A Hybrid Algorithm for Robust Pitch Estimation in Emotional Speech Synthesis

Zineb Hammadi and Mohammed Salah Al-Radhi

hammadi.zineb@edu.bme.hu, malradhi@tmit.bme.hu

Introduction

- Examine how humans experience mixed emotions and its role in emotional speech synthesis.
- Highlight the pitch signal's importance in conveying emotion and its challenge in emotional speech.
- Evaluate PESTO and FCPE on emotional speech datasets, identifying gaps.
- Introduce ESCAPE, new algorithm designed for robust pitch estimation in emotional speech.

Motivation

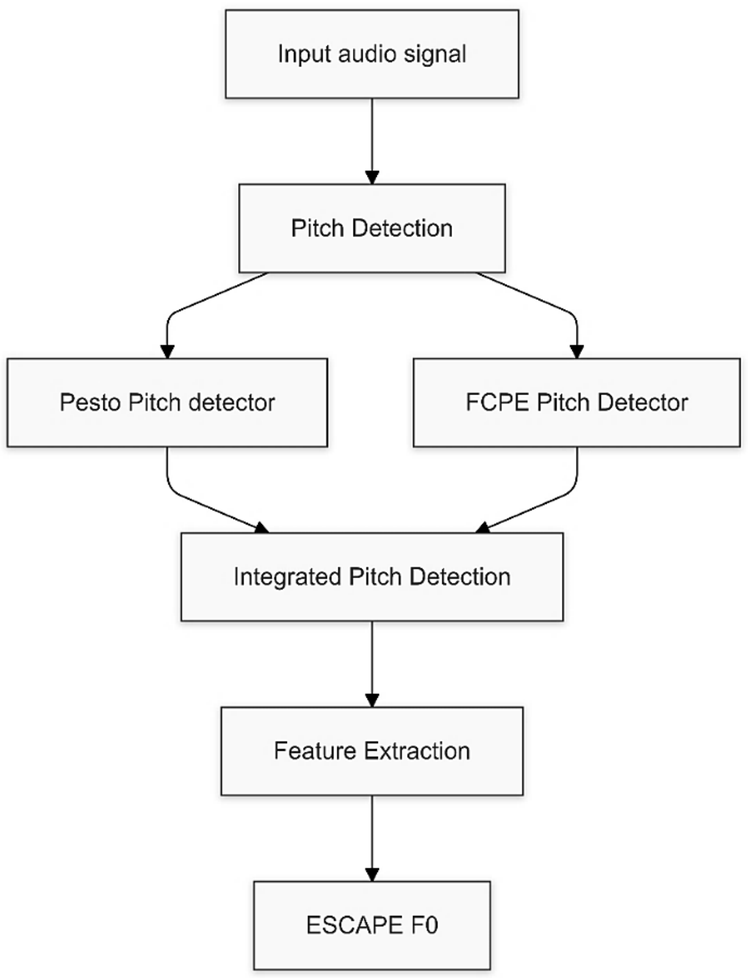
- Explore the challenges of synthesizing human-like emotional expressions in speech, focusing on mixed emotions and their nuances.
- Investigate the limitations of pitch estimation algorithms like PESTO and FCPE for emotional speech datasets.
- Highlight the role of pitch as a key acoustic feature for conveying emotion, emphasizing its variability in emotional contexts.
- Introduce ESCAPE, a hybrid algorithm combining PESTO's precision and FCPE's context-aware processing, designed for emotional speech synthesis.
- Demonstrate ESCAPE's ability to handle rapid pitch transitions, wide frequency variations, and irregular vibrato patterns, achieving robust and efficient pitch tracking.

Methodology

- **Feature extraction:** PESTO applies CQT and ResNet-based encoding, while FCPE uses a conformer-based encoder for context-aware processing.
- **Fusion:** Averages outputs from both models for balanced pitch estimation:

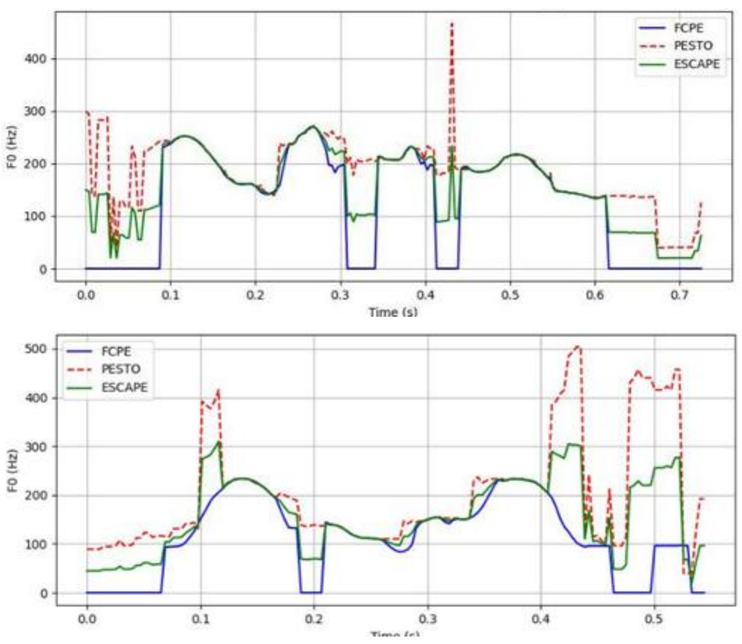
$$ESCAPE_{f0} = 0.5 \times PESTO_{f0} + 0.5 \times FCPE_{f0}$$

- **Training:** Self-supervised learning ensures generalization across emotional contexts.
- **Implementation:** Built in PyTorch for efficient preprocessing, hybrid modeling, and real-time performance.



Results

- **Dataset:** JL-Corpus (2400 sentences, New Zealand English), speech signals were sampled at 44.1 khz with 16-bit resolution.
- ESCAPE outperformed both PESTO and FCPE in pitch contour tracking .
- Our proposed model produced smoother and more naturalistic pitch contours.
- ESCAPE had fewer tracking errors, especially during rapid frequency changes.



- ESCAPE outperforms PESTO and FCPE in both metrics.
- **Gross Pitch Error:** ESCAPE achieves 0.6692, significantly lower than PESTO and FCPE.
- **Root Mean Square Error:** the proposed algorithm shows 1.4541, outperforming both baseline algorithms.

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

Conclusions

- Investigated pitch estimation for emotional speech synthesis.
- Highlighting the limitations of PESTO and FCPE in capturing rapid pitch modulations.
- Proposed ESCAPE, a hybrid algorithm combining self-supervised and context-aware mechanisms.

Future Work

- For our future work we aim to enhance ESCAPE's efficiency for speech synthesis applications.

Budapest University of Technology and Economics
Department of Telecommunications and Artificial Intelligence
Budapest, Hungary