International Scientific Conference 20 UNITECH 26

INTERNATIONAL SCIENTIFIC CONFERENCE
19-21 November 2026
GABROVO

COMPARATIVE EVALUATION OF ACOUSTIC AND TEMPORAL FEATURES FOR FEAR AND BOREDOM RECOGNITION IN SPEECH

Bojan Prlinčević[1], Zoran Milivojević[2], Violeta Stojanović[2], Petar Milić[3], Dijana Kostić[4]

*[1]Kosovo and Metohija Academy of Applied Studies, Dositejeva bb, Leposavić, Serbia**
[2]The Academy of Applied Technical and Preschool Studies, A. Medvedeva 20,Niš, Serbia
[3]University of Pristina - Kosovska Mitrovica, Faculty of Technical Sciences, K. Mikoša 7, K. Mitrovica, Serbia
[4]"Šargan inženjering" d.o.o, Niš, Niš, Serbia

……
**Corresponding author: bojan.prlincevic@akademijakm.edu.rs*

Abstract

This paper investigates speech-based classification of Fear and Boredom using low-dimensional acoustic and temporal features. Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Support Vector Machines (SVM), and a feedforward Artificial Neural Network (ANN) are evaluated on 150 speech utterances using a fixed training–test split. The experiments compare fundamental frequency (F_0), variance, and utterance duration as classification features. The highest test accuracy of 78.57% is achieved by LDA and ANN using F_0 and duration. While ANN produced no Boredom-to-Fear false positives on the evaluated test subset, LDA provides higher Fear sensitivity (65% versus 55%). Adding F_0 variance to the duration-based ANN configuration reduces test accuracy to 59.52%. The results highlight the trade-off between Fear sensitivity and false-alarm control and indicate that feature expansion does not necessarily improve classification performance on limited datasets.

Keywords: Speech Emotion Recognition; Fear and Boredom; Fundamental Frequency; Utterance Duration; Discriminant Analysis; Neural Networks.

INTRODUCTION

Speech emotion recognition (SER) supports affect-aware voice interfaces, telecommunication systems, and human–machine interaction. Modern deep architectures learn rich speech representations, but their computational cost and data requirements motivate investigation of compact, interpretable acoustic features [1], [2].

Pitch, energy, and speech duration are established correlates of vocal emotion [3]. However, fear and boredom may exhibit partially overlapping prosodic patterns, making discrimination difficult when a single static descriptor is used. Discriminant analysis and support vector machines provide useful statistical baselines, while neural networks can represent nonlinear boundaries [5], [6]. CNN–recurrent and transformer-based SER systems are also widely studied [7]–[9], although direct comparisons depend on the corpus and experimental conditions.

Authors in earlier study investigated Equivalent Torque Characteristic (ETC) descriptors derived from changes in fundamental frequency [10]. The present work examines whether combining pitch-related and temporal descriptors improves the separation of fear and boredom. Rather than treating all emotions as low-arousal states, it focuses on their overlap in the selected acoustic feature spaces.

The main contributions of this study are: (i) a comparative evaluation of low-dimensional pitch-related and temporal feature representations for binary fear–boredom classification; (ii) a systematic

This is an open access article licensed under Creative Commons Attribution 4.0 International
Print ISSN: 1313-230X, Online ISSN: 3033-1404

comparison of LDA, QDA, SVM, and feedforward ANN classifiers; and (iii) an analysis of class-dependent recognition performance through overall accuracy, fear sensitivity, and boredom-to-fear false-positive rates. The findings are restricted to the evaluated dataset and training–test partition.

MATERIALS AND METHODS

A. Speech Database and Experimental Protocol

The experiments were conducted using 150 utterances selected from the Berlin emotional speech database [11], including 69 fear and 81 boredom samples. The dataset was randomly divided into a training subset of 108 utterances (49 fear and 59 boredom) and a test subset of 42 utterances (20 fear and 22 boredom).

Fear was defined as the positive class and boredom as the negative class. Classification performance was evaluated using accuracy (ACC), true-positive rate (TPR), and false-positive rate (FPR), calculated as:

$$ACC = \left(TP + TN\right) / \left(TP + TN + FP + FN\right) \tag{1}$$

$$TPR = TP / \left(TP + FN\right) \tag{2}$$

$$FPR = FP / \left(FP + TN\right) \tag{3}$$

The experimental evaluation was conducted using a fixed training–test partition, ensuring consistent conditions for the comparative assessment of the selected classification models.

B. Acoustic Feature Extraction

The acoustic representation is based on the fundamental frequency (F_0) trajectory estimated from each speech utterance. Three utterance-level descriptors are considered: mean fundamental frequency, F_0 variance, and utterance duration. The overall processing framework is illustrated in Fig. 1. The speech signal is represented by utterance-level acoustic and temporal descriptors, which are organized into three evaluated feature configurations. These configurations are subsequently used for fear–boredom classification, and their performance is assessed using ACC, TPR, and FPR. The ETC-related branch illustrates the methodological continuity with previous research [10], providing a broader context for the acoustic feature extraction framework employed in this study.

The investigated feature vectors are defined as:

$$x_1 = \left[mean\left(F_0\right), \sigma^2 \right] \tag{4}$$

$$x_2 = \left[mean\left(F_0\right), T \right] \tag{5}$$

$$x_3 = \left[mean\left(F_0\right), \sigma^2, T \right] \tag{6}$$

where σ^2 denotes the variance of valid F_0 estimates and T represents utterance duration. The first two representations are evaluated using the selected statistical and neural classifiers, while the third is evaluated using ANN.

These configurations enable a direct comparison of pitch dispersion and duration and their influence on classification performance. The ETC-based descriptors provide a methodological foundation for the broader acoustic feature-extraction framework, supporting the systematic investigation of pitch-related and temporal characteristics in speech emotion recognition.

C. Classification models

Four classification algorithms were considered: LDA, QDA, SVM, and a feedforward ANN. LDA and QDA were implemented using the fitcdiscr function in MATLAB, while SVM was implemented using fitcsvm. The ANN was implemented using MATLAB Neural Network Toolbox.

A consistent experimental protocol was applied to evaluate the selected classification models, using the same training and test partitions for each feature configuration to ensure comparability of the obtained results..

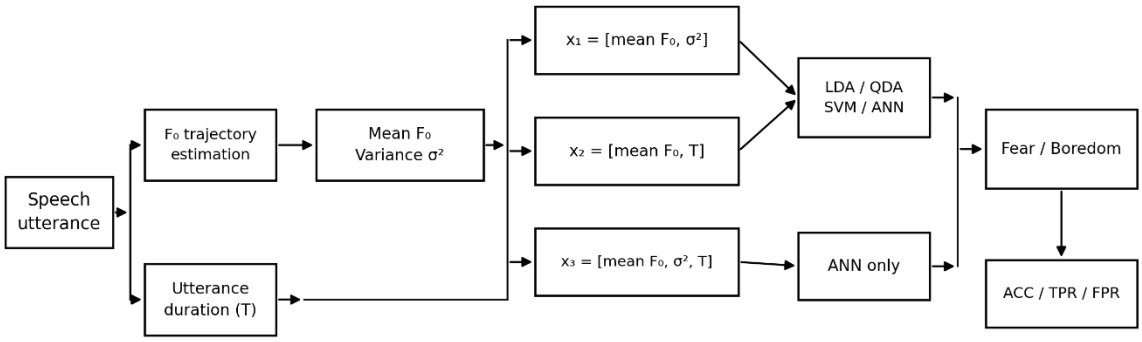


Fig. 1. *Block diagram of the acoustic feature extraction and speech emotion classification framework.*

EXPERIMENTAL RESULTS AND ANALYSIS

A. Pitch and dispersion: (F_0, σ^2)

The classification performance obtained using mean fundamental frequency and its variance is summarized in Table I. Among the evaluated classifiers, ANN achieved the highest test accuracy (76.19%) and fear sensitivity (75.00%), with a false-positive rate of 22.73%. LDA and QDA produced lower false-positive rates (13.64%), whereas SVM exhibited substantial boredom-to-fear confusion (FPR = 54.55%), resulting in the lowest test accuracy (57.14%).

Table I. *Confusion-matrix parameters for Fear–Boredom classification in the (F_0, σ^2) plane.*

Model	Split	ACC	TPR (Fear)	FPR (Boredom→Fear)
LDA	Train	0.7870	0.6735	0.1186
LDA	Test	0.7381	0.6000	0.1364
QDA	Train	0.7963	0.6735	0.1017
QDA	Test	0.7143	0.5500	0.1364
SVM	Train	0.6296	0.7551	0.4746
SVM	Test	0.5714	0.7000	0.5455
ANN	Train	0.9167	0.8776	0.0508
ANN	Test	0.7619	0.7500	0.2273

The test-set decision regions shown in Fig. 2 provide a visual interpretation of these results. The LDA and QDA decision boundaries exhibit relatively conservative separation of the two classes, consistent with their lower false-positive rates. In contrast, the SVM decision region extends further into the boredom sample distribution, explaining its increased frequency of boredom-to-fear misclassifications. The overlap between the two classes indicates that mean F_0 and pitch variance alone do not provide complete discrimination between fear and boredom.

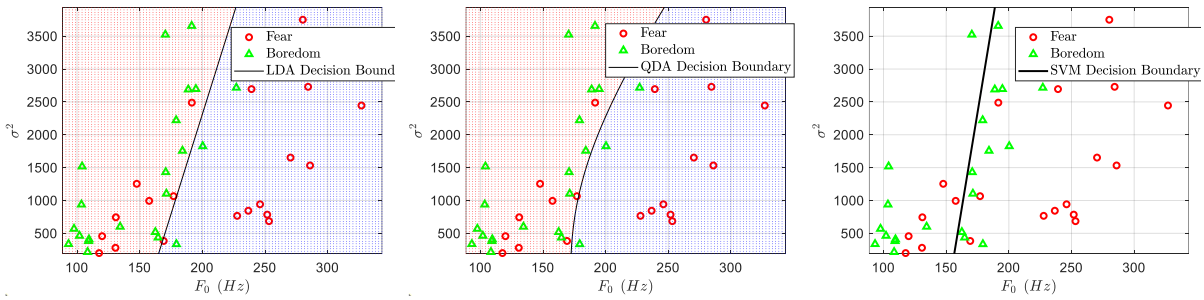


Fig. 2. *Test-set decision regions in the (F0, σ²) plane for LDA, QDA, and SVM.*

B. Pitch and duration: (F_0, T)

The second experiment investigates the effect of replacing pitch variance with utterance duration. The corresponding training and test results are presented in Table II. On the test subset, LDA and ANN achieved the highest accuracy of 78.57%, followed by SVM with 76.19%. However,

the classifiers exhibited different error profiles, particularly in terms of fear sensitivity and boredom-to-fear false alarms..

The duration feature gives a lower test false-alarm rate than the pitch-dispersion representation, particularly for ANN. However, ANN misses more fear instances (TPR = 55.00%) than LDA (TPR = 65.00%).

These findings highlight the importance of considering both fear sensitivity and false-positive rates alongside overall accuracy when evaluating and selecting classification models.

Table II. *Confusion-matrix parameters for Fear–Boredom classification in the (F_0, T) plane.*

Model	Split	ACC	TPR (Fear)	FPR (Boredom→Fear)
LDA	Train	0.7685	0.7347	0.2034
LDA	Test	0.7857	0.6500	0.0909
QDA	Train	N/A	N/A	N/A
QDA	Test	N/A	N/A	N/A
SVM	Train	0.7685	0.6735	0.1525
SVM	Test	0.7619	0.6000	0.0909
ANN	Train	0.8519	0.7755	0.0847
ANN	Test	0.7857	0.5500	0.0000

Although LDA and ANN achieved identical test accuracy (78.57%), their confusion-matrix distributions differed. LDA correctly classified 13 of 20 fear utterances and 20 of 22 boredom utterances, whereas ANN correctly classified 11 fear utterances and all 22 boredom utterances. Thus, ANN produced no boredom-to-fear false positives in the evaluated test subset, but at the expense of two additional missed fear detections compared with LDA. This result demonstrates that overall accuracy alone is insufficient for assessing class-dependent recognition performance.

The decision regions presented in Fig. 3 illustrate the class distribution in the (*F0*, *T*) feature space. Compared with the pitch-variance representation in Fig. 2, the duration-based representation exhibits a different spatial arrangement of fear and boredom samples. The LDA and SVM boundaries are consistent with the relatively low false-positive rates reported in Table II. Nevertheless, the partial overlap between the classes remains evident, indicating that duration alone does not eliminate classification ambiguity.

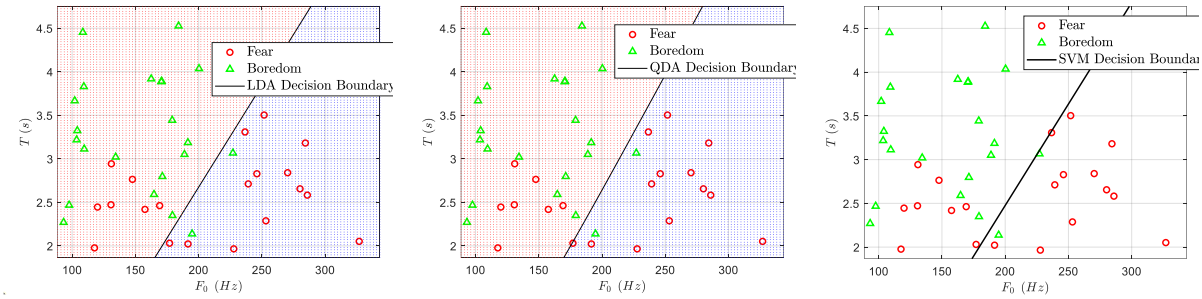


Fig. 3. *Test-set decision regions in the (F0, T) plane for LDA, QDA, and SVM.*

C. Three-dimensional feature representation

The third experiment evaluates the feedforward ANN using the three-dimensional feature vector (F_0, σ^2, T). The resulting classification performance is summarized in Table III. The model achieved 59.52% test accuracy, with 40.00% fear sensitivity and a false-positive rate of 22.73%. Compared with the two-dimensional ANN configurations reported in Tables I and II, the three-dimensional representation resulted in lower classification accuracy and reduced fear sensitivity on the evaluated test subset.

The distribution of fear and boredom samples in the (σ^2, T) plane is illustrated in Fig. 4. The two classes exhibit overlapping regions, suggesting that the combination of pitch dispersion and utterance duration does

not yield a straightforward separation of emotional states. This observation provides further insight into the distribution and interaction of acoustic and temporal features within the three-dimensional ANN classification framework.

Table III. *Confusion-matrix parameters for the 3D ANN configuration (F_0, σ^2,T).*

Model	Split	ACC	TPR (Fear)	FPR (Boredom→Fear)
ANN (3D)	Train	0.7407	0.4898	0.0508
ANN (3D)	Test	0.5952	0.4000	0.2273

Figure 5 presents the three-dimensional distribution of fear and boredom samples in the (F_0, σ^2, T) feature space for the training and test subsets. Partial overlap between the emotional classes is visible in both distributions, with no clear separation along a single feature dimension. These visualizations provide valuable insight into the distribution and interaction of acoustic and temporal features, highlighting the importance of feature selection and dimensionality in achieving effective class discrimination. Future research will focus on extending the experimental evaluation to larger datasets and multiple data partitions, providing a more comprehensive assessment of classification performance and the contribution of different feature configurations.

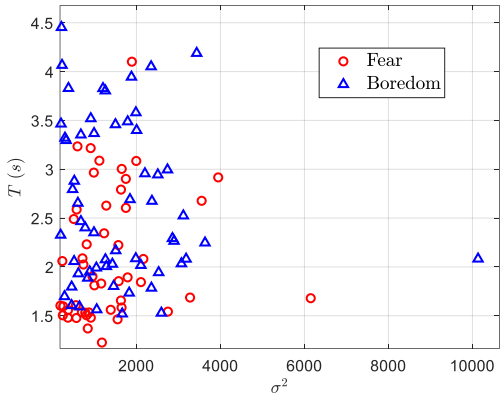
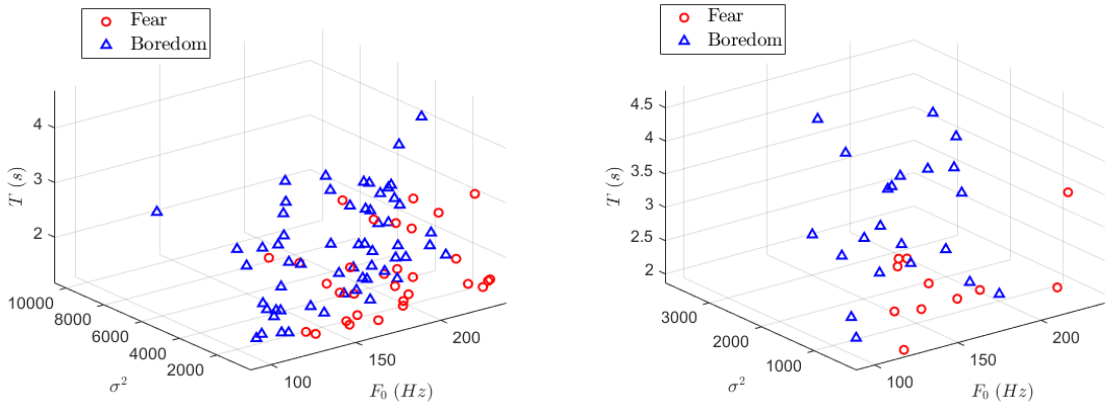


Fig 4. *Distribution of fear and boredom utterances in the (σ^2, T) feature space.*



a) *b)*

Fig 5. *Three-dimensional distribution of fear and boredom utterances in the (F_0, σ^2, T) feature space: (a) training subset; (b) test subset.*

D. Discussion

The experimental results demonstrate that feature selection influences both overall classification accuracy and class-dependent error distribution. As reported in Tables I and II, replacing pitch variance with utterance duration improved the test accuracy of LDA and SVM and reduced boredom-to-fear false-positive rates across the evaluated classifiers. The observed variations in fear sensitivity provide additional insight into the influence of different feature configurations on classification performance.

The decision boundaries shown in Figs. 2 and 3 illustrate partial overlap between fear and boredom in both two-dimensional feature spaces. This overlap helps explain

why improved overall accuracy does not necessarily correspond to improved recognition of both classes. In particular, LDA and ANN achieved identical test accuracy in the duration-based representation but exhibited different confusion-matrix distributions.

The three-dimensional distribution in Fig. 5 and the results in Table III further indicate that adding pitch variance to the duration-based ANN configuration did not improve classification performance on the evaluated partition. These findings suggest that feature relevance and class-dependent errors should be considered alongside model complexity when designing compact speech emotion recognition systems.

The experimental findings provide a foundation for further evaluation using larger datasets, repeated stratified data partitions, and speaker-independent testing, enabling a more comprehensive assessment of classification performance and the robustness of the observed trends.

CONCLUSION

This paper compared LDA, QDA, SVM, and feedforward ANN classifiers for speech-based fear and boredom recognition using low-dimensional acoustic and temporal features. The highest test accuracy of 78.57% was achieved by LDA and ANN in the (F_0, T) feature space.

Despite their identical accuracy, the classifiers exhibited different error distributions. LDA achieved higher fear sensitivity (65%), whereas ANN produced no boredom-to-fear false alarms on the evaluated test subset, with fear sensitivity of 55%. The three-dimensional ANN configuration achieved a classification accuracy of 59.52%, providing additional insight into the influence of feature dimensionality and the interactions between acoustic and temporal descriptors on classification performance.

The results emphasize the importance of feature selection and class-dependent error analysis. Future research will focus on exploring additional temporal descriptors and extending the experimental evaluation to larger datasets and multiple speaker-independent partitions, enabling a more comprehensive assessment of classification performance and model generalization.

ACKNOWLEDGMENTS: *This work was co-funded by the European Union through the Erasmus+ NEST4WB project (Grant Agreement No. 101129321). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.*

REFERENCE

[1] A. Kaya and T. A. Salah, “Deep learning for emotion recognition from speech: A review,” Machine Learning with Applications, vol. 6, art. 100198, 2021.

[2] K. J. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in Proc. INTERSPEECH, 2020, pp. 2232–2236.

[3] E. Tzinis, A. Tsiartas, and S. Narayanan, “Segment-based speech emotion recognition using recurrent neural networks,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 27, no. 8, pp. 1530–1540, 2019.

[4] A. Dhall, J. Joshi, K. Sikka, and R. Goecke, “EmotiW 2020: Audio-video, student engagement and group-level affect prediction,” in Proc. ACM ICMI, 2020.

[5] Y. Huang, M. Wang, and Y. Zhao, “Speech emotion recognition using modified deep residual networks,” Applied Sciences, vol. 10, no. 11, pp. 1–13, 2020.

[6] N. Saha, A. Swetapadma, and M. Mondal, “A brief review on artificial neural network: Network structures and applications,” in Proc. ICACCS, IEEE, 2023, pp. 1974–1979.

[7] M. Li, Y. Yang, and T. Chen, “Lightweight CNN-BiLSTM ensemble for emotion detection in speech,” Scientific Reports, vol. 15, 2025.

[8] Y. Feng, K. Wang, and L. Zhao, “Speech emotion recognition via CNN-Transformer with multi-dimensional attention,” arXiv:2403.04743, 2024.

[9] F. Bao and M. Neumann, “Effective feature fusion in transformer-based speech emotion recognition,” in Proc. ICASSP, 2023.
[10] B. Prlinčević, Z. Milivojević et al., “Speech Emotion Detection Using the Equivalent Torque Characteristic Algorithm and Neural Networks,” in RIM 2025, pp. 1–7.
[11] Technische Universität Berlin, “Emotional speech,” https://www.tu.berlin/en/kw/research/projects/emotional-speech.
[12] J. Volmer, Getriebetechnik, Braunschweig/Berlin, 1978.