

A Hybrid Algorithm for Robust Pitch Estimation in Emotional Speech Synthesis

Hammadi Zineb

Department of Measurement and Information System
Budapest University of Technology and Economics
Budapest, Hungary
hammadi.zineb@edu.bme.hu

Mohammad Salah Al-Radhi

Department of Telecommunications and Artificial Intelligence
Budapest University of Technology and Economics
Budapest, Hungary
malradhi@tmit.bme.hu

Abstract—Emotional intelligence in synthetic speech remains a critical challenge in human-machine interaction, despite significant advances in the naturalness and intelligibility of speech synthesis. Current systems struggle to accurately capture the nuanced emotional expressions characteristic of human speech, including rapid pitch transitions, wide frequency variations, and irregular vibrato patterns. While pitch estimation algorithms like PESTO and FCPE have proven effective for standard speech, their performance on emotionally expressive content remains largely unexplored. In this paper, we present ESCAPE, a novel algorithm specifically designed for emotional speech processing. ESCAPE combines PESTO's precision in handling frequency variations with FCPE's context-aware processing through a hybrid architecture, enabling robust pitch tracking in expressive vocal content. Our approach excels at capturing complex acoustic patterns unique to emotional utterances while maintaining computational efficiency. We provide the first comprehensive evaluation of PESTO and FCPE on emotional speech datasets and demonstrate ESCAPE's transformative potential for pitch estimation in emotionally expressive speech synthesis. The results showcase significant progress in bridging the gap between human-like emotional expression and machine-generated speech, marking a pivotal advancement in emotional speech synthesis technology.

Keywords—*Emotional Speech Synthesis, Pitch Estimation, Self-Supervised Learning, Context-Aware Algorithms, Expressive Speech Processing.*

I. INTRODUCTION

Humans can experience multiple emotional states simultaneously. Consider bittersweet moments, such as recalling a fond memory of a lost love or the first time leaving home for college. These scenarios highlight the co-occurrence of different emotions, even those with opposite valences like joy and sadness. Emotional speech synthesis seeks to incorporate such emotional effects into synthesized speech. The ability to synthesize mixed emotions represents a significant step toward achieving human-like emotional intelligence in speech synthesis, enabling more natural and effective human-machine interactions.

Speech synthesis aims to generate human-like voices from input text. With the advent of deep learning, modern speech synthesis systems have achieved remarkable improvements in naturalness and intelligibility. However, these systems often fail to convey the emotional nuances present in human-human interactions. This lack of expressiveness limits their emotional intelligence, which is essential for applications requiring empathetic and context-aware communication. Emotional

speech synthesis endeavors to address this gap by incorporating emotional contexts into synthetic speech [1] [2].

The pitch signal, also known as the glottal waveform, plays a critical role in conveying emotion. It reflects the tension of the vocal folds and subglottal air pressure, making it a key acoustic feature for emotional expression [3]. While pitch estimation algorithms like PESTO and FCPE have shown promising results on standard speech and music datasets, their performance on emotional or expressive speech remains underexplored. This is a significant gap, as emotional speech poses unique challenges, including highly variable fundamental frequencies, rapid pitch transitions, irregular vibrato patterns, and extreme pitch ranges. These acoustic characteristics often deviate from the patterns observed in neutral speech, making it difficult for existing algorithms to accurately track the fundamental frequency (F0) in such contexts [4] [5].

In this paper, we extend the evaluation of PESTO and FCPE to emotional speech datasets, offering the first comprehensive analysis of their performance on expressive vocal content. Additionally, we introduce ESCAPE (Emotion Self-Supervised Context-Aware Pitch Estimation), a novel algorithm specifically designed to address the dynamic nature of emotional speech. ESCAPE incorporates adaptive processing techniques to account for the wider frequency variations and rapid modulations typical of emotional expressions, achieving robust F0 tracking in expressive speech content.

A. Problem statement

PESTO and FCPE were not originally tested or validated on emotional speech datasets, which presents a significant limitation. Expressive speech exhibits unique characteristics—such as rapid pitch transitions, wide frequency variations, and irregular vibrato patterns—that pose considerable challenges for pitch detection. These algorithms were primarily optimized for the more predictable patterns of standard speech, limiting their applicability to the complexities of emotional speech.

II. RELATED WORK AND BACKGROUND

A. PESTO

Pitch Estimation with Self-Supervised Transposition-Equivariant Objective (PESTO) is a pitch estimation algorithm designed to address the challenges of monophonic audio analysis, offering a computationally efficient solution with minimal overhead. It uses a Siamese neural network architecture that processes pairs of audio signals that have

been pitch-shifted, with these signals represented using the Constant-Q Transform (CQT) [6]. The core innovation in PESTO is its class-based transposition-equivariant objective, which is specifically designed to prevent model collapse in an encoder-only setting, allowing for robust pitch estimation without relying on labeled data.

A key feature of PESTO is its use of the CQT [7], as shown in Figure 1, which transforms a signal by employing a filter bank with bandwidths that vary in proportion to the center frequency of each filter. This design ensures that the Q factor (the ratio of center frequency to bandwidth) remains constant across all filters, enabling precise frequency representation across the audio spectrum. The analysis process involves convolving the signal with a kernel centered at a specific frequency, and this process is repeated for all desired frequencies, producing a time-frequency representation where each frequency bin corresponds to a logarithmically spaced center frequency. This approach allows PESTO to capture fine-grained details in the audio signal that are critical for pitch estimation, especially when dealing with varying frequencies in speech. Another fundamental aspect of PESTO is its ability to enforce equivariance to pitch transpositions, meaning that the model learns to align the pitch probability distributions between pitch-shifted versions of the input. This transposition-equivariant objective ensures that shifts in input lead to corresponding shifts in output, a feature essential for reliable pitch tracking in dynamic audio contexts.

To ensure that this equivariance is preserved across the network, PESTO integrates Toeplitz matrices in its final fully connected layer. These matrices are parameter-efficient and crucial for maintaining the relationship between transposed inputs and their corresponding outputs, minimizing the computational cost while ensuring accurate pitch estimation.

In terms of regularization, PESTO utilizes a variety of loss functions to refine its learning process and improve performance. The Equivariance Loss ensures that shifted inputs result in appropriately shifted output distributions, while the Shifted Cross-Entropy (SCE) loss function aligns the output distributions of pitch-shifted pairs, further preventing model collapse. Additionally, Invariance Loss encourages robustness to pitch-preserving transformations, such as noise and gain adjustments, which are common in real-world audio data. With fewer than 30,000 parameters, PESTO is significantly smaller than many traditional deep learning models, making it particularly suitable for real-time applications and deployment on devices with limited computational resources. This lightweight architecture makes PESTO an attractive choice for applications in mobile devices or other resource-constrained environments, where real-time pitch estimation is required without sacrificing performance [4].

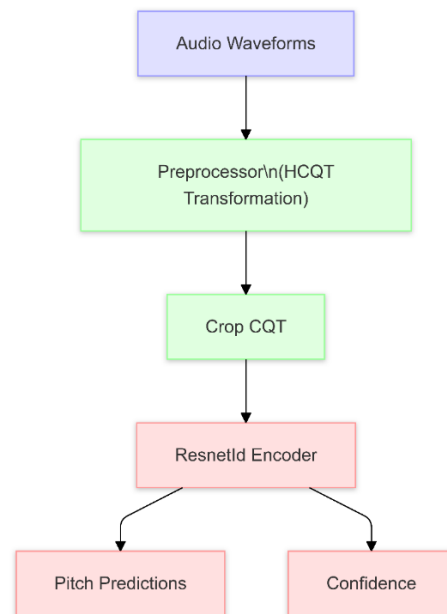


Fig. 1. Diagram representing PESTO model architecture

B. FCPE

Fast Context-based Pitch Estimation (FCPE) is a lightweight and flexible pitch estimation model designed to efficiently process Mel-spectrogram inputs and extract fundamental frequency (F0) information. Developed using PyTorch, FCPE incorporates innovative architectural features and decoding mechanisms tailored for fast and accurate pitch estimation [5].

A central component of FCPE is its Conformer-based Encoder, as shown in Figure 2, which processes Mel-spectrogram inputs using a hybrid architecture that combines convolutional layers with attention mechanisms. This design allows the model to effectively capture both local and global context, making it particularly suited for pitch estimation in audio data that involves complex and variable patterns. FCPE also optionally employs Harmonic Embeddings, which encode harmonic relationships within the audio signal. This feature enhances the model's ability to refine pitch predictions by incorporating higher-level acoustic patterns. The harmonic embeddings are configurable during training, allowing for adaptation to different datasets and applications. For input processing, FCPE uses a stack of convolutional and normalization layers to extract robust feature representations from the mel-spectrogram input. A group normalization step further enhances stability during training, followed by LeakyReLU activation to introduce non-linearity into the model's processing. After extracting features, FCPE employs an output projection and post-processing mechanism. This involves applying a LayerNorm operation to stabilize the output before passing it through a weight-normalized fully connected layer. The resulting latent pitch distribution is then activated using a sigmoid function, producing interpretable probabilities that represent potential pitch values.

FCPE supports multiple decoding mechanisms to convert the latent representations into actual pitch values, either in cents or Hertz. These mechanisms include:

- **Latent-to-Cents Decoding:** This approach converts latent distributions to cent values using global or

local argmax strategies, selecting the most likely pitch estimates based on the latent representation.

- **Gaussian Masking:** To ensure smoother latent-to-cent transformations, FCPE applies Gaussian blurring, which helps to refine the predictions and reduce noise.
- **Cents-to-F0 Conversion:** After decoding to cent values, FCPE translates these values into F0, enabling applications such as MIDI conversion and other pitch-related tasks [5].

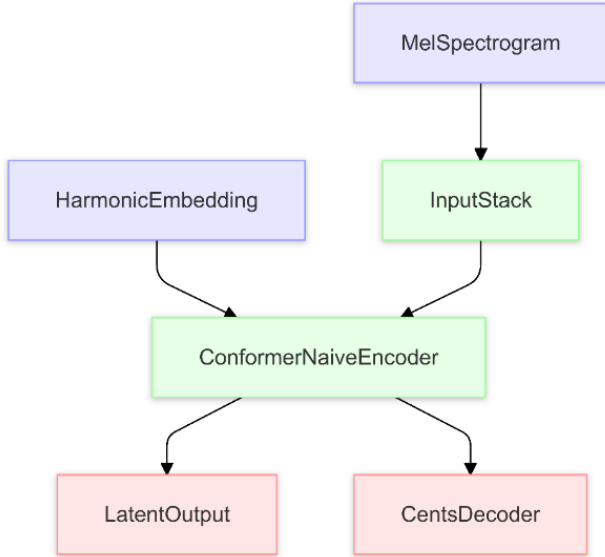


Fig. 2. Diagram representing FCPE model architecture

C. Data Fusion through Concatenation in High-Dimensional Representations

Concatenation is a fundamental operation in data manipulation, particularly in scenarios where multiple data structures (such as arrays, matrices, or sequences) need to be combined into a unified entity. This process aligns the input data along a specified axis or dimension, preserving the order of the original elements [8]. Concatenation is commonly used in various machine learning and signal processing tasks to merge feature sets, combine temporal sequences, or integrate multi-channel information.

Given two matrices A and B , each with shape $(m \times n)$, concatenation along axis 1 (i.e., the horizontal axis) results in a new matrix C with dimensions $(m \times 2n)$, as shown in equation (1):

$$C = [A \mid B] \quad (1)$$

Here, m represents the number of rows, while n refers to the number of columns in each input matrix. The concatenated matrix C has $2n$ columns, which is the sum of the original column counts from A and B . For example, if A and B each have 3 columns, the resulting matrix will have 6 columns.

Key Considerations in Data Fusion:

- **Preservation of Order:** The concatenation operation ensures that the sequence of data remains unchanged, meaning that the internal structure and relationships within the individual arrays or sequences are preserved.

- **Customizable Axis Selection:**

In more complex data structures, such as multidimensional arrays (e.g., NumPy arrays or pandas DataFrames), concatenation can occur along any axis. Typically, concatenation along:

- **axis = 0** corresponds to vertical stacking (i.e., adding rows).
- **axis = 1** corresponds to horizontal merging (i.e., adding columns).

- **Shape Consistency:**

For concatenation to be valid, the input arrays must align along the non-concatenating axes. In practice, this means that the number of rows in the arrays being concatenated must be identical when merging along axis 1, and the number of columns must be the same when merging along axis 0.

Concatenation is thus a versatile tool in data processing, enabling the integration of diverse datasets while ensuring that their inherent structures remain intact, crucial for subsequent analyses or model training.

III. METHODOLOGY

A. INTRODUCING ESCAPE

ESCAPE (Emotion Self-Supervised Context-Aware Pitch Estimation) represents a novel approach to pitch estimation that combines the strengths of PESTO and FCPE while addressing the unique challenges of emotional speech. ESCAPE was designed specifically to improve pitch tracking in expressive speech, where fundamental frequency (F0) variations can be rapid, irregular, and complex due to emotional expression. Unlike existing pitch estimation models, which may struggle with the nuanced pitch dynamics of emotional speech, ESCAPE introduces a hybrid architecture that blends the detailed pitch contour extraction capabilities of PESTO with the real-time processing efficiency and context-awareness of FCPE. The integration of these two models allows ESCAPE to more accurately track pitch across a wide range of emotional speech signals.

The core innovation of ESCAPE lies in its ability to dynamically adjust to the frequency fluctuations and complex patterns typical of emotional speech, while maintaining the computational efficiency required for real-time processing. The model achieves this by incorporating the following key features:

- **Frequency Adaptation:** ESCAPE leverages PESTO's advanced signal processing techniques to handle rapid frequency shifts characteristic of emotional speech.
- **Contextual Awareness:** Drawing from FCPE, ESCAPE incorporates a context-aware processing approach that accounts for both local and global temporal dependencies in the pitch contour.
- **Hybrid Architecture:** By combining the two models, ESCAPE not only improves pitch estimation accuracy but also preserves the computational efficiency necessary for practical deployment.

The resulting system provides robust pitch tracking that can capture the dynamic nature of emotional speech, thereby

offering significant improvements over existing methods in both accuracy and real-time performance.

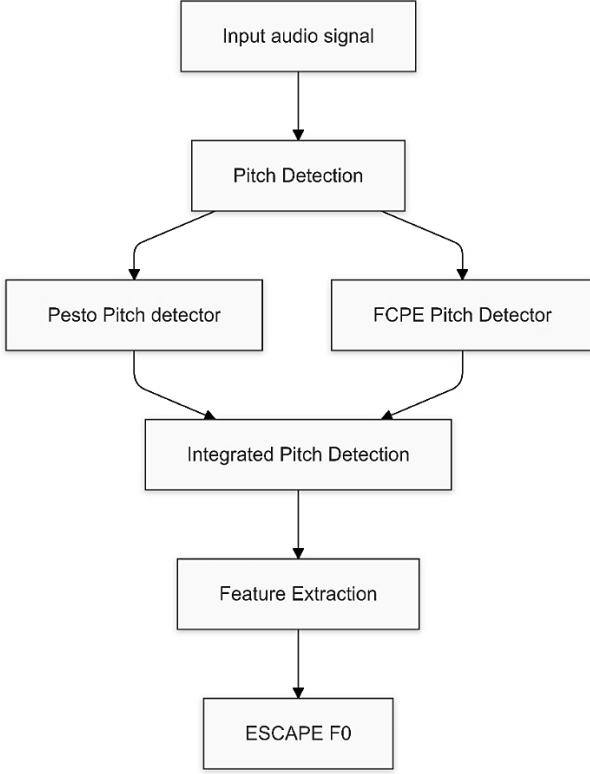


Fig. 3. Diagram representing ESCAPE

B. ESCAPE Architecture

The architecture of ESCAPE in Figure 3 is designed to effectively combine the strengths of PESTO and FCPE, utilizing their complementary features to address the challenges of pitch estimation in emotional speech. ESCAPE integrates both models in a hybrid framework, which processes the outputs of PESTO and FCPE to produce a unified pitch estimate.

- **Feature extraction:** The first step in the ESCAPE pipeline is the integration of outputs from both PESTO and FCPE. This is achieved through a concatenation of the pitch estimates from both models, allowing ESCAPE to leverage the complementary advantages of each. The concatenation is followed by a fusion step, where the two estimates are averaged to produce a final pitch estimate (ESCAPE_f0). This fusion allows ESCAPE to benefit from the fine-grained pitch tracking of PESTO and the context-aware processing of FCPE.

Mathematically, this can be expressed as:

$$ESCAPE_{F0} = 0.5 \times PESTO_{F0} + 0.5 \times FCPE_{F0} \quad (2)$$

This averaging strategy ensures that both pitch estimates contribute equally to the final output, providing a balanced solution for handling the complexity of emotional speech.

- **Pitch Tracking and Context Awareness:** ESCAPE’s architecture is designed to be highly sensitive to the rapid and irregular pitch variations typical of emotional speech. By combining the

techniques from PESTO’s transposition-equivariant objective and FCPE’s context-based processing, ESCAPE can track pitch more accurately in emotional contexts. The model’s ability to handle both fast pitch changes and long-range dependencies within the pitch contour makes it especially suited for expressive speech scenarios.

- **Self-Supervised Training:** ESCAPE is trained in a self-supervised manner, meaning that it does not require explicit labeling of pitch data. Instead, it learns to predict pitch through the use of a self-supervised objective function, which encourages the model to align pitch distributions across different emotional contexts. This training approach enables ESCAPE to generalize well to a wide range of emotional speech data, even with limited labeled data available.

C. Implementation Details

The implementation of ESCAPE leverages the strengths of both PESTO and FCPE, with each model contributing its own set of capabilities to the final pitch estimation system.

- **PESTO:** PESTO’s preprocessing and feature extraction techniques are used to capture the fine-grained frequency variations characteristic of emotional speech. The Constant-Q Transform (CQT) [6] is applied to raw audio, followed by a Toeplitz linear layer and a 1D ResNet encoder to extract pitch-related features. These features are then processed into pitch probabilities, which are used as one of the inputs to ESCAPE.
- **FCPE:** FCPE contributes its efficient processing capabilities, using a Conformer-based encoder to handle mel-spectrogram inputs. This model captures both local and global dependencies in the pitch contour, and optional harmonic embeddings improve pitch prediction in complex audio scenarios.
- **Fusion Mechanism:** The final pitch estimate, ESCAPE_f0, is generated by concatenating the outputs from PESTO and FCPE and averaging them. This fusion process ensures that ESCAPE benefits from the strengths of both models while maintaining a computationally efficient design suitable for real-time applications.

ESCAPE is implemented in PyTorch, ensuring efficient tensor operations and flexible deep learning functionality suitable for audio processing tasks. The implementation consists of three main components:

1. **Preprocessing:** The raw audio is processed using a Constant-Q Transform and feature extraction layers to prepare it for pitch analysis.
2. **Model Architecture:** The hybrid architecture combines the outputs of PESTO and FCPE, followed by the fusion step to generate the final pitch estimate.
3. **Post-Processing:** The pitch estimates are refined and converted to Hertz for practical applications.

IV. EVALUATION RESULTS

A. Data

To evaluate the performance of the pitch estimation algorithms, we used the JL-Corpus as the testing dataset. This emotional speech corpus was chosen due to its balanced and unique design. Unlike many existing corpora, the JL-Corpus includes an equal distribution of four long vowels in New Zealand English, ensuring an unbiased analysis of emotion-related formant and glottal source features. The speech signal was sampled at 44.1kHz and stored as 16-bit numbers. In total, there are 4 speakers, 5 primary emotions, 2 repetitions, 15 sentences and 2 sessions making 1200 primary emotion sentences and 4 speakers, 5 secondary emotions, 2 repetitions, 13 emotion neutral sentences, 2 emotion salient sentences and 2 sessions making 1200 secondary emotion sentences, making a total of 2400 sentences [11].

B. Objective metric

Pitch contour visualizations reveal distinct differences in tracking performance across the three algorithms. As shown in Figure 4 and Figure 5, ESCAPE demonstrates superior pitch tracking capabilities, producing smoother and more consistent pitch contours that better capture the nuanced variations in emotional speech. While both PESTO and FCPE show reasonable tracking abilities, their pitch trajectories exhibit more irregularities and potential tracking errors, particularly during rapid frequency changes characteristic of emotional expressions. ESCAPE's pitch contours appear more naturalistic and stable, with fewer abrupt jumps or discontinuities, suggesting more reliable tracking of the fundamental frequency. This improved performance can be attributed to ESCAPE's hybrid approach, which effectively combines PESTO's ability to handle frequency variations with FCPE's contextual awareness. The visualization particularly highlights ESCAPE's enhanced stability during challenging segments, where emotional speech exhibits extreme pitch modulations [12], confirming its more robust performance in tracking the complex pitch patterns found in expressive speech.

C. Evaluation Metrics

We used two evaluation metrics: RMSE (Root Mean Square Error) and Gross Pitch Error (GPE) to assess ESCAPE's performance.

1. RMSE

The root-mean-square error (RMSE) is calculated as follows [9]:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (f_0^{pred}(i) - f_0^{gt}(i))^2} \quad (3)$$

where $f_0^{pred}(i)$ and $f_0^{gt}(i)$ are the predicted and ground-truth pitch values at frame i , respectively.

2. GPE (Gross Pitch Error)

The GPE counts the frames where the predicted f_0 deviates by more than a certain threshold (e.g., 20%) from the ground-truth f_0 . It is calculated using the following equation [10]:

$$GPE = \frac{1}{N} \sum_{i=1}^N 1(|f_0^{pred}(i) - f_0^{gt}(i)| > \alpha \cdot f_0^{gt}(i)) \quad (4)$$

where $1(\cdot)$ is the indicator function, and α is the threshold (e.g., 0.2 for 20%).

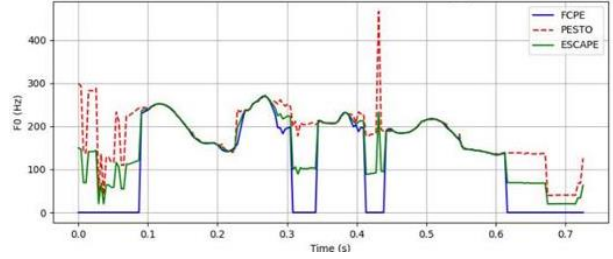


Fig. 4. FCPE vs. PESTO vs. ESCAPE for an angry female WAV file

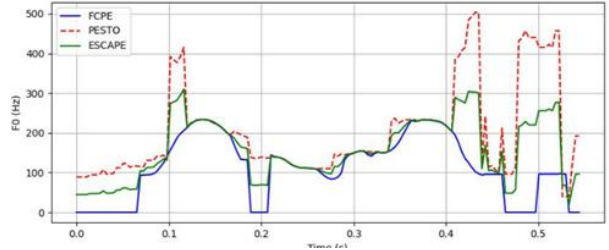


Fig. 5. FCPE vs. PESTO vs. ESCAPE for a sad male WAV file

TABLE 01: COMPARISON OF PESTO, FCPE, AND ESCAPE IN TERMS OF GPE AND RMSE

Metric	PESTO	FCPE	ESCAPE
GPE	0.8759	0.9241	0.6692
RMSE	1.6110	1.5281	1.4541

The results are statistical significance tested. In Table 1 we demonstrate that ESCAPE, the proposed hybrid approach, outperforms both PESTO and FCPE across both metrics. In terms of Gross Pitch Error (GPE), ESCAPE achieves a significantly lower error rate of 0.6692, compared to PESTO's 0.8759 and FCPE's 0.9241. This substantial improvement in GPE indicates that ESCAPE produces fewer significant pitch estimation errors, making it more reliable for emotional speech analysis. Regarding RMSE, ESCAPE also shows superior performance with 1.4541, compared to FCPE's 1.5281 and PESTO's 1.6110. The lower RMSE suggests that ESCAPE's pitch estimates are more accurate and closer to the ground truth values. These results validate the effectiveness of combining PESTO's ability to handle frequency variations with FCPE's contextual processing capabilities in the ESCAPE algorithm, resulting in more robust pitch estimation for emotional speech.

V. CONCLUSION AND DISCUSSION

In conclusion, pitch estimation in speech and audio processing is a complex yet crucial area of research with far-reaching applications in speech synthesis, musical signal processing, and emotion detection. This paper addressed the challenges of pitch estimation in emotional speech synthesis, a domain where conventional methods often fall short due to the intricacies of expressive vocal content. Emotional utterances introduce a layer of variability and rapid pitch modulation that existing algorithms like PESTO and FCPE, despite their effectiveness on standard datasets, struggle to capture accurately. In response to these challenges, we

introduced ESCAPE, a specialized pitch estimation algorithm that leverages self-supervised and context-aware mechanisms to handle the diverse acoustic features of emotional speech. ESCAPE demonstrates enhanced resilience to the dynamic range and modulation patterns inherent in expressive speech, representing a significant advancement toward achieving human-like emotional intelligence in synthesized speech.

For future work, we plan to further improve ESCAPE's efficiency for speech synthesis applications [13] [14]. This will involve training pitch estimation models specifically on emotional speech datasets and refining their architectures to better address the nuanced and dynamic variations that characterize emotional speech. Additionally, we aim to evaluate these algorithms across diverse languages and gender variations, as these factors introduce distinct pitch characteristics. A comparative analysis will help identify potential biases and highlight areas for improvement in pitch accuracy across demographic and linguistic groups. Finally, we intend to develop a multipitch detection algorithm optimized for tracking multiple voices or sounds simultaneously in complex environments.

VI. ACKNOWLEDGEMENTS

This paper is supported by the European Union's HORIZON Research and Innovation Programme under grant agreement No 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI) and by the Ministry of Innovation and Culture and the National Research, Development and Innovation Office of Hungary within the framework of the National Laboratory of Artificial Intelligence. M.S.Al-Radhi's research was supported by the EKÖP-24-4-II-BME-197, through the National Research, Development and Innovation (NKFI) Fund.

REFERENCE

- [1] M.S. Al-Radhi, T.G. Csapó, and G. Németh, "Noise and acoustic modeling with waveform generator in text-to-speech and neutral speech conversion," *Multimed. Tools Appl.*, vol. 80, no. 2, pp. 1969–1994, Jan. 2021.
- [2] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "EmoMix: Emotion Mixing via Diffusion Models for Emotional Speech Synthesis," *Ping An Technology (Shenzhen) Co., Ltd., Shenzhen, China; University of Science and Technology of China, Hefei, China. Submitted to INTERSPEECH*, Jun. 1, 2023.
- [3] A. Kroon, "Comparing Conventional Pitch Detection Algorithms with a Neural Network Approach," *ECSE 523 Speech Communications Final Project*, Dept. of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada, 29 Jun 2022.
- [4] A. Riou, S. Lattner, G. Hadjeres, and G. Peeters, "PESTO: Pitch Estimation with Self-Supervised Transposition-Equivariant Objective," *LTCI, Télécom-Paris, Institut Polytechnique de Paris, France; Sony Computer Science Laboratories, Paris, France; Sony AI*, 5 Sep 2023.
- [5] CNChTu, "FCPE," [Online]. Available: <https://github.com/CNChTu/FCPE.git>.
- [6] C. Schorkhuber and A. Klapuri, "Constant-Q Transform Toolbox for Music Processing" *Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz. Centre for Digital Music, Queen Mary University of London*.
- [7] Sony CSL Paris, "PESTO," [Online]. Available: <https://github.com/SonyCSLParis/pesto.git>.
- [8] NumPy Developers, "numpy.concatenate — NumPy v2.1 Manual," [Online]. Available: <https://numpy.org/doc/stable/reference/generated/numpy.concatenate.html>.
- [9] C. J. Willmott and K. Matsuura, "Advantages of the Mean Absolute Error (MAE) Over the Root Mean Square Error (RMSE) in Assessing Average Model Performance.," *climate research*, vol. 30, No 1, December 19 2005, pp. 79-82 (4 pages).
- [10] M.S. Al-Radhi, T.G. Csanó, C. Zainkó and G. Németh, "Towards Parametric Speech Synthesis Using Gaussian-Markov Model of Spectral Envelope and Wavelet-Based Decomposition of F0," 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 2022, pp. 1150-1154.
- [11] J. James, L. Tian, C. Watson, "An Open Source Emotional Speech Corpus for Human Robot Interaction Applications", in *Proc. Interspeech*, 2018.
- [12] M. Morrison, C. Hsieh, N. Pruyne, and B. Pardo, "Cross-domain Neural Pitch and Periodicity Estimation," *arXiv:2301.12258*, 2024.
- [13] A.R. Mandeel, M.S. Al-Radhi, and T.G. Csapó, "Speaker adaptation experiments with limited data for end-to-end text-to-speech synthesis using tacotron2," *Infocommunications Journal*, vol. 14, pp. 55–62, 2022, doi: 10.36244/ICJ.2022.3.7.
- [14] E. S. Kumar, K. J. Surya, K. Y. Varma, A. Akash, and K. N. Reddy, "Noise Reduction in Audio File Using Spectral Gating and FFT by Python Modules," *Recent Developments in Electronics and Communication Systems*, IOS Press, 2023, pp. 510–515.