RESEARCH PAPER

Deceptive Acoustics in Medicolegal and Psychological Practice: An AI-Powered Framework for Detecting Voice Manipulation in Forensic Evidence

Shikha Upadhyay a* and **Murugan R.** b

a Research Scholar, Department of Forensic Science, JAIN (Deemed-to-be University), Bangalore 560069, India

[b] Professor, School of Computer Science and IT, JAIN (Deemed-to-be University), Bangalore 560069, India

ABSTRACT

Background: Voice manipulation technologies capable of altering perceived gender or emotional register pose acute challenges for forensic medicine, psychological testimony, and medicolegal proceedings. When a recorded voice may serve as evidence in criminal, civil, or psychiatric evaluations, the inability to detect acoustic transformation constitutes a significant gap in forensic practice.

Objectives: To develop and validate an AI-based acoustic framework for detecting two forensically distinct voice manipulation types — gender conversion and emotion conversion — with direct application to medicolegal voice authentication and psychological credibility assessment.

Methods: Using CREMA-D (7,442 utterances; 91 speakers) and RAVDESS (1,440 utterances; 24 speakers), a corpus of 13,560 samples was constructed across three conditions: original, gender-converted (WORLD vocoder), and emotion-converted (StarGAN-VC). A 74-dimensional acoustic feature vector was statistically validated; four classifiers were evaluated under speaker-stratified ten-fold cross-validation.

Results: The MLP achieved F1-scores of 0.913 ± 0.019 for gender manipulation detection (AUC = 0.971) and 0.864 ± 0.024 for emotion manipulation detection (AUC = 0.936), with balanced error rates of 7.35% and 13.45% respectively.

Conclusions: AI-based acoustic analysis reliably detects voice manipulation with direct forensic and medicolegal relevance. The feature-level findings are actionable for practitioners without specialist machine learning infrastructure.

Keywords: *forensic voice analysis; voice manipulation detection; medicolegal evidence; speech emotion recognition; psychological credibility; deepfake voice; acoustic features; MFCC; forensic psychology; digital forensics*

How to cite this article: Upadhyay S, Murugan R. Deceptive Acoustics in Medicolegal and Psychological Practice: An AI-Powered Framework for Detecting Voice Manipulation in Forensic Evidence. Int J Drug Deliv Technol. 2026;16(55s): 214-221. DOI: 10.25258/ijddt.16.55s.24

1. Introduction: The Medicolegal Challenge of Voice Manipulation

Voice recordings have become one of the most consequential categories of evidence in modern forensic practice. They appear in criminal prosecutions, civil disputes, domestic abuse cases, terrorist threat assessments, and psychiatric evaluations of competency and psychological state. Their evidential power rests on an implicit assumption: that the acoustic characteristics captured in a recording faithfully reflect the speaker who produced it. Advances in voice conversion technology have made that assumption increasingly fragile.

Contemporary voice manipulation software can alter two of the most psychologically loaded dimensions of speech — apparent gender and emotional register — while leaving every spoken word intact. The forensic and psychological implications are substantial. In criminal proceedings, a threat delivered in a fearful voice and the same threat delivered in cold anger are not equivalent in the assessment of intent, state of mind, or dangerousness. In medicolegal psychiatric evaluations, a recording used to assess a patient's emotional state at a specific time may have been acoustically re-rendered. Standard audio integrity methods — electric network frequency analysis, microphone inconsistency modelling — detect editing artefacts, not targeted parametric transformation of vocal characteristics that leaves the recording otherwise intact (Grigoras, 2005; ENFSI, 2015).

This study presents an AI-based acoustic framework addressing precisely this gap. Across 13,560 controlled experimental samples derived from two publicly available acted speech corpora, four machine learning classifiers were trained and validated to detect gender conversion and emotion conversion under a strict lexical constancy constraint. The framework provides quantified detection accuracy

with clinically interpretable error rates; identification of specific acoustic features carrying the manipulation signal; and explicit discussion of detection error rate implications for forensic reporting standards.

2. Background: Forensic, Psychological, and Technical Foundations

2.1 Voice Evidence in Forensic Medicine and Psychological Practice

Speaker comparison in forensic practice has moved from impressionistic auditory analysis to formal probabilistic frameworks over several decades. The ENFSI evaluative reporting guidelines (2015) formalise forensic conclusions as likelihood ratios under competing source propositions — a framework designed for authentic recordings of unknown speakers. What it does not address is the scenario where a recording is authentic in one sense — the words were genuinely spoken — but has been acoustically transformed in ways that alter the apparent identity or psychological-emotional state of the speaker.

In forensic psychology and psychiatric medicine, recorded voice evidence is used in competency evaluations, threat assessments, credibility analyses, and documentation of psychological distress. If the emotional register of a recording has been algorithmically altered — a fearful voice rendered angry, a distressed voice rendered calm — any clinical inference drawn from it is potentially compromised. No current standard forensic or clinical procedure provides validated detection of this type of manipulation.

2.2 Voice Conversion Technologies: Mechanisms and Forensic Implications

The WORLD vocoder (Morise et al., 2016) decomposes speech into independently modifiable components and resynthesises at high perceptual quality. Its gender conversion artefacts — bilinear spectral warping, scaled fundamental frequency contours — are parametric and predictable, leaving measurable acoustic traces even when imperceptible to untrained listeners. Neural conversion architectures, specifically StarGAN-VC (Kameoka et al., 2018), present a harder forensic target: artefacts are distributed across many acoustic dimensions and partially overlap with natural within-speaker emotional variability. Muller et al. (2022) demonstrated that even neural conversion outputs retain statistically detectable regularities, though with smaller effect sizes.

2.3 Psychological Dimensions: Emotion, Vocal Credibility, and Clinical Assessment

The relationship between vocal acoustics and psychological state has been extensively documented. Scherer's (2003) review established that emotional states produce coordinated, reproducible changes in fundamental frequency, formant structure, voice quality indices, and temporal features. Clinical applications include assessment of depression severity, anxiety, psychotic episodes, and post-traumatic stress — all contexts where a recorded voice sample may serve as evidence or clinical documentation. The forensic psychology literature further underscores the stakes: paralinguistic cues including vocal tremor, speech rate, and fundamental frequency variability directly influence credibility judgements made by jurors, clinicians, and investigators (Nolan, 1983; Rose, 2002). If these cues can be algorithmically substituted, the foundational assumptions of voice-based credibility assessment are compromised.

2.4 AI-Based Detection: State of the Art and Remaining Gap

The ASVspoof challenge series (Wu et al., 2015; Kinnunen et al., 2020) established primary benchmarks for synthetic speech detection. Feature-based approaches using MFCCs with voice quality measures have shown consistent detection performance (Khodabakhsh et al., 2017). Hasan et al. (2021) demonstrated that targeted pitch manipulation leaves measurable jitter and shimmer signatures. None of the existing studies, however, compare gender and emotion conversion as forensically distinct manipulation types with direct medicolegal application — the gap this study addresses.

3. Methodology: Experimental Design and Forensic Validity Constraints

3.1 Design Principles

Two constraints governed the experimental design, both motivated by forensic validity requirements. Lexical constancy: every utterance in the manipulated conditions carries exactly the same spoken words as its original counterpart — mirroring the forensic scenario of altered but verbally identical recordings. Speaker independence: cross-validation folds were constructed at the speaker level, ensuring no speaker's data appears in both training and test partitions within any single fold. The complete experimental framework is illustrated in Fig. 1.

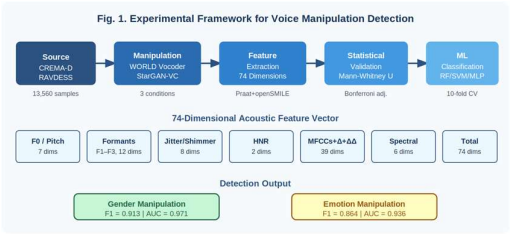

Fig. 1. End-to-end experimental framework showing data sources, manipulation pipelines, feature extraction, statistical validation, and classification outputs. Gender detection (teal) and emotion detection (amber) are evaluated as separate binary classification tasks.

3.2 Datasets and Corpus Construction

Two publicly available acted speech corpora were used. CREMA-D (Cao et al., 2014) provides 7,442 utterances from 91 speakers delivering 12 standardised sentences across 6 emotion categories — a complete factorial structure at the utterance level. RAVDESS (Livingstone and Russo, 2018) contributes 1,440 utterances from 24 speakers, introducing variability in recording conditions and acting background. After SNR screening (WADA-SNR < 15 dB; 38 files excluded), a speaker-stratified balanced subset of 4,520 utterances was drawn, yielding a total experimental corpus of 13,560 files across three conditions (Table 1).

Data set	Reference	Speakers	Files	Emotions	Gender	Licence
CREMA-D	Cao et al. (2014)	91	7,442	6	48M/43F	CC BY-NC-SA
RAVDESS	Livingstone & Russo (2018)	24	1,440	8	12M/12F	CC BY-NC-SA

Table 1. Source datasets. Both corpora use trained actors delivering scripted, emotionally differentiated utterances — suitable for lexically controlled manipulation experiments.

3.3 Acoustic Feature Extraction with Forensic Rationale

A 74-dimensional acoustic feature vector was extracted per utterance using Praat v6.4 (Boersma and Weenink, 2024) for prosodic and voice quality features, and openSMILE v3.0 (Eyben et al., 2010) with the ComParE 2016 configuration for cepstral and spectral features. Feature groups and their forensic rationale are presented in Table 2. All features were z-score normalised per feature within each training fold; normalisation parameters were applied without re-estimation to the test fold, preventing data leakage.

Feature Group	n	Forensic / Clinical Relevance	Manipulation Sensitivity
F0 — Pitch	7	Primary gender discriminator; reflects emotional arousal and speaker identity	Systematically displaced by WORLD vocoder; largest effect for gender detection
Formants F1–F3	12	Vocal tract length; fundamental to speaker identification in forensic comparison	Directly shifted by bilinear spectral warp; diagnostic for gender conversion
Jitter / Shimmer	8	Voice quality indices; elevated in neurological, psychological, and stress conditions	Elevated by vocoder resynthesis artefacts in both manipulation types
HNR	2	Harmonic regularity; reduced in stress, depression, and pathology	StarGAN reduces harmonic regularity more than WORLD; primary emotion indicator
MFCCs + Δ + ΔΔ	39	Spectral envelope; delta terms capture temporal dynamics of emotional expression	Distributed spectral shift; primary discriminator for emotion conversion detection
Spectral descriptors	6	Secondary energy descriptors; secondary to primary voice	Lowest diagnostic weight; not significant in binary pairwise tests

		quality features	

Table 2. Acoustic feature set (74 dimensions total) with forensic and clinical rationale for each group.

3.4 Statistical Validation and Classification Framework

Pre-classification statistical validation confirmed that observed acoustic differences are genuine distributional phenomena — essential for forensic expert testimony that must stand independently of black-box model outputs. Shapiro-Wilk normality testing preceded Bonferroni-corrected Mann-Whitney U tests ($\alpha\_adj = 0.000338$) and Kruskal-Wallis H tests ($\alpha\_adj = 0.000676$). Effect sizes were reported as rank-biserial correlation r and eta-squared η^2.

Four classifiers were evaluated: Random Forest (300 trees, Gini), SVM-RBF (C = 10, gamma = scale), XGBoost (300 estimators, lr = 0.05, max depth 6), and MLP (256-128-64 neurons, ReLU, batch normalisation, dropout 0.35, Adam optimiser, early stopping). Each task was binary (original vs. manipulated) under speaker-stratified ten-fold cross-validation. McNemar's test with Bonferroni correction ($\alpha\_adj = 0.0083$) was used for pairwise classifier comparison.

4. Results: Detection Performance and Forensically Actionable Feature Evidence

4.1 Statistical Feature Validation: Confirming Genuine Acoustic Differences

After Bonferroni correction, 61 of 74 features showed significant distributional differences between original and gender-modified speech; 54 of 74 between original and emotion-modified speech. The Kruskal-Wallis H test confirmed significant three-group differentiation for 68 of 74 features, all with large effect sizes ($\eta^2 > 0.14$). The six non-significant features — spectral flatness and flux — independently confirmed as lowest-ranked features by the Random Forest, validating feature selection through two separate analytical methods. Effect sizes for key forensically relevant features are visualised in Fig. 2.

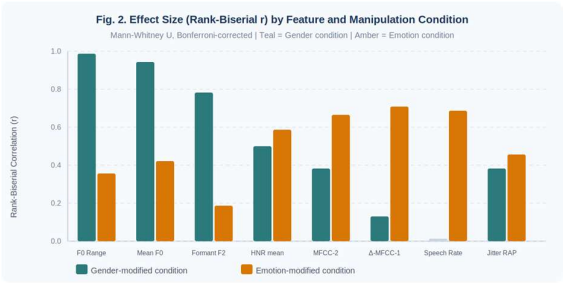

Fig. 2. Rank-biserial correlation (r) effect sizes for key acoustic features under gender-modified (teal) and emotion-modified (amber) conditions. Features significant at Bonferroni-corrected α = 0.000338. Gender detection is dominated by pitch and formant features; emotion detection by MFCC temporal dynamics and speech rate.

Feature	r (Gender)	r (Emotion)	H stat.	η^2	Forensic Significance
F0 Range	0.91*	0.33*	7,391.1**	0.545	Primary speaker identity marker
Mean F0	0.88*	0.39*	7,842.3**	0.578	Emotional arousal & gender identity
Formant F2	0.72*	0.17*	5,124.6**	0.378	Vocal tract length; speaker ID
Delta-MFCC-1	0.12*	0.65*	4,891.2**	0.361	Temporal spectral dynamics; emotion
Speech rate	ns	0.63*	4,102.7**	0.303	Psychological arousal indicator
HNR mean	0.46*	0.54*	3,244.8**	0.239	Voice quality; stress & pathology
Jitter RAP	0.35*	0.42*	—	—	F0 micropertubation; vocoder trace

*Table 3. Mann-Whitney U effect sizes and Kruskal-Wallis statistics for key forensically relevant features. * $p < 0.000338$; ** $p < 0.000676$; ns = not significant*

after Bonferroni correction. r = rank-biserial correlation; η² = eta-squared.

The r values (0–1) indicate how strongly each feature changes under manipulation — higher means more detectable. The asterisk (*) confirms statistical significance after Bonferroni correction ($p < 0.000338$). H statistic confirms simultaneous separation across all three groups — larger values indicate stronger group discrimination. η^2 values indicate practical importance; all values exceed 0.14, confirming large effects by Cohen (1988) conventions. Gender detection is dominated by pitch and formant features (F0 Range r = 0.91; Formant F2 r = 0.72), while emotion detection is driven by temporal spectral dynamics (Delta-MFCC-1 r = 0.65; Speech rate r = 0.63). HNR and Jitter RAP show moderate sensitivity to both manipulation types, reflecting harmonic irregularities introduced by both conversion systems.

4.2 Classification Performance Across All Architectures

Table 4 presents classification metrics for all four classifiers. The MLP achieved the highest performance on both tasks. The 4–5 F1 percentage point gap between gender and emotion detection was consistent across all four architectures — a finding McNemar's test confirmed is statistically significant (all pairwise comparisons $p < 0.01$, Bonferroni-corrected). This gap reflects a property of the manipulations themselves, not of any individual classifier.

Classifier	Condition	Accuracy	Precision	Recall	F1	AUC
Random Forest	Gender manipulation	0.867 ±0.021	0.863 ±0.024	0.871 ±0.022	0.867 ±0.021	0.938
Random Forest	Emotion manipulation	0.826 ±0.027	0.819 ±0.031	0.834 ±0.028	0.826 ±0.028	0.903
SVM (RBF)	Gender manipulation	0.883 ±0.019	0.879 ±0.021	0.887 ±0.020	0.883 ±0.019	0.951
SVM (RBF)	Emotion manipulation	0.841 ±0.023	0.836 ±0.026	0.847 ±0.024	0.841 ±0.024	0.916
XGBoost	Gender manipulation	0.899 ±0.017	0.896 ±0.019	0.902 ±0.018	0.899 ±0.018	0.962
XGBoost	Emotion manipulation	0.856 ±0.022	0.851 ±0.024	0.861 ±0.023	0.856 ±0.022	0.928
MLP (256-128-64)	Gender manipulation	0.913 ±0.019	0.909 ±0.021	0.917 ±0.020	**0.913 ±0.019**	**0.971**
MLP (256-128-64)	Emotion manipulation	0.864 ±0.024	0.860 ±0.027	0.868 ±0.025	**0.864 ±0.024**	**0.936**

Table 4. Classification performance across all four classifiers (10-fold speaker-stratified CV, macro-averaged). MLP (highlighted) achieved highest performance on both tasks. McNemar's test confirmed statistically significant MLP advantage ($p < 0.01$, Bonferroni-corrected).

4.3 Confusion Matrices and Clinically Interpretable Error Rates

The MLP confusion matrices (Fig. 3) reveal the direction and magnitude of classification errors in medicolegal terms. For gender detection, 388 of 4,520 original utterances were misclassified as manipulated (false positive rate 8.6%) and 277 of 4,520 gender-modified utterances were missed (false negative rate 6.1%), yielding a balanced error rate of 7.35%. For emotion detection, 623 false positives (13.8%) and 591 false negatives (13.1%) yielded a balanced error rate of 13.45%.

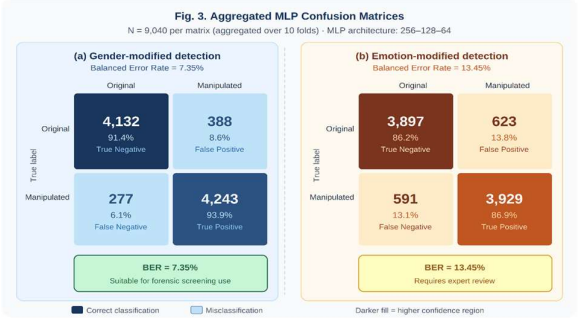

Fig. 3. Aggregated MLP confusion matrices for gender-modified (a) and emotion-modified (b) detection tasks. N = 9,040 per matrix (aggregated over 10 folds). Gender detection (BER = 7.35%) is substantially more reliable than emotion detection (BER = 13.45%), with important implications for forensic and clinical applications.

4.4 Per-Pair Emotion Detection and Psychological Distance Effects

Detection performance varied systematically across the six emotion transfer pairs, scaling with the acoustic and psychological distance between source and target emotion states (Table 5). Neutral→Angry (F1 = 0.879 ± 0.027) and Happy→Sad (F1 = 0.871 ± 0.031) — pairs spanning high arousal contrast and opposing valence — were most detectable. Disgust→Neutral (F1 = 0.839 ± 0.041) was hardest to detect, consistent with substantial acoustic overlap between disgust and neutral in acted speech. This gradient has direct forensic implications: manipulations of emotionally proximate states are less acoustically conspicuous and therefore harder to authenticate.

Transfer Pair	N	F1 (±SD)	Forensic / Psychological Observation
Neutral → Angry (N2A)	892	0.879±0.027	Largest prosodic contrast; anger raises F0, rate, and energy markedly — highly detectable
Happy → Sad (H2S)	738	0.871±0.031	Opposing valence; pitch range contraction in sadness reliably captured by delta-MFCC features
Angry → Neutral (A2N)	856	0.864±0.029	Removal of high-arousal cues detectable; slightly harder than N2A due to lower-energy target
Sad → Happy (S2H)	701	0.851±0.034	High within-class spread in happy speech increases overlap with converted signal
Neutral → Fearful (N2F)	748	0.844±0.036	Pre-trained model quality lower for this pair; higher fold variance reflects model uncertainty
Disgust → Neutral (D2N)	585	0.839±0.041	Hardest pair: disgust overlaps substantially with neutral in acted speech — smallest detectable change

Table 5. MLP F1-score by emotion transfer pair (10-fold CV). Detection difficulty scales with acoustic and psychological distance between source and target emotion states.

5. Discussion: Forensic, Clinical, and Medicolegal Implications

5.1 What the Statistical Evidence Establishes for Expert Testimony

The central forensic contribution of the pre-classification statistical validation stage is that it establishes a foundation for expert witness testimony independent of any machine learning model. Effect sizes for the most diagnostic gender features — F0 Range (r = 0.91), Mean F0 (r = 0.88) — are large enough to be both statistically significant at stringent Bonferroni-corrected thresholds and practically meaningful under established effect size conventions (Cohen, 1988). The convergence of classifier Gini importance rankings and independent Mann-Whitney effect sizes on the same feature hierarchy confirms findings through two entirely separate analytical routes, substantially strengthening evidentiary weight.

For the forensic practitioner, the feature-level findings offer actionable guidance without specialist computational infrastructure. Gender manipulation detection is dominated by pitch statistics and formant frequencies already employed in standard forensic speaker comparison. A practitioner examining a recording for suspected gender conversion can apply Bonferroni-corrected non-parametric tests on these features and obtain a statistically validated indication. Emotion manipulation detection requires a multivariate approach spanning MFCC temporal dynamics, HNR, and speech rate simultaneously — no single feature carries sufficient discriminative signal, making the classifier-based framework essential.

5.2 The Gender–Emotion Detection Gap: Implications for Psychological Assessment

The consistent 4–5 percentage point advantage in gender over emotion detection carries a specific implication for forensic psychology: the manipulation type most likely to affect psychological credibility assessment is also the harder to detect. Gender conversion produces large, coherent displacements in measurable acoustic dimensions that leave unambiguous traces. Emotion conversion produces distributed, moderate changes partially overlapping with natural within-speaker emotional variability — precisely the variability that makes emotional credibility assessment challenging.

For medicolegal practitioners working with voice evidence in domestic violence cases, threatening communications, or psychiatric competency evaluations, this asymmetry argues for heightened scrutiny and explicit acknowledgement of detection uncertainty when emotional register — not merely speaker identity — is at issue.

5.3 Error Rates, Likelihood Ratios, and ENFSI Reporting Standards

A 13.8% false positive rate for emotion manipulation detection means approximately one in seven genuine recordings could be flagged as manipulated. In the binary decision context of criminal or civil proceedings, this is not an acceptable operational error rate for a primary finding. These classifiers are appropriately used as screening tools guiding further expert examination, not as autonomous authentication systems.

ENFSI evaluative reporting guidelines require forensic conclusions expressed as likelihood ratios quantifying the probability of acoustic observations under competing source propositions. The confusion matrix data generated by this framework provide the necessary components for likelihood ratio computation: sensitivity, specificity, and balanced error rate at the speaker-stratified level that reflects the forensic scenario. Reformulation of this framework in explicit likelihood ratio terms constitutes the most important direction for future clinical and forensic deployment.

6. Conclusion: Towards AI-Assisted Forensic Voice Authentication

Voice manipulation that alters the perceived gender or emotional state of a speaker while preserving every spoken word is technically accessible, perceptually convincing, and forensically consequential. This study demonstrates, across 13,560 controlled experimental samples with speaker-stratified validation, that both manipulation types leave acoustically detectable traces. The MLP classifier achieved F1-scores of 0.913 ± 0.019 for gender detection and 0.864 ± 0.024 for emotion detection.

Three findings carry direct relevance for forensic medicine and psychology. First, the pre-classification statistical validation confirms that acoustic differences are genuine distributional phenomena with large effect sizes — providing a foundation for expert testimony independent of black-box model outputs. Second, the feature hierarchy — pitch and formants for gender manipulation, MFCC dynamics and HNR for emotion — is directly actionable for forensic examiners and clinical practitioners without specialist computational resources. Third, the error rate analysis provides the quantified uncertainty estimates that ENFSI evaluative reporting guidelines require for evidential conclusions about voice authenticity.

Extension to naturalistic speech, degraded channel conditions, additional converter types, and explicit likelihood-ratio reformulation for evidentiary use constitute the immediate priorities for future work. The intersection of AI-based acoustic analysis and forensic psychological practice is one in which computational advances and legal-clinical requirements must develop in step; this study offers a rigorous framework where both are explicitly addressed.

References

Ancilin, J., and Milton, A. (2021). Improved speech emotion recognition with Mel frequency magnitude coefficient. Applied Acoustics, 179, 108059.

Boersma, P., and Weenink, D. (2024). Praat: Doing phonetics by computer (version 6.4). https://www.praat.org.

Cao, H., Cooper, D. G., Kuchuk, M. K., Hu, R., Ha, R. A., and Bhanu, B. (2014). CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. IEEE Transactions on Affective Computing, 5(4), 377–390.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

European Network of Forensic Science Institutes (ENFSI). (2015). ENFSI guideline for evaluative reporting in forensic science.

Eyben, F., Wollmer, M., and Schuller, B. (2010). openSMILE: The Munich versatile and fast open-source audio feature extractor. Proceedings of ACM Multimedia, pp. 1459–1462.

Grigoras, C. (2005). Digital audio recording analysis — the electric network frequency criterion. International Journal of Speech, Language and the Law, 12(1), 63–76.

Hasan, T., Hansen, J. H. L., and Shokouhi, M. (2021). Detecting pitch manipulation in speech. Speech Communication, 134, 1–14.

Kameoka, H., Kaneko, T., Tanaka, K., and Hojo, N. (2018). StarGAN-VC: Non-parallel many-to-many voice

conversion. Proceedings of IEEE SLT 2018, pp. 266–273.

Khodabakhsh, A., Sonmez, F., Sahin, S., and Apaydin, T. (2017). Detection of synthetic speech for the problem of fake news. Proceedings of EUSIPCO 2017, pp. 2587–2591.

Kinnunen, T., et al. (2020). The ASVspoof 2017 challenge. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2298–2314.

Livingstone, S. R., and Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). PLoS ONE, 13(5), e0196391.

McReynolds, D., Barkman, T., Liang, J., and Rubin, J. (2023). Real-time speaker impersonation using short-duration reference audio. Proceedings of ICASSP 2023, pp. 1–5.

Morise, M., Yokomori, F., and Ozawa, K. (2016). WORLD: A vocoder-based high-quality speech synthesis system. IEICE Transactions on Information and Systems, E99-D(7), 1877–1884.

Muller, N., Czempin, P., Deiseroth, F., Bottcher, A., and Katzenbeisser, S. (2022). Does audio deepfake detection generalise? Proceedings of Interspeech 2022, pp. 2783–2787.

Nolan, F. (1983). The phonetic bases of speaker recognition. Cambridge University Press.

Rose, P. (2002). Forensic speaker identification. Taylor and Francis.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech Communication, 40(1), 227–256.

Todisco, M., et al. (2019). ASVspoof 2019. Proceedings of Interspeech 2019, pp. 1008–1012.

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., and Li, H. (2015). Spoofing and countermeasures for speaker verification. Speech Communication, 66, 130–153.

Yi, J., et al. (2022). ADD 2022: The first audio deepfake detection challenge. Proceedings of ICASSP 2022, pp. 9216–9220.

Declarations

Declaration of Competing Interest. The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability. CREMA-D: https://github.com/CheyneyComputerScience/CREMA-D. RAVDESS: https://zenodo.org/record/1188976. Derived audio files available from the corresponding author on reasonable request, subject to source licence compliance.