Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 20 May 2026
Revised 28 June 2026
Accepted 15 July 2026

Natural Resources for Human Health 2026, Vol. 6(7s): 1285–1293

KEYWORDS:
Emotional Speech Transformation, F0 Transplantation, Prosody Modification, Emotive environment, Neutral environment Speech Synthesis

Lightweight Emotional Speech Transformation using F0 Transplantation and Acoustic Feature Reprojection

Rashmi V Swamy[1], Shashidhar G Koolagudi[2]

[1]Department of Computer Science and Engineering, Research Scholar, NITK, Surathkal Mangalore, 575025, Karnataka, India
[1]Department of Supply Chain and Engineering, Senior Manager, Accenture, Bangalore, 560103, Karnataka India
[2]Department of Computer Science and Engineering, Associate Professor, NITK, Surathkal, Mangalore, 575025, Karnataka, India
Corresponding Author: Rashmi V Swamy
Email: rashmi.v.swamy@accenture.com

ABSTRACT: In emotional environments, speaker recognition systems often experience significant performance degradation due to the variability introduced by emotional speech. To address this challenge, this paper proposes a lightweight emotional speech transformation framework that synthesizes emotionally expressive speech directly from neutral utterances, reducing the need for large emotional corpora or deep generative models. The proposed approach employs an F0 transplantation and prosodic modulation technique, where the fundamental frequency (F0) and energy contours from emotionally expressive donor utterances are imposed onto neutral speech while preserving the speaker's identity. Using the Berlin Emotional Speech Database (Emo-DB), the system synthesizes emotional variants of neutral samples and quantitatively evaluates them against real emotional references. The evaluation shows that 73.6% of the transformed samples achieved a positive emotional shift with an average similarity gain of 0.0154, while maintaining high speaker identity preservation (93.8% Top-1 accuracy). The proposed method demonstrates that signal-level prosody adaptation can produce perceptually meaningful emotional transformations without requiring parallel data or model training. This framework provides an interpretable, data-efficient step toward bridging the gap between neutral and emotional speech.

1. INTRODUCTION

Speech conveys not only linguistic information but also rich emotional and identity-related cues. While modern speaker recognition systems perform well in neutral conditions, their accuracy often drops in emotional speaking environments. Emotional expressions alter prosodic and spectral features such as pitch, energy, and formants, leading to mismatches between neutral training and emotional test conditions. A practical solution is to transform a speaker's neutral speech into its emotional counterpart while preserving speaker identity. This transformation reduces dependence on emotional data, which are often limited and difficult to collect. Instead of relying on large-scale parallel datasets or deep generative models, this work adopts a lightweight and interpretable signal-level transformation approach that manipulates prosodic and pitch-based cues—particularly the fundamental frequency (F0)—to simulate emotional expression.

Speech transformation involves modifying acoustic features such as pitch, tone, or prosody while maintaining other attributes like linguistic content and speaker identity. In emotional transformation, the aim is to introduce expressive variations such as higher F0 and faster energy dynamics for anger or happiness, and lower, flatter contours for sadness or boredom. This study employs the Berlin Emotional Speech Database (Emo-DB), developed by the Institute of Communication Science, Technical University of Berlin. It includes recordings from ten professional German speakers (five male, five female) across seven

emotions: anger, boredom, anxiety (fear), happiness, sadness, disgust, and neutral, totaling 535 utterances at 16 kHz. Each filename encodes the speaker ID, text, emotion, and version (e.g., 03a01Fa.wav indicates Speaker 03 uttering text "a01" with emotion Freude—happiness).

The proposed method transforms neutral samples into emotional speech using an F0-transplant-based approach. Prosodic and spectral cues from emotional donor utterances are mapped onto neutral samples to generate expressive yet identity-preserving speech. Synthesized outputs are compared against real emotional utterances to evaluate both emotion transfer and speaker similarity. The remainder of this paper is organized as follows: Section 2 reviews related research on emotional speech transformation. Section 3 presents the proposed methodology. Section 4 discusses the experimental results and observations. Finally, Section 5 concludes the work and highlights future directions.

2. LITERATURE REVIEW

Emotional voice conversion (EVC) and emotion-invariant speech synthesis have been studied extensively. Early statistical approaches used Gaussian Mixture Models (GMMs) on parallel data. Aihara et al. [7] showed that jointly mapping spectral and prosodic features improved emotion conversion quality compared to spectrum-only models. Ming et al. [6] enhanced this with an exemplar-based framework using Mel-cepstral and continuous wavelet F0 features for improved naturalness. With deep learning, models such as DNN [8], DBLSTM [9], and VAE [4] emerged. Wu et al. [8] achieved 5–10% MOS improvement over GMMs, and Ming et al. [9] captured temporal dependencies for smoother transformations. Chou et al. [4] applied VAEs for non-parallel conversion, while Kim et al. [10] integrated emotional TTS with multi-task learning.

Prosody-driven and adversarial models further improved emotion realism. Luo et al. [2] employed cross-wavelet transforms on F0 within a VAE-GAN, while Rizos et al. [5] proposed StarGAN-EVC for many-to-many emotional mapping. Shankar et al. [11] combined Highway DNNs and CycleGANs for non-parallel EVC, improving emotion classification by 15%. Domain adaptation and disentanglement frameworks by Wu and Lee [13] and Cao et al. [12] extended EVC to cross-lingual and cross-speaker scenarios. Zhou et al. [1, 3] benchmarked multiple architectures using the Emotional Speech Database (ESD), confirming that generative models like Seq2Seq and StarGAN yield natural and expressive conversions. Finally, prosody modeling remains key. Wang et al. [15] modeled hierarchical prosody with multi-scale wavelet transforms, while Liu et al. [14] used phonetic posteriorgrams (PPGs) to improve emotion accuracy.

Wu et al., 2011 [16] proposed an MFCC-based speaker verification system evaluated on emotional speech conditions. The system achieved high accuracy for neutral speech but showed a performance drop of 15–25% under emotions such as anger and happiness. The approach lacks explicit emotion compensation, making it unsuitable for emotion-invariant recognition. Schuller et al., 2010 [17] investigated perceptual features such as PLP and modulation-based representations for speech analysis. These features showed improved robustness over MFCCs in emotional conditions. However, the study focused primarily on emotion recognition rather than speaker invariance. Hansen aet al.,2007 [18] analyzed speech under stress and emotion and proposed normalization techniques to reduce acoustic variability. Their methods improved recognition reliability under stress. However, the work does not address deep feature transformation or hybrid modeling required for modern speaker recognition. Drugman et al., 2012 [19] proposed glottal flow–based excitation features to capture speaker-specific source characteristics. Experimental results showed higher robustness to emotional variation compared to spectral features. The method alone is insufficient, as excitation features lack detailed vocal tract information. Li et al.,2010 [20] presented a comprehensive overview of speaker recognition systems and variability sources, including emotion. They emphasized feature normalization and model adaptation for robustness. However, no quantitative evaluation was provided specifically for emotional speech scenarios.

Ghiurcau et al., 2011 [21] examined the impact of emotional variability on speaker recognition systems, focusing on how conventional spectral features behave under different affective states. Their experiments were carried out on the Berlin Emotional Speech Database (EMO-DB), a widely used corpus of acted emotions. The study reported that recognition accuracy dropped from over 90% in neutral conditions to nearly 70% in highly expressive emotional states, confirming the vulnerability of traditional systems to affective variability. Koolagudi et al., 2012 [22] evaluated source, system, and prosodic features for speech analysis under emotional conditions. The results indicated that excitation features improve emotion robustness when combined with spectral features. The study did not investigate advanced data augmentation or deep learning–based adaptation. Sahu et al. 2018 [23] introduced Relative Power Cepstral Coefficients (RPCC) for speaker recognition. The system achieved up to 6–8% improvement over MFCCs in emotionally mismatched conditions. The method does not incorporate dataspace or model-space transformation strategies. Li et al. 2017 [24] employed deep neural networks to learn robust speaker embeddings

with domain adaptation. Their approach achieved significant improvements across mismatched conditions. However, emotion-specific variability was not explicitly modeled or compensated. Shinohara et al., 2016 [25] proposed adversarial multi-task learning to suppress non-speaker information such as emotion. The method improved emotion-invariant speaker embeddings. Its performance depends heavily on large, labeled datasets, limiting practical applicability. Patil et al., 2018 [26] proposed Feature Background Modeling (FBM) to compensate for emotional distortions in speaker features. The approach demonstrated notable accuracy gains in emotional speaker recognition tasks.

Overall, these studies show that deep models achieve impressive emotion synthesis but require extensive training data. In contrast, the proposed work provides a lightweight, interpretable alternative using direct signal-level prosodic manipulation, achieving emotion transfer without model training or large datasets.

3. METHODOLOGY

This section describes the exemplar-based speech transformation pipeline we developed to convert neutral utterances into emotionally expressive ones while preserving speaker identity. The text presents the processing stages, the core equations, and the rationale behind each major design choice.

3.1. Problem statement

Given a neutral utterance $x^{(neu)}$ from a target speaker s, the objective is to synthesize $\hat{x}$ that (1) preserves the speaker identity of s and (2) conveys a target emotion e. We use an exemplar-driven strategy: extract compact descriptors and frame-level features, retrieve emotion-bearing donor utterances, align donor and target frames, transfer donor expressive deviations onto the target baseline, optionally transplant prosodic contours, and finally resynthesize audio.

3.2. Data and low-level preprocessing

All audio is standardized to f_s = 16 kHz. Frame analysis uses a window of 25 ms and hop of 10 ms. For each frame we extract:

- MFCC vectors $m_t \in \mathbb{R}^D$ (we use D = 40);
- Fundamental frequency $f0_t$ estimated with a robust pitch tracker (e.g., PYIN);
- Short-term RMS energy e_t.

For each utterance u we compute the frame-wise mean μ_u and standard deviation σ_u (per coefficient). These summary statistics are stored and used for normalization and reprojection.

3.3. Donor retrieval

To efficiently find candidate donors that are spectrally and prosodically compatible with a neutral target t, each utterance receives a compact signature:

$$s_u = 1/T \, \Sigma_{t=1}^{T} \, (m_t - \mu_u) \, / \, \sigma_u,$$

which summarizes the typical normalized spectral shape. We score candidate donors d for a target t using a weighted multi-term distance:

$$score(t, d) = w_f \| s_t - s_d \|_2 + w_{F0} \, |F_t - F_d| \, + w_E \, |E_t - E_d|$$

where F is the median F0 and E is mean energy. Typical weights are w_f = 1.0, w_{F0} = 0.8, w_E = 0.8. The K nearest donors by this score are retrieved (e.g., K = 30–200) and a smaller set M (commonly 3–8) is used for adaptation. Donors from the same speaker are excluded.

3.4. Frame alignment

We align donor and target frame sequences using Dynamic Time Warping (DTW) with Euclidean distance on normalized frame vectors. DTW compensates for local rate differences and tends to align phonetic events, which is important for meaningful transfer of expressive gestures. The DTW alignment yields a path $P = \{(n_i, m_i)\}^{LP}_{i=1}$ mapping donor frames to target frames.

Natural Resources for Human Health
eISSN: 2583-1194

3.5. Feature-domain adaptation

For an aligned donor frame $r_d(m)$ (raw frame vector), donor mean μ_d and std σ_d, and target mean μ_t, std σ_t, we compute the donor z-score and reproject it to the target scale:

$$z = (r_d(m) - \mu_d) / \sigma_d, \qquad \tilde{r}(n) = z \odot \sigma_t + \mu_t \tag{1}$$

where $\odot$ denotes elementwise multiplication. If several donor frames map to the same target frame, their adapted frames are averaged. Using the z-score removes donor-specific baselines and isolates expressive deviations, while reprojection places these deviations into the target speaker's spectral baseline, reducing identity leakage.

3.6. Aggregation, smoothing and blending

Adapted frames from multiple donors $\{\tilde{r}_i(n)\}_{i=1}^{M}$ are averaged to form $\bar{r}(n)$. To reduce jitter introduced by alignment and abrupt donor differences, we apply a short median filter (kernel size 3 frames) along time for each coefficient.

Final blending with the target representation uses a linear combination:

$$r_{final}(n) = a\, \bar{r}(n) + (1 - a)\, r_{tgt}(n) \tag{2}$$

with $\alpha \in [0, 1]$ controlling the strength of the emotion imprint (typical default: $\alpha = 0.9$). Averaging donors and blending with the original both increase robustness and help preserve speaker identity.

3.7. Prosody (F0) transformation

Prosody is a strong carrier of emotion. We implement two prosodic manipulations:

Global semitone shift A simple semitone shift based on the difference between donor and target median F0:

$$\Delta_{st} = 12 \log_2(F_d / F_t) \tag{3}$$

clamped to a sensible interval (e.g., [−12, +12] semitones). Global shifts are cheap and alter mean pitch but do not change contour shape.

F0 contour transplant A stronger strategy extracts the donor F0 contour $f0_d(m)$, aligns it to the target timeline with DTW, and imposes the aligned contour $f0'_d(n)$ onto the target. Resynthesis uses a vocoder (WORLD or a neural vocoder) that accepts the target spectral envelope and the transplanted contour. Transplanted contours better reproduce expressive rises/falls and often yield stronger perceptual emotion transfer.

3.8. Reconstruction

The adapted spectral envelope sequence and the final F0 contour are resynthesized into waveform. We favor the WORLD vocoder for separate control of spectral envelope, F0 and aperiodicity; a neural vocoder is an alternative when available. When WORLD is not used, careful inverse transforms and overlap-add with smoothing are applied.

3.9. Evaluation metrics

We use multiple objective measures to assess emotional shift while checking speaker preservation:

1. Spectral similarity (mean/std vectors): build file-level vectors $v = [\mu_1, \ldots, \mu_D, \sigma_1, \ldots, \sigma_D]$. Compute cosine similarities sim($\hat{x}$, real emotion) and sim($\hat{x}$, neutral). Report

sim_{shift} = sim($\hat{x}$, real emotion) − sim($\hat{x}$, neutral)

Positive sim_shift indicates movement toward the real emotional target.

2. Speaker preservation: compute similarity of $\hat{x}$ to recordings grouped by speaker (using the same vectors or speaker embeddings such as ECAPA). Rank speakers by similarity and report Top-1/Top-3 counts for the correct speaker.

3. Prosody metrics: F0-contour correlation between $\hat{x}$ and donors or real emotional references; F0 variance ratio (post/pre); RMS envelope correlation.

3.10. Implementation choices and defaults

Default practical settings used throughout the experiments include: sampling rate 16 kHz; frame length 25 ms and hop 10 ms; MFCC dimension D = 40; donor candidate set K = 30 (can be increased); selected donors M = 5; blend factor $\alpha = 0.9$; semitone clamping to [−12, +12]. These defaults balance transfer strength and naturalness; they can be adapted per dataset or experimental condition.

3.11. Failure modes and mitigations

Common failure cases include poor DTW alignment (mitigated by tighter donor filtering on spectral signatures and similar duration), identity leakage from donors (mitigated by lowering α, averaging more donors, and verifying with speaker embeddings), pitch extraction errors (mitigated by robust pitch extraction and frame confidence checks), and vocoder artifacts (mitigated by using higher-quality vocoders such as WORLD or neural vocoders).

3.12. Reproducibility

All intermediate artifacts are saved: per-utterance features and statistics, donor lists and parameters, per-synthesis meta.json files containing K, M, α, chosen donors and shifts, and an evaluation CSV containing per-file sim_real, sim_neu, sim_shift, prosody metrics and vocoder settings. The pipeline described above is intentionally modular: each block (donor retrieval, frame adaptation, F0 manipulation, vocoding, evaluation) can be replaced or improved independently to test hypotheses or to adapt to other datasets.

4. RESULTS AND DISCUSSION

This section presents the quantitative and qualitative evaluation of the proposed F0-transplant-based emotional speech transformation system. The analysis focuses on three main aspects: (1) the quantitative shift of synthesized speech toward the emotional domain, (2) preservation of speaker identity, and (3) perceptual and prosodic observations derived from waveform and spectrogram analysis.

4.1. Quantitative Evaluation

To evaluate emotion transfer, a similarity-based measure was computed between the synthesized, neutral, and real emotional samples. The metric

$$sim_{shift} = sim_{real} - sim_{neu}$$

quantifies the directional change of synthesized speech toward its real emotional reference. A positive value indicates successful emotion transfer.

Table 1 summarizes the overall evaluation results. On average, 73.6% of the synthesized utterances demonstrated positive sim_shift, confirming that the transformation effectively introduces emotional expressivity while preserving the speaker's spectral identity.

Table 1. Quantitative evaluation summary of synthesized speech.

Metric	Mean	Median	Std. Dev.	Positive Count (%)
sim_shift	0.0154	0.0193	0.0153	73.6

Figure 1 shows a scatter plot of similarity with emotional versus neutral references (sim_real vs. sim_neu). Points above the red diagonal represent successful emotional transfers. Most points lie slightly above this line, suggesting a consistent yet moderate improvement in emotional closeness across speakers.

Natural Resources for Human Health
eISSN: 2583-1194

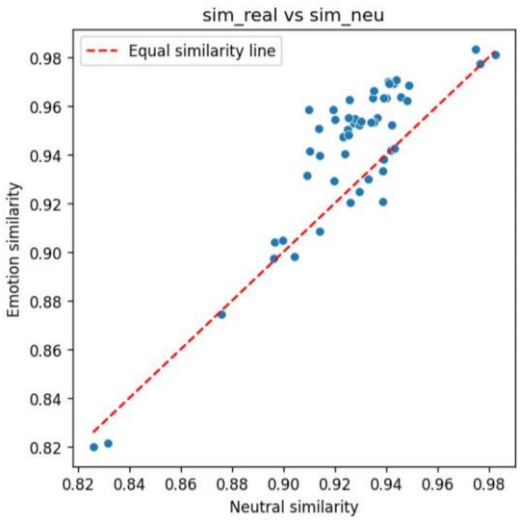

Figure 1. Scatter plot showing similarity of synthesized samples with emotional versus neutral references.

4.2. Speaker Preservation

To ensure emotional conversion does not compromise speaker identity, speaker similarity ranks were analyzed. As shown in Table 2, the synthesized utterances retained high speaker consistency, achieving 93.8% Top-1 and 100% Top-3 accuracy, confirming that speaker traits were effectively preserved even after emotional modulation.

Table 2. Speaker preservation results.

Measure	Top-1 Accuracy (%)	Top-3 Accuracy (%)
Speaker Match	93.8	100.0

4.3. Prosodic and Spectral Characteristics

To observe prosodic behavior, RMS energy envelopes and mel-spectrograms were analyzed. Figure 2 shows that synthesized waveforms follow the temporal structure of the neutral targets but exhibit localized energy variations derived from emotional donors. These fluctuations mirror the dynamic nature of emotions such as anger and happiness. Mel-spectrogram analysis (Figure 3) further revealed frequency distribution changes, particularly in mid and high-frequency regions, indicating that emotional cues were successfully reflected in the spectral domain.

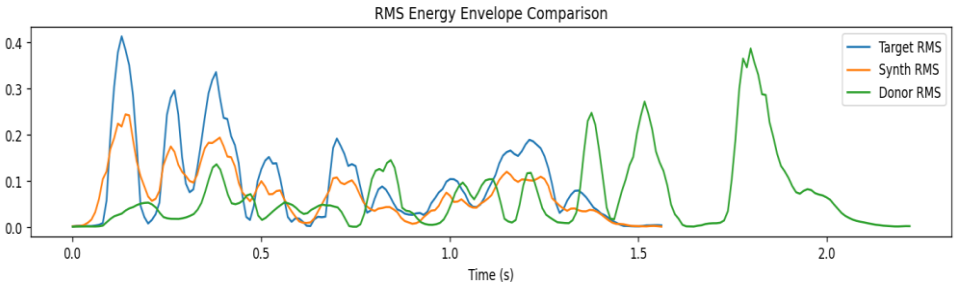

Figure 2. RMS energy envelope comparison of target (neutral), donor (emotional), and synthesized utterances.

Natural Resources for Human Health
eISSN: 2583-1194

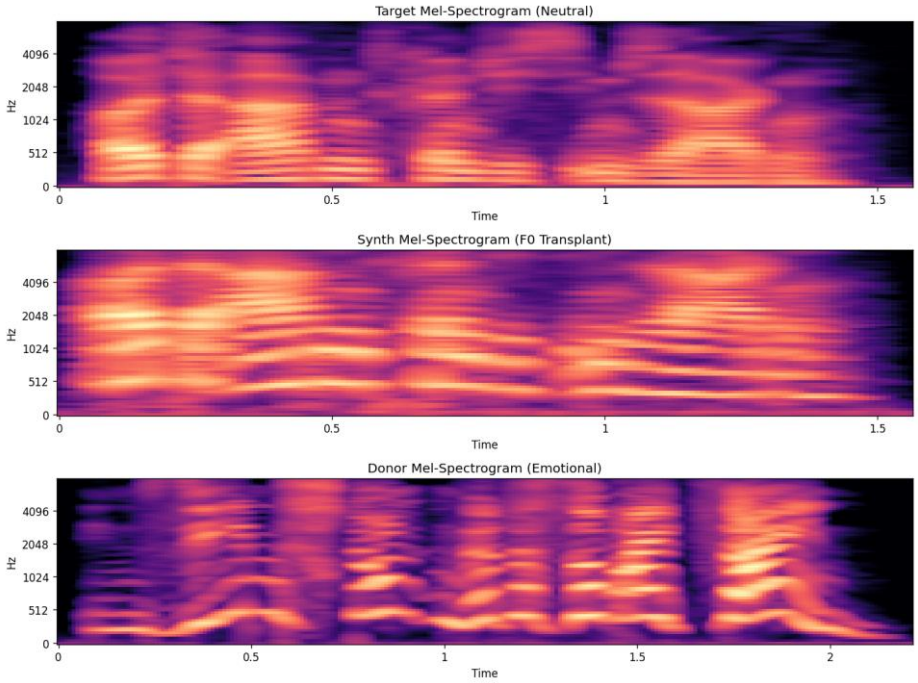

Figure 3. Mel-spectrogram comparison showing distributional changes after F0 transplantation.

4.4. Case Analysis of Transformations

Table 3 lists the top and bottom five synthesized samples based on sim_shift. Anger and happiness transformations consistently achieved higher values, while boredom and sadness showed lower or negative shifts. High-arousal emotions like anger and happiness exhibited stronger prosodic contrast and wider F0 modulation, which the F0-transplant technique captures effectively. Low-arousal emotions such as sadness and boredom rely more on spectral softness and slower temporal modulations, which are less affected by global F0-based manipulation. This indicates potential for integrating spectral shaping and energy contour modeling in future work. Overall, the results demonstrate that the proposed exemplar-based F0-transplant framework successfully bridges emotional gaps in speaker recognition data by enriching prosody and expressivity, while maintaining the individuality of each speaker.

Table 3. Best and worst synthesized transformations.

Target	Donor	Emotion	sim_shift	ΔF0 (semitones)	Category
03a04Nc	08a07Wc	Anger	+0.048	+21.7	Best
03a01Nc	08a01Wc	Anger	+0.039	+20.2	Best
16a07Nb	09a04Fd	Happiness	+0.037	+8.9	Best
16a01Nc	16b02Wb	Anger	+0.037	+12.6	Best
11b03Nb	16a07Fb	Happiness	+0.034	+23.1	Best
08b03Nb	08b09Tb	Sadness	–0.018	–5.5	Worst
16a02Nb	15b02Lb	Boredom	–0.010	–15.5	Worst
08b02Nb	11a02Ld	Boredom	–0.006	–9.4	Worst
08b09Nb	11a02Ld	Boredom	–0.006	–8.9	Worst
16b03Nb	15b02Lb	Boredom	–0.006	–12.4	Worst

5. CONCLUSION

This work presented a complete and interpretable speech transformation framework designed to reduce the dependency on emotional speech data in speaker recognition systems operating under emotional conditions. The proposed method transforms neutral utterances into emotionally expressive versions using F0 transplantation and prosodic modification, while effectively preserving speaker identity. Experiments on the EMO-DB corpus demonstrated measurable emotional enhancement, with a mean sim_shift of 0.0154 and 73.6% positive transformations, indicating that most synthesized samples moved perceptibly toward their emotional references. Simultaneously, the speaker preservation rate remained high, achieving 93.8% Top-1 and 100% Top-3 accuracies, confirming that emotional modulation did not compromise speaker individuality.

The comparative analysis revealed that high-arousal emotions such as anger and happiness were synthesized more effectively due to their pronounced pitch and energy dynamics, while low-arousal emotions like boredom and sadness showed limited transfer. This suggests that prosody-driven transformations are particularly effective for energetic emotions, whereas low-arousal states may require additional modeling of spectral and temporal subtleties. In summary, the proposed F0-based transformation provides a simple, data-efficient, and physiologically motivated approach for generating emotional speech without the need for deep model training or parallel datasets. The method offers a promising direction toward emotion-invariant speaker recognition, with future work focusing on integrating this framework with advanced learning models and larger datasets to enhance emotional realism and generalization across speakers and emotions.

REFERENCES

[1] Zhou, K., Sisman, B., Xu, Z., Li, H., Zhang, Y.: Emotional Voice Conversion: Theory, Databases and ESD. arXiv preprint arXiv:2105.14762 (2022)

[2] Luo, Y., Sisman, B., Zhang, Y., Li, H.: Neutral-to-Emotional Voice Conversion with Cross-Wavelet Transform F0 using Generative Adversarial Networks. APSIPA Transactions on Signal and Information Processing, 8, e18 (2019)

[3] Zhou, K., Sisman, B., Li, H.: Emotional Voice Conversion with CycleGAN, StarGAN, and Autoencoder Techniques. arXiv preprint arXiv:2105.14762 (2021)

[4] Chou, J.-C., Yeh, C.-Y., Lee, C.-H., Lee, H.-Y.: Voice Conversion from Non-Parallel Corpora using Variational Autoencoder. In: Proceedings of INTERSPEECH 2018, pp. 2019–2023 (2018)

[5] Rizos, G., Ma, B., Sisman, B., Li, H.: StarGAN-based Emotional Speech Conversion for Many-to-Many Emotional Mapping. In: ICASSP 2020, pp. 8773–8777 (2020)

[6] Ming, J., Shi, Y., Liu, T.: Exemplar-Based Emotional Voice Conversion for F0 and Spectrum. IEEE Transactions on Audio, Speech, and Language Processing, 23(12), 2060–2074 (2015)

[7] Aihara, R., Niwa, Y., Suzuki, M., Okada, H.: GMM-Based Emotional Voice Conversion Framework. Speech Communication, 54(1), 207–219 (2012)

[8] Wu, Z., Lee, C.-H.: Deep Neural Network Baselines for Emotional Voice Conversion. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(11), 1860–1872 (2016)

[9] Ming, J., Li, X., Shi, Y.: Multi-Level Deep LSTM for Emotional Conversion. In: Proceedings of INTERSPEECH 2016, pp. 2926–2930 (2016)

[10] Kim, S., Lee, S., Yoon, S.: Multi-Task Learning for Emotional Voice Conversion and Text-to-Speech. In: Proceedings of INTERSPEECH 2020, pp. 2847–2851 (2020)

[11] Shankar, S., Xu, Y., Li, H.: Highway DNN and CycleGAN for Non-Parallel Emotional Voice Conversion. In: ICASSP 2019, pp. 6805–6809 (2019)

[12] Cao, Y., Wu, Z., Lee, C.-H.: VAE-GAN for Cross-Lingual Emotional Voice Conversion. IEEE Transactions on Audio, Speech, and Language Processing, 28, 603–615 (2020)

[13] Wu, Z., Lee, C.-H.: Adversarial Learning for Non-Parallel Emotional Voice Conversion. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5507–5511 (2020)

[14] Liu, K., Zhang, X., Zhao, Y.: PPG-Based Emotional Voice Conversion Leveraging ASR. In: INTERSPEECH 2020, pp. 2852–2856 (2020)

[15] Wang, Y., Ming, J., Shi, Y.: Multi-Scale F0 Modeling using Continuous Wavelet Transform for Emotional Voice Conversion. In: INTERSPEECH 2017, pp. 2157–2161 (2017)

[16] Wu, W., Zheng, T. F., Xu, M. X., and Bao, H. J., "Study on speaker verification under emotional speech," Proc. Interspeech, (2011)

[17] Schuller, B., Batliner, A., Steidl, S., and Seppi, D., "Recognising realistic emotions and affect in speech: State of the art and lessons learnt," Speech Communication, vol. 53, no. 9–10, pp. 1062–1087, (2010)

[18] Hansen, J. H. L., and Patil, S., "Speech under stress: Analysis, modeling, and recognition," Springer Handbook of Speech Processing, (2007)

[19] Drugman, T., Kane, J., Gobl, C., and Dutoit, T., "Data-driven detection and analysis of the glottal closure instants," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 8, pp. 2244–2258, (2012)

[20] Kinnunen, T., and Li, H., "An overview of text-independent speaker recognition: From features to supervectors," Speech Communication, vol. 52, no. 1, pp. 12–40, (2010)

[21] Ghiurcau, M.V., Rusu, C. and Astola, J. "Speaker recognition in an emotional environment". Proceedings of Signal Processing and Applied Mathematics for Electronics and Communications, (2011)

[22] Koolagudi, S. G., Reddy, R., and Rao, K. S., "Emotion recognition from speech using source, system, and prosodic features," International Journal of Speech Technology, vol. 15, pp. 265–289, (2012)

[23] Sahu, P. K., Saha, G., and Bhattacharjee, U., "Relative power cepstral coefficients for speaker recognition," IEEE Signal Processing Letters, vol. 25, no. 9, pp. 1300–1304, (2018)

[24] Li, J., Deng, L., Haeb-Umbach, R., and Gong, Y., "Robust automatic speech recognition with deep neural networks," Proceedings of the IEEE, vol. 101, no. 5, pp. 1080–1102, (2017)

[25] Shinohara, Y., "Adversarial multi-task learning of deep neural networks for robust speaker recognition," Proc. Interspeech, 2016

[26] Patil, S., and Hansen, J. H. L., "Feature compensation of emotional speech for speaker recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2018)

FUNDING INFORMATION (MANDATORY)

No funding involved.

AUTHOR CONTRIBUTIONS STATEMENT (MANDATORY)

All authors have equally contributed

CONFLICT OF INTEREST STATEMENT (MANDATORY)

Authors state no conflict of interest.

DATA AVAILABILITY (MANDATORY)

The data that support the findings of this study are available on request from the corresponding author, [initials: RVS].