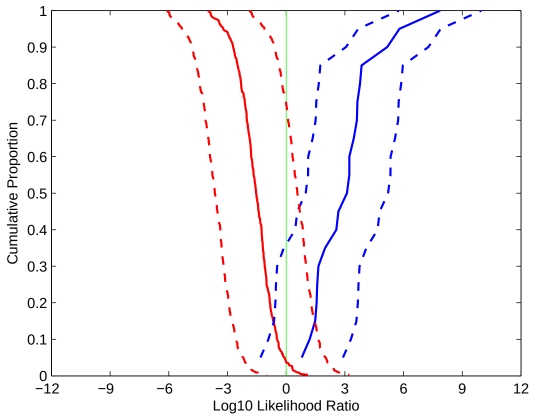

Figure 10: Tippett plot of the baseline fully-automatic MFCC-based system (mobile-to-landline v mobile-to-landline recordings).

4. Conclusion

The present paper evaluated the use of voice-source features extracted by GLOTTEX® as part of a forensic-voice-comparison system. Features were extracted from tokens of a pause filler, /n/, in a database of recordings of female speakers of Standard Chinese. We were not able to obtain any substantial improvement in performance over a baseline MFCC GMM-UBM system in any of the three channel conditions tested (high-quality v high-quality recordings, mobile-to-landline v mobile-to-landline recordings, and mobile-to-landline v high-quality recordings).

Potential reasons for failing to find improvement could be that for some sessions of some speakers, the number of tokens was very small. Also, although the manufacturer of GLOTTEX® assured us that the procedure could be applied to nasals, the inverse vocal-tract filtering in GLOTTEX® appears to assume an all-pole model of speech production, and it may be that it does not adequately account for zeros in the spectra of nasals. The codecs in mobile-telephone systems also use all

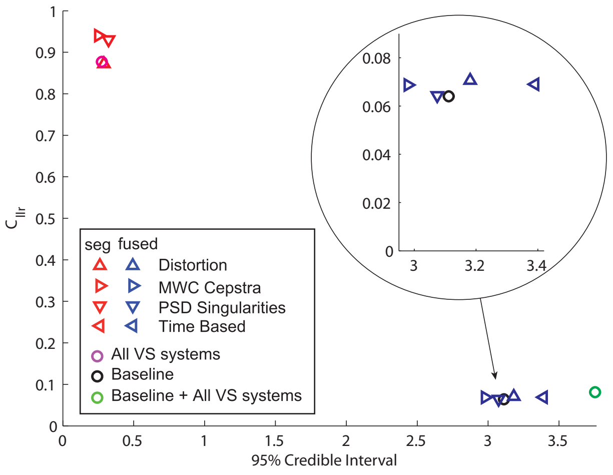

Figure 11: Measures for validity (C_{llr}) and reliability ($log10$ 95% credible interval) for the voice source feature systems individually (*red*) as well as after fusion with the generic fully-automatic baseline system (*blue*) (mobile-to-landline v high-quality recordings).

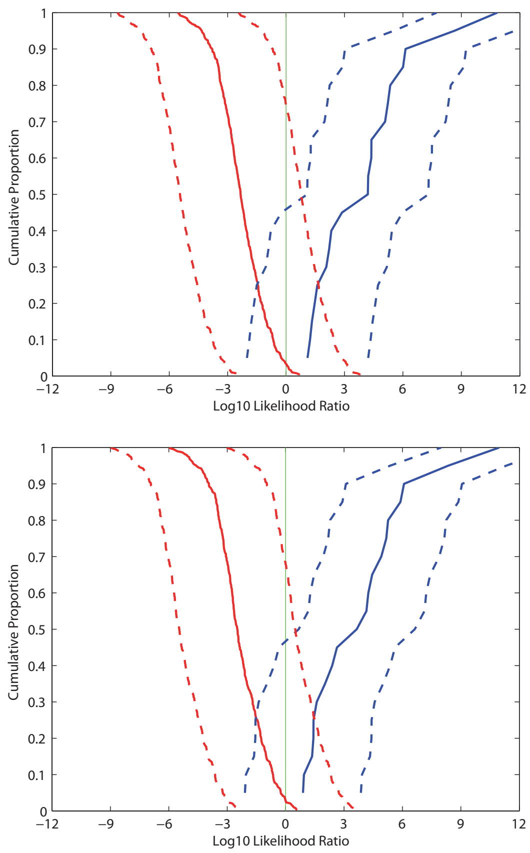

Figure 12: Tippett plot of the baseline fully-automatic MFCC-based system (top) and after fusion with the MWC cepstra system (bottom) (mobile-to-landline v high-quality recordings).

-pole models which would not effectively model spectral zeros, and hence an all-pole assumption in GLOTTEX® may not be problematic in the mobile-to-landline channel conditions. One should also bear in mind that useful information could be extracted even if there was a mismatch between theory and practice.

Although we have not found any evidence to support Gómez-Vilda et al.'s [6] proposal that GLOTTEX® would be effective in forensic voice comparison, they presented some empirical results which supported their proposal. Gómez-Vilda et al. [6] lacks details of their experimental methodology, but the results appear to be the same as those reported in Gómez-Vilda et al. [30] which provides a little more detail on the methodology. The features that they report using do not appear to be exactly the same as those output by the current version of GLOTTEX® . The database of voice recordings they used consisted of phonetically-balanced read speech from apparently a single recording session per speaker [31], and apparently only high-quality recordings were analyzed. Their data were therefore highly unrealistic compared to what one would expect in forensic casework. Our data, non-contemporaneous recordings of spontaneous telephone speech, comes much closer to being forensically realistic. As best we can tell, they sampled the voice-source features using 32-ms wide sliding windows over all the voiced sections of their recordings, rather than sampling only within tokens of a particular phonetic-unit as we did. Rather than using logistic-regression fusion as we did, they concatenated voice-source features with baseline regular MFCCs in a GMM-UBM system. As an additional test of GLOTTEX® for forensic voice comparison, we emulated this basic methodology in part by extracting features from the whole recordings. Given the restrictions of the software, we partitioned the speech active portion of the recordings into smaller segments of 200 ms duration and used GLOTTEX® for feature extraction using the same setting as before. Due to processing constraints we used the same GMM-UBM implementation as in the baseline MFCC system using diagonal covariances. As before, the number of Gaussians in the mixture was determined based on tests on the development set (64, 128, 256, 512, 1024).

Figure 13 shows the results for the high-quality v high-quality tests. Performance after fusion with the baseline resulted in substantial increases in validity with small decreases in reliability. In the high-quality v high-quality condition, this approach, does appear to result in better performance than the segmental approach (results for other conditions were not available as of the submission deadline for the present paper).

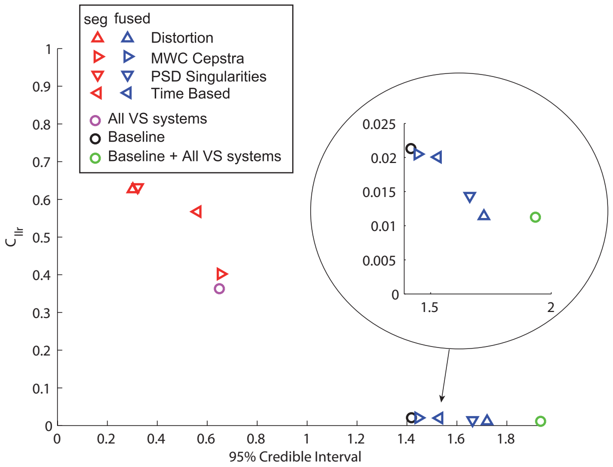

Figure 13: Measures for validity (C_{llr}) and reliability ($log10$ 95% credible interval) for the voice source feature systems applied to the total speech-active portion of the recordings individually (*red*) as well as after fusion with the generic fully-automatic baseline system (*blue*) (high-quality v high-quality recordings).

5. References

[1] O. Köster, M. Jessen, F. Khairi, and H. Eckert, "Auditory-perceptual identification of voice quality by expert and non-expert listeners," in *Proceedings of the XVI International Congress of the Phonetic Sciences (ICPhS)*, Saarbrcken, Germany, 2007, pp. 1845–1848.

[2] A. Hirson and M. Duckworth, "Glottal fry and voice disguise: a case study in forensic phonetics," *J. Biomed. Eng.*, vol. 15, pp. 193–200, 1993.

[3] J.S. Gruber and P. Poza, *Voicegram identifiaction evidence*, vol. 54 of *American Jurisprudence Trials*, Westlaw, 1995.

[4] P. Gómez-Vilda, R. Fernández-Baillo, V. Rodellar, V. Nieto, A. Álvarez, L.M. Mazaira-Fernándeza, R. Martínez, and J.I. Godino-Llorenteb, "Glottal source biometrical signature for voice pathology detection," *Speech Communication*, vol. 51, no. 9, pp. 759–781, 2009.

[5] P. Gómez-Vilda, R. Fernández-Baillo, V. Nieto, F. Díaz, F.J. Fernández-Camacho, V. Rodellar, A. Álvarez, and R. Martínez, "Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters," *Journal of Voice*, vol. 21, no. 4, pp. 450–476, 2006.

[6] P. Gómez-Vilda, A. Álvarez, L.M. Mazaira, R. Fernández-Baillo, V. Nieto, R. Martínez, C. Muñoz, and V. Rodellar, "Decoupling vocal tract from glottal source estimates in speaker's identification," *Language Design (Special Issue)*, pp. 111–118, 2008.

[7] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *Proc. Interspeech*, Antwerp, Belgium, August 2007, pp. 778–781.

[8] M.D. Plumpe, T.F. Quatieri, and D.A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Proc.*, vol. 7, no. 5, pp. 569–586, 1999.

[9] J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in *Proc. ICASSP*. IEEE, 2008, pp. 4821–4824.

[10] C. Zhang and G.S. Morrison, "Forensic database of audio recordings of 68 female speakers of standard chinese," 2011, Available: http://databases.forensic-voice-comparison.net/.

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

[12] A. Ní Chasaide and C. Gobl, "Voice source variation," in *The Handbook of Phonetic Sciences*, W.J. Hardcastle and J. Laver, Eds., pp. 427–461. Blackwell, Oxford, 2010.

[13] H. Hirose, "Investigating the physiology of laryngeal structures," in *The Handbook of Phonetic Sciences*, W.J.

Hardcastle and J. Laver, Eds., pp. 130–152. Blackwell, Oxford, 2010.

[14] M. Hirano, “Morphological structure of the vocal cord as a vibrator and its variations,” *Folia phoniatrica*, vol. 26, no. 2, pp. 89–94, 1974.

[15] B.H. Story, “An overview of the physiology, physics and modeling of the sound source for vowels,” *Acoustical Science and Technology*, vol. 23, no. 4, pp. 195–206, 2002.

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

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

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

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

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

[21] N. Brümmer, “Tools for fusion and calibration of automatic speaker detection systems,” 2005.

[22] G. S. Morrison, “Robust version of train_llr_fusion.m from Niko Brümmer’s FoCaL Toolbox (release 2009-07-02),” 2009.

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

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

[25] G. S. Morrison, “Measuring the validity and reliability of forensic likelihood-ratio systems,” *Science & Justice*, 2011.

[26] G. S. Morrison, T. Thiruvaran, and J. Epps, “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*, H. Cernocký and L. Burget, Eds., Brno, Czech Republic, 2010, International Speech Communication Association, pp. 63–70.

[27] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET curve in assessment of detection task performance,” in *Proc. Eurospeech ’97*, 1997, pp. 1895–1898.

[28] G. S. Morrison, “A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM-UBM),” *Speech Communication*, vol. 53, pp. 242–256, 2010.

[29] G. S. Morrison, *Expert Evidence*, chapter Forensic voice comparison, Thomson Reuters, Sydney, Australia, 2010.

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

[31] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.B. Mariño, and C. Nadeu, “Albayzin speech database: Design of the phonetic corpus,” in *Proceedings of Eurospeech ’93*, Berlin, Germany, 1993, International Speech Communication Association, pp. 175–178.