

# Multimodal Emotion Recognition from Speech and Text: Comparing Audio-Only and Multimodal CNN Approaches

Sara Jose Roig

Gothenburg University Library

Maskininlärning för statistisk datalingvistik: avancerad

gusjosersa@student.gu.se

## Abstract

This project attempts to investigate what features are important in speech emotion recognition. To do so, it compares 2 different simple architectures: one convolutional neural network (CNN) with audio features only, and a multimodal CNN with audio and text features. The two models were trained with open-source datasets: CREMA-D, which contains emotional acted speech; and later on, was also trained with MELD, which features conversational dialogues. The baseline model aims to learn emotional cues only from the audio input, while the multimodal model learns audio and text separately and concatenates both types of features before classification. Both models were trained and evaluated using standard metrics such as accuracy, F1-score and confusion matrices. Then, those results were compared. Results show that multimodal effectiveness depends strongly on dataset characteristics and emotion class distribution. In controlled and simple scenarios, the multimodal model showed little to no improvement. In more complex scenarios, the multimodal model achieved an improvement but was not consistent over all the classes classified. The limitations and inconsistencies are likely due to the type of architecture and simplicity of the fusion mechanism of features used in the multimodal model. Therefore, this project provides with future directions and improvements that could be applied to achieve better results. Moreover, this study provides a reproducible baseline for future work in multimodal emotion recognition and demonstrates the practicality of integrating textual features into audio-based emotion classifiers.

## 1 Introduction

Human speech is inherently multimodal and conveys a lot of underlying information. Aside from the textual meaning, variation in features like pitch, intensity, rhythm and timbre are really important

to identify emotions. Therefore, understanding how this acoustic and textual information interact is a step towards more natural aware systems. In speech emotion recognition (SER), recognising these queues (textual and auditory) is of the upmost importance. This task can help in a wide range of fields, such as, virtual assistants and mental health monitoring to customer service and interactive learning systems.

This investigation is done by comparing to models with CNN architectures:

1. **Baseline CNN (audio-only):** a model trained exclusively on acoustic features derived from speech signals.
2. **Multimodal CNN (audio and text):** a model that processes both audio and text inputs through separate branches and fuses them at the feature level.

The project evaluates these models on two commonly used datasets: CREMA-D, containing acted emotional speech across six emotion categories, and MELD, featuring natural conversations with seven emotion classes. These datasets provide complementary perspectives: CREMA-D offers clean, controlled emotional expressions; and MELD represents spontaneous, contextual dialogue. Together, they enable a more detailed evaluation of unimodal versus multimodal emotion recognition.

The multimodal model architecture has a simple fusion strategy: audio features are extracted as spectrograms and passed through convolutional layers, while text features are represented through embeddings and processed via dense layers. The resulting feature vectors are concatenated and classified into emotion categories. Performance is assessed by evaluating accuracy, F1-score, and confusion matrices to examine both overall and individual class differences.

The final objective of this project is not to achieve state-of-the art results in emotion recognition but to establish a clean, interpretable and reproducible baseline for studying how textual features complement audio information. For that reason, a transparent data pipeline will be provided. The aim is to be a baseline for exploring richer fusion strategies in the future and more nuanced modelling of prosodic and semantic in expressive speech.

## 2 Deviation from Project Proposal

There has been a significant deviation and simplification from the initial project. The initial project aimed to build a small pipeline for prosody extraction. More precisely, to estimate low-level acoustic features such as duration, fundamental frequency and energy from text. For that task, I initially proposed to use the LibriTTS dataset, which provides high-quality speech recordings paired with transcriptions, and to obtain word-level timestamps through forced alignment using the Montreal Forced Aligner (MFA). Once aligned, per-word prosody features would be extracted from the audio and used as ground-truth targets for a regression model predicting these values from textual and semantic features. However, this task turned to be more complex than anticipated and technical challenges emerged. They were two main challenges: the complexity of achieving reliable word-level alignments and the model evaluation. Using MFA required extensive data cleaning, adjustments and manual verification of boundaries. Minor transcription inconsistencies or non-standard pronunciations frequently caused misalignments. All of that exceeded the time scope of this project. In addition, the extraction of accurate ground-truth prosodic features proved difficult to automate. Even small errors in alignment or unvoiced segments significantly affected these statistics, compromising the reliability of evaluation metrics such as RMSE and correlation. Moreover, Libri TTS consists primarily of read speech which offered a limited range of variation in prosody.

Therefore, the initial part of this project which consisted of exploring expressivity was maintained, but instead of prosody extraction the project was reframed to emotion recognition, which is closely related but uses tagged and well annotated datasets. In that way, the need of forced alignment or any kind of manual task was eliminated. Moreover, the

dataset was changed as week first for CREMA-D and later for MELD. These datasets already provide emotion labels paired with transcripts and ready to use audio files.

Therefore, the task was changed to a categorical classification. This new approach ensured that the project remained achievable within available time and computational resources, while still contributing valuable insights into multimodal representations of speech expressivity.

## 3 Ethical Implications, Limitations and Expectations

The datasets used for this project are open source and well known. Nevertheless, ethical considerations to consider. Even though all the data is publicly available and collected with informed consent from participants or actors, we must keep in mind that speech recordings inherently contain personally identifiable and sensitive information, such as vocal characteristics that can reveal gender, age or ethnicity. Moreover, these datasets are generated by American English speakers and may reflect cultural and contextual biases in emotional expression. Therefore, it is critical to interpret the results as dataset-specific and to avoid using them as a real world emotional assessment without further validation.

### 3.1 Limitations

This project has several limitations. Firstly, as briefly mentioned in the introduction the fusion mechanism in the multimodal CNN is deliberately simple, consisting of straightforward concatenation of audio and text embeddings. This approach offers interpretability, but it may fail to capture complex interdependencies between modalities that more advanced techniques, such as attention based fusion or transformer architectures, could achieve.

Secondly, the text features available in both datasets are often brief, consisting of short utterances or phrases. Thus, this restricts the information given by the text features. Additionally, the word embeddings were obtained using shallow methods instead of a pretrained large language model, which makes them less exact.

### 3.2 Expectations

With these limitations in mind, the project sets realistic expectations: we believe the baseline model may have an acceptable performance as acoustic

features are the primary carriers of emotion. The multimodal model may show an incremental improvement but not dramatic gains. Nevertheless, the outcomes are expected to demonstrate the value of multimodal learning even in small scale setups.

## 4 Project Background

Speech Emotion Recognition aims to automatically identify the emotional state of a speaker from speech signals through lexical and paralinguistic features as pitch, loudness or rhythm. Understanding these cues is essential for improving human-computer interaction and assistive technologies (Scherer, 2003).

Early approaches to SER relied on features such as Mel-Frequency Cepstral Coefficients (MFCCs) and formants, which were modelled using machine learning algorithms such as Support Vector Machines (SVM), Gaussian Mixture Models (GMM), or Hidden Markov Models (HMM) (Lee and Narayanan, 2005; El Ayadi et al., 2011). The main issue with these models is that they depended heavily on feature engineering and lacked robustness across different speakers and recording conditions.

Later, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) started being used to learn representations directly from spectrograms or raw waveforms (Trigeorgis et al., 2016). CNNs are particularly effective at capturing local time-frequency patterns that correspond to prosodic variations in emotional speech. Architectures such as Deep CNNs and attention-based models have since become standard for SER tasks (Zhao et al., 2019).

More recently, there have been some approaches that move beyond audio only models towards a multimodal emotion recognition, integrating complementary information from text, visual and physiological signals. This shift originates from the observation that emotion is inherently multimodal. In that way the same word can express different emotions depending on tone, and textual context can disambiguate acoustic ambiguity (Poria et al., 2018).

Multimodal systems commonly employ either feature-level fusion (concatenating embeddings before classification) or decision-level fusion (combining separate model predictions). Moreover, recent works further explore attention-based cross-modal fusion and transformer architectures that dy-

namically weigh modalities based on context (Han et al., 2021).

## 5 Data Resources and Requirements

This project uses two publicly available corpora: CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) and MELD (Multimodal EmotionLines Dataset).

CREMA-D (Cao et al., 2014) consists of 7,442 audio clips from 91 actors (48 male and 43 female) performing 12 sentences with six targeted emotions: Angry, Disgust, Fear, Happy, Neutral, and Sad. The recordings were produced under controlled conditions and validated by humans. Each utterance is stored as an audio file and accompanied by metadata specifying the emotion label, actor ID and sentence text.

MELD (Poria et al., 2019) is derived from the Friends TV series and includes multimodal dialogue turns annotated with seven emotions: Anger, Disgust, Fear, Joy, Neutral, Sadness, and Surprise. Unlike CREMA-D, MELD features spontaneous, showing more natural conversations. But also, it is more challenging dataset, due to background noise and overlapping speech. The dataset includes aligned audio files, textual transcriptions, and speaker IDs, which allows experiments that incorporate both linguistic and paralinguistic information.

### 5.1 Data Preprocessing

For audio processing each file was pre-processed, transformed and reshaped using Librosa library. For text preprocessing, transcriptions were tokenized and converted to embeddings. For both datasets, the data was split into training, validation and testing subsets. Features and scalers were serialized for reproducibility and intermediate results were stored.

### 5.2 Hardware and Software Requirements

The project was developed in Python 3.8+ using common machine learning and signal-processing libraries such as numpy, pandas, librosa, tensorflow, and Scikit-learn and Keras. All dependencies are listed in the requirements.txt file provided in the repository. Training and evaluation were performed locally on a system equipped with an NVIDIA RTX 3050 GPU

## 6 Methods

This project follows a modular and reproducible design in which feature extraction, model setup, training and evaluation are handled by separate scripts. Two main architectures were implemented: a Baseline CNN that relies only on acoustic information, and a Multimodal CNN that processes both audio and textual features in parallel. The following subsections describe the complete end to end process.

### 6.1 Feature Extraction

Audio preprocessing was performed, and audio files were normalized, resampled and converted into spectrograms. Features were standardized, reshaped for one-dimensional convolution, scaled and saved as a serialized file.

Text features were derived from the transcriptions by lowercasing, removing punctuation, tokenizing and converting to embeddings. Given the short utterances, lightweight embeddings were sufficient. Both modalities, paired with emotion labels, were stored as NumPy arrays.

### 6.2 Model Architectures

Two CNN architectures were implemented. The Baseline CNN processes the features through two Conv1D layers, followed by max pooling, dropout and global average pooling, a dense layer and a finally, a Softmax output.

The Multimodal CNN, extends the baseline by adding a parallel text branch with two dense layers that transform the text features into representations. Audio and text embeddings are concatenated and passed through a dense layer, and finally through Softmax classification. This concatenation fusion treats both modalities equally without any adaptive weighting.

For MELD, both models were enhanced with batch normalization, increased and class weighting was added to handle severe class imbalance. This change was done due to poor results obtained with the previous architecture.

### 6.3 Training Procedure

Training used the Adam optimizer and categorical cross-entropy loss. Early stopping was applied after five epochs without improvement, and 15% of data split was used for validation. The model weights were saved. Learning curves for accuracy and loss were tracked for both datasets.

### 6.4 Evaluation and Visualization

Models were evaluated on held-out test sets using overall accuracy, macro-F1, and class specific F1-scores. Confusion matrices, ROC curves, and F1-comparison plots were generated to visualize class performance and multimodal differences.

## 7 Implementation and Reproducibility

An essential component of this project is its open and reproducible implementation pipeline. The full codebase is publicly available in the GitHub repository at [github.com/sarajose/Multimodal-Emotion-Recognition](https://github.com/sarajose/Multimodal-Emotion-Recognition). This is done to allow researchers to replicate, inspect and extend the current project. The main stages of implementation and how reproducibility is addressed are described below.

The repository is organised in a modular way, enabling a separation between data processing, model definition, training and evaluation. The feature extraction scripts implement the preprocessing for both datasets, handling audio and text feature extraction. The models are defined in a dedicated file that includes both the baseline CNN and the multimodal CNN architectures. Training and evaluation are handled in separate scripts for each dataset (CREMA-D and MELD). These scripts include the main experimental configuration, hyperparameters, and routines for saving models. Additional scripts handle evaluation by loading the trained models, computing accuracy and F1-scores, and generating visualisations such as confusion matrices and ROC curves which are saved in the results folders.

To ensure reproducibility and transparency, several practices were implemented. Firstly, random seeds and data splits were fixed across all scripts to make training outcomes as consistent as possible. Secondly, scalers and feature sets were serialized as external files to ensure consistent scaling across experiments. Thirdly, all dependencies were documented in a requirements.txt file, allowing any user to recreate the same Python environment with the correct versions of the packages. Fourth, the project follows open science principles by using an open-source license that permits reuse, modification, and redistribution for academic purposes. Finally, visualisations and reports, including confusion matrices and per-class F1-score plots are stored in the repository and directly correspond to the figures discussed in this report.

In practical terms, reproducing the key results in-

volves a straightforward process. The user needs to download the target dataset (CREMA-D or MELD), run the corresponding preprocessing script to generate features, then execute the training script to build and train the model. When the training has finished, the evaluation script computes all the relevant metrics and produces the visualisations necessary to assess and visualize performance. The output results can be compared directly with those presented in this report to verify consistency.

This open and transparent implementation approach strengthens the reliability of the results and provides a solid foundation for future work. Other researchers can use the same structure to test alternative fusion methods, integrate different datasets and develop more complex architectures such as attention-based or transformer-based models. However, some degree of variability may still appear across different hardware configurations, specially using different settings.

In summary, this project prioritises reproducibility by providing modular, well-documented, and open-source code. Through consistent preprocessing, fixed random seeds, and detailed documentation of dependencies. So it ensures that other researchers can easily replicate or extend the findings. This approach promotes transparency and positions the project as a strong baseline for future multimodal emotion recognition studies.

## 8 Results

The results were evaluated in both datasets. On CREMA-D, the baseline CNN achieved 44.59% accuracy and a 41.41% F1-score, while enhanced model achieved slightly lower overall performance with 43.79% accuracy and 41.31% F1-score. This represents a modest decrease of 0.80 percentage points in accuracy. However, ROC-AUC analysis revealed a marginal improvement for the multimodal approach: the baseline achieved 0.788 AUC while the multimodal model reached 0.790 AUC, as shown in Figure 1.

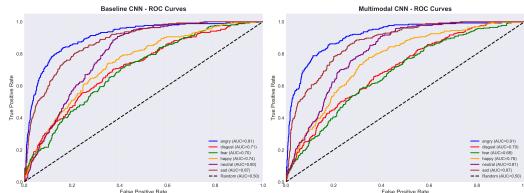


Figure 1: ROC curves comparison for CREMA-D dataset.

Class analysis showed mixed results. The multimodal model improved F1-score for fear detection by a small margin, suggesting that text features provided useful complementary information for this specific emotion. However, performance on angry, disgust, happy, neutral, and sad emotions showed slight decreases, with the most notable drop in the happy category.

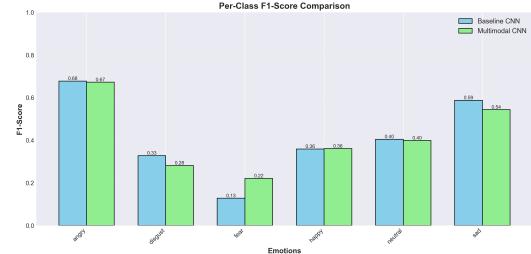


Figure 2: Per-Class F1-score comparison illustrates these per-class performance differences between models.

The confusion matrices (Figure 3) reveal that both models struggle with similar misclassification patterns, particularly confusing happy with neutral and fear with sad.

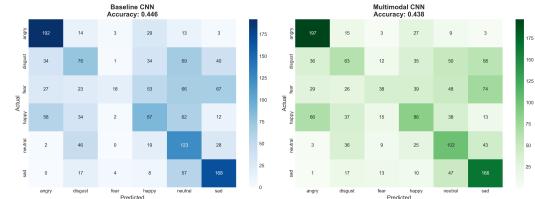


Figure 3: Confusion matrices comparison between models for CREMA-D dataset.

On MELD, the results demonstrate a notable improvement with the multimodal approach. The baseline CNN achieved 23.10% accuracy and 25.85% F1-score. The multimodal CNN showed a 38.70% accuracy and 35.72% F1-score. These results were achieved after implementing architectural improvements including batch normalization, increased model capacity and class weighting to handle classes imbalance, higher dropout rates for regularization and reduced learning rates for better convergence.

However, per-class F1-scores reveal mixed results, as illustrated in Figure 4. The enhanced model shows an improvement for majority classes like neutral and joy but performs worse on minority classes such as anger, disgust or fear. The confusion matrices in Figure 5 further demonstrate this pattern, showing better discrimination for well-

represented emotion classes while struggling with minority classes. This highlights the persistent challenge of recognizing underrepresented emotions in highly imbalanced datasets.

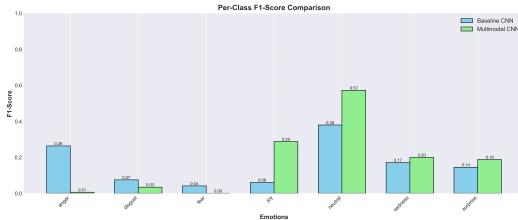


Figure 4: Per-class F1-score comparison for MELD dataset showing improvements in majority classes (neutral, joy) but degradation in minority classes (anger, disgust, fear).

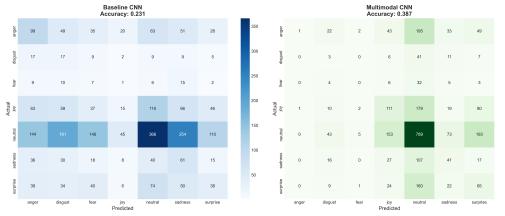


Figure 5: Confusion matrices comparison for MELD dataset between baseline and multimodal CNN models.

## 9 Relevant Analysis

The results show that effectiveness of multimodal emotion recognition depends strongly on dataset characteristics and model architecture. Therefore, we can see that depends a lot on the type of dataset and information in that dataset.

For CREMA-D, the multimodal approach led to minimal overall improvement and even a slight decrease in accuracy compared to the audio only baseline. While the enhanced model improved F1-score for fear, it reduced performance for most other emotions. This is likely because CREMA-D consists of acted, clearly articulated speech where audio features alone are highly informative and where text offers really little value.

On MELD, the multimodal model showed an improvement over the baseline, with accuracy rising from 23.10% to 38.70% and F1-score. However, this gain was not uniform across all emotion categories. The model performed better on majority classes such as neutral and joy, but its performance on minority classes like anger, disgust, and fear actually declined. This suggests that while multimodal fusion can help in more complex, conver-

sational settings, especially for well-represented emotions.

Overall, these findings indicate the enhanced model is not always beneficial. For controlled and acted speech, audio features may be enough, while in more naturalistic and imbalanced datasets, multimodal approaches can offer moderate improvements, particularly for the most common emotion categories. Nevertheless, more advanced feature representations and fusion techniques are likely needed to achieve consistent gains across all classes.

## 10 Error Analysis and Future Improvements

If we have a closer look at the misclassifications in the confusion matrices, we can see a few reasons behind the limited improvements of the multimodal model. In the CREMA-D dataset, most errors occurred between similar emotions such as happy and neutral or fear and sad, where acoustic cues overlap. Because the sentences in CREMA-D are fixed and semantically neutral, the textual features added little value, explaining why multimodal fusion did not improve results significantly.

In contrast, the MELD dataset presented a more complex setting with spontaneous conversations and unbalanced emotion classes. Here, the multimodal model improved performance for common emotions such as neutral and joy but underperformed for rarer classes, such as, disgust and fear. This suggests that the model’s simple concatenation fusion could not adaptively weigh modalities or overcome data imbalance.

These errors point to several directions for future work. Incorporating contextual text embeddings (e.g., BERT or RoBERTa) could help capture semantic nuances. Attention-based or gating fusion mechanisms could dynamically prioritize audio or text depending on reliability. Finally, class balancing or data augmentation should help reduce bias toward the emotions with bigger representation. Addressing these issues would lead to more consistent and generalizable multimodal emotion recognition systems.

## 11 Discussion

The findings reveal that enhanced model effectiveness depends critically on dataset characteristics and architectural design. The strong baseline CNN performance on CREMA-D aligns with established

literature showing that prosodic and spectral features correlate strongly with emotional expression (El Ayadi et al., 2011; Zhao et al., 2019).

The slight performance decrease mentioned in results can also be attributed at the simple concatenation between features. Unlike attention-based approaches in recent studies (Poria et al., 2018; Han et al., 2021), this project’s architecture lacks mechanisms to adaptively weight modalities based on their reliability.

On MELD, the overall improvement demonstrates that textual context can help disambiguate emotions when acoustic signals are degraded by noise, overlapping speakers or spontaneous speech. However, the performance disparities between classes with different amounts of data reveal a critical weakness. The decline in minority class performance suggests the model learned textual patterns that generalize poorly to underrepresented emotions.

Several factors limited multimodal effectiveness. The shallow text features lack the semantic richness of transformer-based embeddings like BERT, and the concatenation strategy treats all the features as equally reliable rather than dynamically emphasizing the most informative one. These limitations point to clear directions. Firstly, incorporating pre-trained encoders should be added. Secondly, adaptive fusion mechanisms should be developed. And finally, class imbalance should be addressed.

## 12 Conclusion

This project evaluates the integration of audio and text features for speech emotion recognition, comparing baseline and multimodal CNN architectures across two datasets with distinct characteristics. The experiments revealed that multimodal CNN is not always beneficial, but rather depends on data quality, emotion class distribution and architectural design choices.

The empirical findings demonstrate the potential and limitations of current approaches. While multimodal integration has the capability of improving overall performance in challenging conversational settings, it is not always the case. These results show that simply adding features is not always a good idea and does not guarantee improvement.

This project also gives a reproducible pipeline which encompasses, feature extraction, model training, and comprehensive evaluation. Moreover, it provides a transparent and usable data pipeline that

can be the base for future research. Showing in which cases a multimodal approach is likely to succeed or fail. Additionally, it identifies key limitations and provides concrete directions for improvement.

As a conclusion, the path to a more robust multimodal systems for emotion recognition it is recommended, with richer feature representations obtained from pretrained models and more adaptive fusion strategies. Moreover, it is important to have access to more diverse and bigger datasets with balanced classes to ensure fairness of results.

In summary, this project provides a clear and interpretable baseline for studying multimodal emotion recognition. It highlights when multimodal fusion is likely to succeed, when it may fail, and why model design and data characteristics matter. The findings, together with the open-source implementation, can work as a foundation for future experiments which could aim to build more reliable and fair emotion recognition systems.

## References

- H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. 2014. [Crema-d: Crowd-sourced emotional multimodal actors dataset](#). *IEEE Transactions on Affective Computing*, 5(4):377–390.
- M. El Ayadi, M. S. Kamel, and F. Karray. 2011. [Survey on speech emotion recognition: Features, classification schemes, and databases](#). *Pattern Recognition*, 44(3):572–587.
- K. Han, D. Yu, I. Tashev, and L. Lu. 2021. [Speech emotion recognition: Datasets, features, and challenges](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6289–6293. IEEE.
- C.-C. Lee and S. S. Narayanan. 2005. [Toward detecting emotions in spoken dialogs](#). *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303.
- S. Poria, E. Cambria, D. Hazarika, and P. Vij. 2018. [Affective computing for speech-based emotion recognition and sentiment analysis: Review, trends, and challenges](#). *Information Fusion*, 44:103–119.
- S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea. 2019. [Meld: A multimodal multi-party dataset for emotion recognition in conversations](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.
- K. R. Scherer. 2003. [Vocal communication of emotion: A review of research paradigms](#). *Speech Communication*, 40(1–2):227–256.

G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. 2016. **Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network.** In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, Shanghai, China. IEEE.

J. Zhao, X. Mao, and L. Chen. 2019. **Speech emotion recognition using deep 1d & 2d cnn lstm networks.** *Biomedical Signal Processing and Control*, 47:312–323.