Applied Computer Science 2025, vol. 21, no. 4, pp. 110–126
https://doi.org/10.35784/acs_8130

Submitted: 2025-07-21 | Revised: 2025-09-01 | Accepted: 2025-09-10

Keywords: features fusion, speech emotion recognition, Kronecker convolution, speech signal processing

Pawel POWROZNIK [1*], *Maria SKUBLEWSKA-PASZKOWSKA* [1]
[1] Lublin University of Technology, Poland, p.powroznik@pollub.pl, maria.paszkowska@pollub.pl
[*] Corresponding author: p.powroznik@pollub.pl

K4F-Net: Lightweight multi-view speech emotion recognition with Kronecker convolution and cross-language robustness

Abstract

Speech emotion recognition has been gaining importance for years, but most of the existing models are based on a single signal representation or conventional convolutional layers with a large number of parameters. In this study, we propose a compact multi-representation architecture that combines four images of the speech signal: spectrogram, MFCC features, wavelet scalogram, and fuzzy transform maps. Furthermore, the application of Kronecker convolution for efficient feature extraction with an extended receptive field is shown. Another novelty is cross-fusion, a mechanism that models interactions between branches without significantly increasing complexity. The core of the network is complemented by a transformer-based block and language-independent adversarial learning. The model is evaluated in a scenario of quadruple cross-lingual tests covering four data corpora for four languages: English, German, Polish and Danish. It is trained on three languages and tested on the fourth, achieving a weighted accuracy of 96.3%. In addition, the influence of selected activation functions on the classification quality is investigated. Ablation analysis shows that removing the Kronecker convolution reduces the efficiency by 5.6%, and removing the fuzzy transform representation by 4.7%. The obtained results indicate that the combination of Kronecker convolution, multi-channel fusion, and adversarial learning is a promising direction for building universal, language-independent emotion recognition systems.

1. INTRODUCTION

Verbal communication involves the exchange of information, ideas, moods, or feelings through the use of words and sounds (George & Ilyas, 2024). It contains a wealth of information about the speakers, the content of their message, and their emotional state. Speech emotion recognition (SER) has been widely applied to detect various emotional states that affect attention, motivation, and relationships with other people. It accompanies people in their daily lives, including vocal instructions, smartphones, synthesized speech, and criminal investigations (Ezzameli & Mahersia, 2023; Ntalampiras et al., 2009). In particular, studies on acoustic data have recently received much attention in the context of the rapid development of machine learning and deep neural networks. Classification and regression are applied to voice identification as well as to emotion recognition. A large number of studies concern different types of SER based on utterance-based (Zhao et al., 2014) context-aware (Kakuba & Han, 2022; Trigeorgis et al., 2016) cross-corpus (Latif et al., 2020) cross-linguistic (Song et al., 2016) and cross-cultural systems (Kamaruddin et al., 2012). These supervised machine learning techniques provide both traditional handcrafted feature extraction and automated processes to extract specific information from signals (Jin et al., 2015). Speech preprocessing transforms the signal into different types of features. Mel Frequency Cepstral Coefficients (MFCC), Gamma Tone Cepstral Coefficients, Linear Predictive Cepstral Coefficients, Bark Frequency Cepstral Coefficients and Mel-Spectrograms are prominent for SER (Chwaleba & Wach, 2024; George & Ilyas, 2024; Motamed et al., 2017; Prasomphan, 2015).

Many studies deal with SER using Support Vector Machine (SVM), Neural Network (NN), Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) (El Ayadi et al., 2011; Ververidis & Kotropoulos, 2006). Deep learning models have achieved remarkable performance in SER. Convolutional Neural Networks (CNNs) are widely used with automatic feature extraction (Chowdhury et al., 2025; Madanian et al., 2025).

CNN-based models using transformers with multi-dimensional attention mechanisms are said to have great potential for SER in human-machine interaction (Ahn et al., 2025; Tang et al., 2025).

1.1. Motivation

The models used for SER typically have a single signal representation or standard convolutional layers with a large number of parameters. Each signal representation conveys different types of features on which recognition is based. In this study, we apply multimodality to extract the most important features of the signal for speech recognition. This approach provides more accurate information about human emotion by using the signal representation from different modalities. Another goal is to test whether Kronecker convolution can improve the efficiency of feature extraction with an extended receptive field. The combination of CNN with transformer-based blocks provides better performance by exploiting local and global signal dependencies. This type of architecture can improve SER detection.

1.2. Contribution

This study focuses on enhancing the speech emotion recognition. The main aim is to propose a compact, multi-representation architecture that combines four images of the speech signal: spectrogram, MFCC features, wavelet scalogram and fuzzy-transform maps. Thus, the main contributions of this study are as follows:

- Four corpora of emotional speech representing Polish, English, German and Danish are included in this study. They are chosen for their linguistic diversity and realistic acoustic variability.
- The Kronecker Convolution with Four Feature Modalities (K4F-Net) model is proposed for SER. Four signal representations are used for signal preprocessing. Spectrogram, MFCC features, wavelet scalogram and fuzzy transform maps provide different types of relevant characterization of speech signals. The fusion of these modalities provides both local and global signal dependencies. Standard convolution is replaced by Kronecker operation. This approach provides a more comprehensive capture of local features. The novelty is cross-fusion, a mechanism that models interactions between branches without significantly increasing complexity. The core of the network is complemented by a transformer-based block and language-independent adversarial learning.
- The proposed model has been verified using the following metrics: weighted accuracy, macro precision, macro recall, and macro F1 score.
- In order to quantify the importance of individual components of the proposed model, a series of ablation experiments are performed on the Polish-held-out fold. The ablation analysis shows that removing the Kronecker convolution reduces the efficiency by 5.6%, and removing the fuzzy transformation representation reduces the efficiency by 4.7%.

2. RELATED WORKS

Speech emotion recognition (SER) has become a hot topic in affective computing because prosodic cues convey a substantial amount of human intent. Behavioral studies attribute nearly half of communicative content to vocal tone, rather than facial or gestural signals (Johanson et al., 2021; Madanian et al., 2023). Correctly identifying a speaker's affective state is therefore essential for natural human-computer interaction, prompting intense efforts to design deep learning systems that can perceive and respond to users' emotions with increasing accuracy (Johanson et al., 2021; Mishra et al., 2025). The task remains challenging due to background noise, channel distortion, and the linguistic diversity that characterizes real-world deployments (Ibrahim et al., 2024; Powroźnik & Czerwiński, 2016).

Modern SER pipelines rely on both one-dimensional waveforms and two-dimensional time-frequency images (Abdel-Hamid et al., 2014; Echim et al., 2024). Waveform level features are typically processed by 1-D CNNs, often cascaded with recurrent layers such as LSTM or GRU. However, some studies have attempted to process speech signals in a feedforward manner (Avots et al., 2019; Powroźnik, 2014).Using multilayer perceptrons or combining spectral and prosodic features improves robustness to noise (Czerwinski & Powroźnik, 2018) especially when focusing on a specific language.

Although CNNs still dominate SER, Vision Transformer (ViT) architectures have gained momentum due to their self-attention mechanism, which models long-range spatial dependencies more efficiently than convolutions (Akinpelu et al., 2024; Khasgiwala & Tailor, 2021). Hybrid CNN/ViT solutions have improved

classification accuracy by up to 17.7% (Khasgiwala & Tailor, 2021). CNN backbones such as ResNet-50 are often used to suppress noise (Kim & Lee, 2025) while lightweight variants of ViT (e.g., l-ViT) can outperform traditional CNNs in accuracy (Akinpelu et al., 2024).

Pre-trained ViT and BEiT backbones tested for human-robot interaction (Mishra et al., 2025). A dual-path fusion of MaxViT and MViTv2 with an MLP head (MaxMViT-MLP) achieved state-of-the-art results by combining CQT and Mel-STFT spectrograms, thus exploiting both logarithmic and linear frequency scales (Ong et al., 2024). SepTr decouples temporal and spectral attention via separable transformer blocks (Ristea et al., 2022) while the Audio Spectrogram Transformer (AST) performs well with either random initialization or AudioSet pre-training (AST).(Gong et al., 2021). Vertically segmented patches of log-Mel spectrograms further push accuracy (Kim & Lee, 2024).

For Arabic and English corpora, transformer-based models outperform ViT, wave2vec, and other baselines. CoordViT, which concatenates coordinate information, also performs well (Mohamed et al., 2024). Combining ViT patching along the time axis with parallel CNN feature extractors improves performance (Hashemi & Asgari, 2023). Compact Convolutional Transformers (CCTs) Demonstrate Robustness in Cross-Corpus Environments (Arezzo & Berretti, 2022). In a large Romanian-German benchmark, CvT and AlexNet outperformed alternatives such as CNN-LSTM, VGG-16, ViT, and LeViT; Grad-CAM++ maps revealed the most salient spectral regions (Echim et al., 2024).

Interest is shifting to models that fuse multiple modalities. Transformer architectures that combine facial landmarks, action units, head pose, and MFCCs improve robustness (Chumachenko et al., 2022). Joint speech-face embeddings have been explored with xlsr-Wav2Vec 2.0 (Luna-Jiménez et al., 2021) while the Fuzzy Multimodal Transformer (FMMT) integrates audio, visual, and textual cues for a richer affective context (Liu et al., 2025).

Noting that there is no single large, language-universal corpus, recent work also explores the merging of heterogeneous datasets to improve generalization across languages and speaking styles (Ibrahim et al., 2024).

3. MATERIALS AND METHODS

3.1. Datasets

To ensure both linguistic diversity and realistic acoustic variability, the experimental material combines four publicly available corpora of emotional speech in Polish, English, German and Danish.

The Polish corpus (Database A) consists of 240 studio-quality utterances recorded by eight professional actors, evenly divided by gender, expressing joy, anger, sadness, fear, boredom and a neutral mood. It was compiled at the Lodz University of Technology and is described in (Kaminska et al., 2013).

The English corpus (Database B) was collected at the Center for Strategic Technology Research and contains 700 utterances covering anger, happiness, sadness, fear and neutrality. Thirty actors produced the material, which was then validated by a separate panel of thirty listeners (Petrushin, 2000).

The German corpus (Database C) is the Berlin Emotional Speech Database, which consists of a total of 535 carefully selected recordings depicting anxiety, fear, boredom, joy, anger, and indignation (Hareli & Hess, 2012). Twenty trained actors contributed the speech, and only items with a recognition rate of at least 80% in a 20-listener evaluation were retained.

Finally, the Danish corpus (Database D) - the Danish Emotional Speech Database (DES) - comes from the Center for Personal Communication at Aalborg University (Engberg & Hansen, 1996). It includes isolated words, short sentences, two passages, and eighteen longer segments representing anger, sadness, joy, surprise, and a neutral tone. Recordings of four actors were included and subsequently reviewed by a group of twenty listeners ranging in age from 18 to 58.

To increase acoustic diversity and minimize model overfitting on small, acted corpora, a number of different speech enhancement techniques were used in this study. Each sample was randomly modified with parameters chosen from continuous intervals to preserve natural speech. A time shift of up to ±50ms was introduced, training the network to be robust to small decalibrations of the STFT windows. A speed/pitch perturbation was applied: a factor of 0.9-1.1 changed the tempo without significantly affecting intelligibility, while shifting the formants and enriching the acoustic space. Background noise with an SNR of 10-25 dB, randomly selected from office, street, and white noise recordings, was added to increase robustness to typical field listening conditions. In selected samples, reverberation was synthetically simulated by convolution with

a randomly selected room impulse response, and bandpass filters mimicking cell phone or laptop speakers were applied to 20% of the examples. At the spectral level, time-frequency masking was applied, randomly excluding narrow bands or short temporal fragments, forcing the search for generalized patterns. Finally, rare emotion classes were augmented by mixing two examples of the same label, maintaining the balance between languages and emotions without artificially duplicating identical recordings. In total, the augmentation increased the effective size of the training set by more than five times. The final structure of the applied data set is shown in Table 1.

Tab. 1. The final structure of used datasets

Corpus (Language)	No. of speakers	Emotions	No. of items before augmentation	No. of items after augmentation
A (Polish)	8	joy, anger, sadness, fear, boredom, neutral	240	1112
B - SAVEE (English)	30	anger, joy, sadness, fear, neutral	700	3326
C – Emo-DB (German)	10	anxiety, fear, boredom, joy, anger, indignation	535	2675
D – DES (Danish)	4	anger, sadness, joy, surprise, neutral	250	1232

3.2. Discrete fourier transform

The spectrogram representation is one of the four signal views integrated into our multi-view classifier for speech emotion recognition. A spectrogram is a visual representation of the frequency content of a signal as it varies with time, obtained by the Short-Time Fourier Transform (STFT), a time-dependent adaptation of the Discrete Fourier Transform (DFT) (Kozieł et al., 2024).

The Discrete Fourier Transform of a Finite Discrete-Time Signal $x(n)$where $n = 0,1,\dots,N-1$transforms it from the time domain to the frequency domain, and is expressed mathematically as Eq. 1 (Powroźnik et al., 2021):

$$X(k) = \sum_{n=0}^{N-1} x[n]\, e^{-j2\pi\frac{kn}{N}}, \qquad k = 0,1,\dots,N-1 \tag{1}$$

where $X(k)$is the frequency domain representation, and $N$is the total number of samples in the analyzed segment. The resulting complex-valued $X(k)$encodes both amplitude and phase information of the frequency components.

To capture the temporal evolution of the frequency content, the STFT is applied. The STFT divides the speech signal into short overlapping segments, applies a suitable windowing function $w(n)$(in this case, a Hamming window) to each segment and computes the DFT separately for each windowed segment. Mathematically, the STFT of the signal $x(n)$is given as equation 2 (Allen & Rabiner, 1977):

$$X(m,k) = \sum x(n)w(n-mR)e^{-j2\pi\frac{kn}{N}} \tag{2}$$

where $m$indexes the discrete time frame, $R$is the hop size between neighboring frames, and $k$indexes the frequency bins.

The spectrogram $S(m,k)$ is obtained by computing the squared magnitude of the STFT, effectively discarding phase information and providing the distribution of energy across frequencies and time frames (Eq. 3):

$$S(m,k) = |X(m,k)|^2 \tag{3}$$

3.3. Mel-frequency cepstral coefficients

One of the signal representations used in our multi-view speech emotion recognition model is based on MFCC, which provides a perceptually relevant characterization of speech signals. The MFCC representation is derived from the short-term power spectrum of speech, modified by a perceptually motivated frequency

distortion known as the Mel scale, which reflects the human auditory perception mechanism (Abdul & Al-Talabani, 2022).

To compute MFCCs (Equation 4), a given speech frame $x(n)$ first undergoes a windowing operation (usually a Hamming window) to mitigate boundary effects, resulting in windowed frames $x_w(n)$:

$$x_w(n) = x(n) \cdot w(n) \qquad 0 \le n \le N-1 \tag{4}$$

where $w(n)$is the window function, typically defined as in Equation 5:

$$w(n) = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right) \tag{5}$$

The next step is to calculate the Discrete Fourier Transform of each windowed frame to obtain the power spectrum $X(k)$ Eq. 1. Then, a set of triangular Mel-scaled filter banks $H_m(k)$ is applied to the power spectrum to generate the Mel-spectrum coefficients Eq. 6:

$$S(m) = \sum_{k=0}^{N-1} |X(k)|^2 H_m(k), \qquad m = 1,2,\dots,M \tag{6}$$

where the filter banks are spaced according to the Mel scale f_{mel}, defined as Eq. 7:

$$f_{mel} = 2595log_{10}\left(1 + \frac{f}{700}\right) \tag{7}$$

with f being the linear frequency (in Hz). This nonlinear transformation reflects human ear perception, emphasizing lower frequencies.

Finally, the MFCC coefficients $c(l)$ are computed by applying the Discrete Cosine Transform on the logarithmic Mel spectrum coefficients $S(m)$ defined by Eq. 8:

$$c(l) = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} log[S(m)] cos\left[\frac{\pi l}{M}\left(m - \frac{1}{2}\right)\right], \qquad l = 1,2,\dots,L \tag{8}$$

Typically, only the first 13 coefficients are retained, as higher order coefficients represent rapid spectral changes and often carry less meaningful information about speech emotion characteristics (Zheng et al., 2015).

3.4. Wavelet scalograms

A wavelet scalogram is a two-dimensional energy map that illustrates how the spectral content of a signal varies over time. It is obtained from the continuous wavelet transform (CWT), which projects a real-valued speech waveform $x(t)$ onto a family of scaled and translated copies of a mother wavelet $\psi(t)$.

The CWT coefficient at scale $a > 0$ and translation $b \in R$ is defined by Eq. 9.

$$W_x(a,b) = \frac{1}{\sqrt{|a|}} \int_{\infty}^{\infty} x(t)\ \psi^*\left(\frac{t-b}{a}\right) dt \tag{9}$$

where $\psi^*(\cdot)$denotes complex conjugation, and the factor $1/\sqrt{|a|}$ equalizes energy across scales (Mallat, 2009). Small values of $a$analyzes high-frequency, short-duration phenomena, while large emphasizes slowly varying, low-frequency structures, giving the CWT its multiresolution capability. The scalogram itself is the squared modulus of the CWT coefficients (Eq. 10):

$$P_x(a,b) = |W_x(a,b)|^2 \tag{10}$$

Equations (9)-(10) describe the continuous time CWT for a waveform $x(t)$. In practice, all corpora are resampled to 16 kHz and processed as discrete-time sequences $x[n] = x(nT_s)$ with $T_s = 1/16{,}000s$. The implemented CWT therefore uses discrete scales ak and translations $b_m = mH$ (skip Hin samples), yielding coefficients $W_x[a_k, b_m]$ that populate the scalogram image. This aligns the notation with Eq. (1), where the STFT/DFT is assumed to be x_n. The continuous forms are retained for clarity. All downstream operations use their standard discrete counterparts.

For a digitally sampled signal at rate f_s, scale is converted to a pseudo-frequency $f = \frac{f_c}{a}$, with f_c the centre frequency of the chosen mother wavelet (Powrożnik et al., 2021). Discrete scales a_m $(m = 1, \dots, M)$ and translations $b_n = \frac{nR}{f_s}, n = 1, \dots, N$, where R is the hop size, yield a matrix representation $\mathrm{P_{m,n}} = P_x(a_m, b_n) \in R^{M \times N}$, whose entries become pixel intensities in the scalogram image.

3.5. Fuzzy-transform images

A fuzzy transform (F-T) image encodes the temporal evolution of low-dimensional fuzzy coefficients that approximate a signal by a fuzzy partition of its domain. Let $x(t)$ is a real-valued speech waveform defined on a compact interval $\Omega \subset R$(e.g. a short time frame) and let $\mathcal{A} = \{A_1, \dots, A_K\}$ is a fuzzy partition of Ω.

Each fuzzy set A_k is characterised by a membership function (Eq. 11):

$$\mu_k(t) = A_k(t): \Omega \longrightarrow [0,1], \qquad k = 1, \dots, K, \tag{11}$$

Satisfies the α- normalization and Ruspini conditions (eq. 12):

$$\sum_{k=1}^{K} \mu_k(t) = 1, \qquad \forall\, t \in \Omega \tag{12}$$

In practice, triangular or other "hat" functions are chosen as prototype membership functions for simplicity and local support (Eq. 13):

$$\mu_k(t) = \max\left\{0, 1 - \frac{|t - c_k|}{h}\right\}, \tag{13}$$

where c_k is the center of $A_k$and $h$is half the base width, ensuring overlap with neighboring fuzzy sets.

The k-th fuzzy coefficient of $x(t)$is obtained as the membership-weighted average defined in equation 14:

$$F_k = \frac{\int_\Omega x(t)\mu_k(t)dt}{\int_\Omega \mu_k(t)dt}, \quad k = 1, \dots, K \tag{14}$$

which can be interpreted as a localized, smoothed sample of the signal (Perfilieva, 2006). For a uniformly sampled frame $\{x_n\}_{n=1}^{N}$ with sample period $\Delta t$the discrete form is as in equation 15:

$$F_k = \frac{\sum_{n=1}^{N} x_n \mu_k(t_n)}{\sum_{n=1}^{N} \mu_k(t_n)}, \tag{15}$$

where $t_n = n\Delta t$.

For each STFT-like frame of length N, the K-dimensional vector $\mathrm{F^{(m)}} = \left[F_1^{(m)}, \dots, F_K^{(m)}\right]^{\mathsf{T}}$ is computed. Stacking these column-wise over successive frames $m = 1, \dots, M$ produces a matrix $\mathrm{F} = \left[\mathrm{F^{(1)}}\ \mathrm{F^{(2)}}\ \cdots\ \mathrm{F^{(M)}}\right] \in R^{K \times M}$, whose entries are linearly rescaled to ($[0,255]$) and rendered as a grey-scale bitmap.

This F-transform image becomes the fourth parallel input of K4F-Net, complementing spectrograms, mel-spectrum maps, and wavelet scalograms. Because fuzzy coefficients integrate local information through overlapping membership functions, the image captures smooth temporal-spectral trends that have proven discriminative in speech emotion tasks, while exhibiting robustness to noise and speaker variability.

The membership-driven locality of the F-transform provides tunable resolution: narrower support ($h! \downarrow$)emphasize rapid energy fluctuations associated with high arousal, while broader supports emphasize slower prosodic drifts associated with low arousal or neutral affect. These complementary features, when fused with the other three representations in the cross-fusion layers, significantly enhance recognition accuracy.

3.6. Kronecker convolution

Kronecker Convolution (KC) extends the receptive field of a standard convolution without introducing additional learnable parameters by factorizing the kernel by a fixed Kronecker product. Let the input feature map be $X \in R^{C_{in} \times H \times W}$ and the learnable base kernel is $G \in R^{C_{out} \times C_{\mathrm{in}} \times k \times k}$.

A standard $k \times k$ convolution produces the output activation at spatial index $p \in Z^{\mathbb{2}}$ as Eq. 16 (Patro et al., 2023):

$$Y(p) = \sum_{c=1}^{C_{\text{in}}} \sum_{q \in \mathcal{R}_{k}} G_{c}^{c}(q) X_{c}^{c}(p-q), \tag{16}$$

where $\mathcal{R}_{k} = \{-\lfloor k/2 \rfloor, \ldots, \lfloor k/2 \rfloor\}^2$ indexes the local receptive field.

Kronecker convolution replaces the original kernel by (Eq. 17):

$$W^{\text{K}} = G \otimes T_{r,s} \tag{17}$$

where "$\otimes$'' denotes the Kronecker product and $T_{r,s} \in \{0,1\}^{(rs)\times(rs)}$ is a fixed binary transformation matrix that expands G both geometrically and structurally (Wu et al., 2019):

$$T_{r,s} = \begin{pmatrix} J_s & 0 & \cdots & 0 \\ 0 & J_s & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & J_s \end{pmatrix}, \qquad J_s = 1_{s\times s},$$

with $J_s$one $s \times s$All-One Block, 0 is a zero block of the same size, $r \in N^{+}$is the inter-dilatation factor that controls the spacing between blocks, and $s \in N^{+}$is the intra-sharing factor that controls the size of each block.

The effective kernel size thus increases from $k$to $k' = kr$which gives a receptive field identical to that of an atrous (dilated) convolution with rate $r$but the number of parameters remains $C_{\text{out}}C_{\text{in}}k^2$ because $T_{r,s}$ is not trainable.

Using the extended kernel, the KC output is as in equation 18:

$$Y(p) = \sum_{c=1}^{C_{\text{in}}} \sum_{q \in \mathcal{R}_{k'}} \text{b}W_{c}^{\text{KC}}(q) X_{c}^{c}(p-q), \tag{18}$$

where non-zero entries W^{K} are divided into $s \times s$sub-regions, allowing the operator to capture both long-range context (via r) and dense local detail (via s). The proportion of input locations that contribute to each output, called the Valid-Feature-Ratio (VFR), is $\text{VFR}(r,s) = \frac{s^2}{(RS)^2} = \frac{1}{r^2}$, which exceeds that of a dilated convolution at the same rate ($s = 1$) for any $s > 1$to reduce information loss in sparse sampling. The setting $(r,s) = (1,1)$ recovers the standard convolution, while $(r > 1, s = 1)$degenerates to atrous convolution. Kronecker convolution thus generalizes both operators within a unified parameter-efficient formulation.

3.7. K4F-Net architecture

The proposed network ingests four synchronized time-frequency images extracted from a multi-speaker signal: a wavelet scalogram, an STFT spectrogram, a Mel-frequency spectrogram, and a fuzzy transform map. Each modality is resampled to the common spatial resolution $R^{1\times224\times224}$ and standardized. A dedicated representation branch processes each image with three successive Kronecker Convolution Blocks (KCBs) whose purpose is to enlarge the receptive field without increasing the number of parameters. The entire architecture is summarized in Fig. 1 and detailed in Table 2 (per-branch flow) and Table 3 (shared trunk).

Each branch uses three KCBs with channel widths of $(1 \to 32 \to 64 \to 128)$. The first two blocks use an intermediate dilation factor of $r = 2$while the third uses $r = 1$. An internal division factor $s = 2$is kept constant. Each KCB ends with a 2×2max pooling layer, so the spatial resolution shrinks from $224 \times 224$to 112×112then $56 \times 56$and finally 28×28. The resulting tensor of each branch is therefore denoted by $Z_i \in R^{128x28x28}, i \in \{1,\ldots,4\}$.

Tab. 2. The general structure of view processing blocks. KC denotes Kronecker convolution

Layer No	Layer/Block	Main Params	Output (single branch)
1.	KC-Conv2D	filters no: 32, kernel: 3×3, $r = 2$, $s = 2$, $SiLU$	$224 \times 224 \times 32$
2.	KC-Conv2D	filters no: 32, kernel: 3×3, $r = 2$, $s = 2$, $SiLU$	$224 \times 224 \times 32$
3.	MaxPooling2D	2×2	$112 \times 112 \times 32$
4.	KC-Conv2D	filters no: 64, kernel: 3×3, $r = 2$, $s = 2$, $SiLU$	$112 \times 112 \times 64$
5.	KC-Conv2D	filters no: 64, kernel: 3×3, $r = 2$, $s = 2$, $SiLU$	$112 \times 112 \times 64$
6.	MaxPooling2D	2×2	$56 \times 56 \times 64$
7.	KC-Conv2D	filters no: 128, kernel: 3×3, $r = 2$, $s = 2$, $SiLU$	$56 \times 56 \times 128$
8.	KC-Conv2D	filters no: 128, kernel: 3×3, $r = 2$, $s = 2$, $SiLU$	$56 \times 56 \times 128$
9.	MaxPooling2D	2×2	$28 \times 28 \times 128$

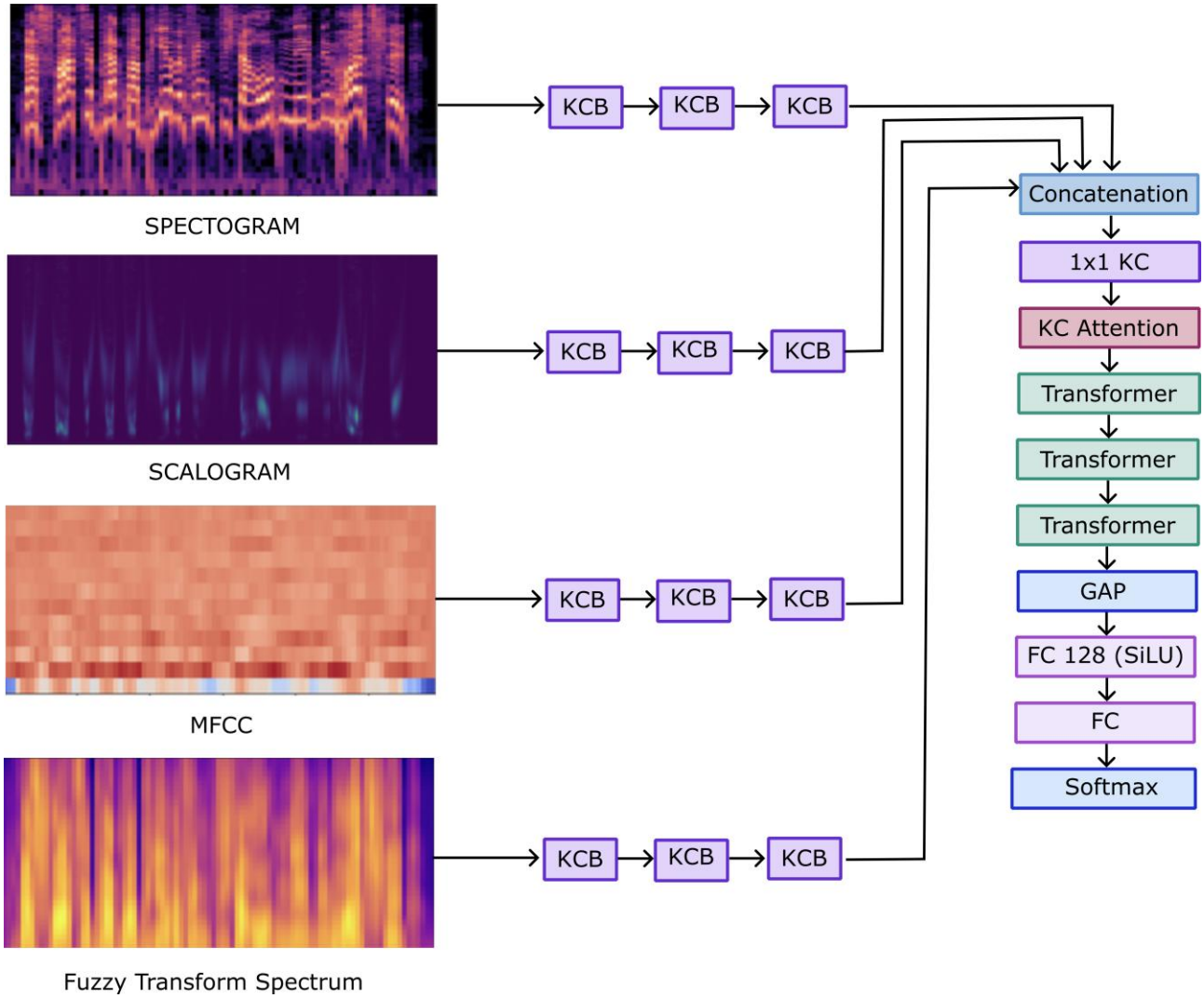

Fig. 1. The overall architecture of proposed K4F-Net

Tab. 3. The general structure of concatenated views processing blocks. KC denotes Kronecker convolution

Layer No	Layer/Block	Main Params	Output (after fusion)
10.	Concatenate	-	$28 \times 28 \times 512$
11.	KC-Conv2D	filters no: 512, kernel: $1 \times 1, r = 2, s = 2,$ $SiLU$	$28 \times 28 \times 512$
12.	KC - Attn	8 heads	$28 \times 28 \times 512$
13.	Transformer	3 blocks, $d = 512$, 8 heads	$28 \times 28 \times 512$
14.	GlobalAvgPooling2D	-	512
15.	Dense	$512 \rightarrow 128, SiLU$	128
16.	Dense	$512 \rightarrow Class\ no, SoftMax$	

In Table 1 and Table 2 KC stands for Kronecker Convolution. Effective kernel size $k \cdot r \times k \cdot r$is derived from the base $k \times k$ by the Kronecker product with the matrix $T_{r,s}$. The number of learning weights remains the same $C_{out}C_{in}k^2$.

The four tensors are concatenated along the channel axis to give $Z \in R^{512\times H'\times W'}$and are then represented by a 1×1 Kronecker Convolution in Queries Q, keys and values $V$of dimension $d = 512$. Self-attention augmented with Kronecker structure is performed as in Eq. 19:

$$KC - Attn(Q, K, V) = SoftMax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right) \otimes T_{2,2V} \tag{19}$$

so that the local neighbourhood cues provided by $T_{2,2}$ to preserve fine-grained detail while attention weights model long-range intermodal relationships.

Three stacked encoder blocks follow, each consisting of layer normalization, eight-headed KC attention (KC-Attn) of width $d = 512$, a position-wise feed-forward network of size $2d$ and remaining links. The Kronecker pattern embedded in the projection matrices maintains the low parameter number, but the receptive field, already expanded by the previous KCBs, now covers the entire joint feature map.

Global spatial averaging reduces the tensor to a 512-dimensional vector. A dropout layer with probability 0.2 mitigates co-adaptation. A fully connected layer maps $512 \rightarrow 128$ with Swish enabled, and the last layer projects $128 \rightarrow Class\ No$ followed by a SoftMax that estimates posterior probabilities.

3.8. Language-independent adversarial learning

The multi-view front-end described in Sections 3.2 - 3.6 yields a joint feature tensor $\mathbf{H} \in R^{d\times 28\times 28}$ which encodes spectral-temporal cues from four complementary signal images. Although these cues are effective for emotion recognition, they still carry language-specific idiosyncrasies that interfere with cross-linguistic generalization. To suppress such disruptive information, we employ an adversarial strategy analogous to domain-adversarial neural networks (Ganin & Lempitsky, 2014; Xia et al., 2019).

The Shared Feature Extractor $\mathcal{F}_\theta$includes all layers up to and including the transformer encoder. Two task-dependent heads are connected in parallel: an emotion classifier $\mathcal{C}_\phi^{emo}$with SoftMax output of size $K_{\text{emo}} =$ 5and a language discriminator $\mathcal{C}_\psi^{long}$ with SoftMax Size $K_{\text{long}} = 4$. The latter is preceded by a gradient reversal layer $\mathcal{R}_\lambda$ which multiplies the backward signal by $-\lambda < 0$ during parameter updates, thereby inducing an adversarial objective without changing the forward pass.

To construct a cost function, optimization objectives must be defined. Let $(\boldsymbol{x}, y^{emo}, y^{long})$ denotes an input batch with an emotion label $y^{emo}$and language label y^{long}. The two cross entropy losses are defined by Eq. 20 and Eq. 21:

$$\mathcal{L}_{emo} = -\sum_{k=1}^{K_{\text{emo}}} 1[y^{\text{emo}} = k] \log \mathcal{C}_\phi^{emo}\left(\mathcal{F}_\theta(x)\right)_k, \tag{20}$$

$$\mathcal{L}_{lang} = -\sum_{\text{l}=1}^{K_{\text{lang}}} 1\left[y^{\text{lang}} = \text{l}\right] \log \mathcal{C}_\psi^{lang}\left(\mathcal{R}_\lambda \circ \mathcal{F}_\theta(x)\right)_\text{l}. \tag{21}$$

The overall objective to be minimized is therefore $\mathcal{L}_{total} = \mathcal{L}_{emo} + \beta\mathcal{L}_{long}$ where $\beta > 0$ controls the trade-off between emotion fidelity and language invariance. Because the gradient of $\mathcal{L}_{long}$ is reversed when

it reaches θthe extractor $\mathcal{F}_\theta$ is trained to maximize the error of the discriminator, i.e. to produce features that disguise language identity, while ψtries to minimize the same error.

The resulting min-max game converges to a saddle point where $\mathcal{F}_\theta$ retains information essential for emotion prediction, but discards language-dependent artifacts, so that $\mathcal{C}_\phi^{emo}$ generalizes to English, German, Polish and Danish.

To ensure the fastest possible convergence of the model, we empirically set $\lambda = 1$ and anneal βoff 0.0 to 0.5according to schedule $\beta_t = 0.5\left(1 + \cos\left(\frac{\pi t}{T}\right)\right)$ over the first $T = 10$ epochs, after which it remains constant. This warm-start stabilizes learning in the early iterations and yielded the highest average weighted accuracy on the evolutionary folds.

The adversarial branch is purely auxiliary and does not alter the forward path of the emotion classifier. Consequently, it is fully compatible with the multi-view Kronecker convolutional backbone, the cross-fusion mechanism, and the transformer coder described in the previous sections. All tensor forms remain unchanged: $H = \mathcal{F}_\theta(x)$ has shape $512 \times 28 \times 28$ before global averaging, exactly as in Table 2, and the additional parameters $\mathcal{C}_\psi^{long}$ increase the total footprint by less than 1%.

3.9. Evaluation metrics

The effectiveness of the proposed classifier is evaluated using four class-level measures that are robust to label imbalance (Powers, 2020; Sokolova & Lapalme, 2009). Let us $\hat{y}_i \in \{1, \ldots, K\}$denote the predicted emotion and $y_i$is the ground truth for sample $i = 1, \ldots, N$ then true positive (TP_c), false positive (FP_c), false negative (FN_c) can be defined as Eq. 22, Eq. 23, Eq. 24, n_c(Eq. 21) denotes the number of all examples in the test set whose ground truth belongs to class *c*:

$$n_c = \sum_{i=1}^{N} 1[y_i = c], \tag{21}$$

$$TP_c = \sum_{i=1}^{N} 1[y_i = c]\, 1[\hat{y}_i = c], \tag{22}$$

$$FP_c = \sum_{i=1}^{N} 1[y_i \neq c]\, 1[\hat{y}_i = c], \tag{23}$$

$$FN_c = \sum_{i=1}^{N} 1[y_i = c] 1[\hat{y}_i \neq c]. \tag{24}$$

Weighted Accuracy (WA) gives the proportion of correctly classified instances while respecting the natural class priors Eq. 25:

$$\mathrm{WA} = \frac{\sum_{c=1}^{K} \mathrm{TP_c}}{\sum_{c=1}^{K} \mathrm{n_c}} \tag{25}$$

Macro Precision and Macro Recall are obtained by first calculating the per-class values defined in Equations 26 and 27:

$$Prec_c = \frac{TP_c}{TP_c + FP_c} \tag{26}$$

$$Rec_c = \frac{TP_c}{TP_c + FN_c} \tag{27}$$

and then averaged uniformly over the K classes:

$$Precision_{\mathrm{macro}} = \frac{1}{K} \sum_{c=1}^{K} Prec_c \tag{28}$$

$$Recall_{\mathrm{macro}} = \frac{1}{K} \sum_{c=1}^{K} Rec_c \tag{29}$$

Macro F1 is the harmonic mean of the two macro quantities as defined in Equation 30:

$$F1_{\mathrm{macro}} = 2\,\frac{Precision_{\mathrm{macro}} * Recall_{\mathrm{macro}}}{Precision_{\mathrm{macro}} + Recall_{\mathrm{macro}}} \tag{30}$$

While Equation 25 reflects overall operational accuracy in mission scenarios, the macro metrics Equations 26 - 30 weigh each emotion equally and therefore highlight performance in minority conditions.

4. EXPERIMENTS AND RESULTS

All experiments follow a leave-one-language-out protocol in which three languages provide the training material, while the fourth language forms a strictly unseen test set. The procedure is repeated four times, so that English, German, Polish, and Danish each serve as the target language once. Within each training batch, the data is divided by speaker into 80% training and 20% development subsets. Mini-folds contain 16 utterances; each utterance is divided on-the-fly into the four 224×224 images as described in Section 3 and extended with the scheme of Section 3.1. The model is optimized with AdamW (initial learning rate 3×10^{-4} Weight Decay 10^{-4}) and cosine annealing decay. Training stops after ten epochs with no loss or improvement in accuracy. Otherwise it continues for 100 epochs. All runs are performed under PyTorch~2.1 on Ubuntu~22.04 with an Intel i9-13900K CPU, 64GB RAM, and a single NVIDIA RTX 4070 GPU. Performance is reported using the metrics described in Section 3.9.

4.1. Cross-language performance

Table 4 shows the cross-lingual results obtained with the proposed K4F-Net when each of the four languages is omitted in turn and evaluated as an unseen target. The metric values are macro-averaged across the emotion classes so that each class contributes equally, regardless of its prior frequency in the corpus. Since different emotions occur in all sets (datasets A-D), the classes used for testing are those that occur in the test set and at least one training set. In all four folds, the network achieves an accuracy above 95%, while maintaining a balanced trade-off between precision and recall. The average weighted accuracy over the four folds is 96.3%.

Tab. 4. Cross-language emotion recognition performance of K4F-Net. The model is trained on three languages and evaluated on the fourth. Numbers are macro-averaged across the tested classes and expressed in %

Target language	Accuracy	Precision	Recall	F1-score
English	95.9	95.4	95.1	95.2
German	96.8	96.3	96.6	96.4
Polish	97.1	96.8	97.2	97.0
Danish	95.6	95.0	95.3	95.1
Mean	**96.3**	**95.9**	**96.1**	**95.9**

A companion experiment with a parameter-matched four-branch ResNet-34 serves as the main baseline. As summarized in Table 5, K4F-Net delivers consistent gains for each language, with an average improvement of 4.8 percentage points in accuracy and 4.2 points in macro-F1. The largest accuracy gain, +6.1%, occurs when Polish is the target language, supporting the hypothesis that multi-view representation and Kronecker cross-fusion benefit highly inflected languages with large prosodic variation.

Our leave-one-language-out protocol operates on the intersection of available emotion labels between the training set and the held-out target (Table 3). Thus, English and Danish folds (with 4-5 emotions available for testing) are evaluated on a smaller label set than Polish and German folds (6-7 emotions). Despite this variance, accuracy remains above 95% for all targets, suggesting that K4F-Net's gains are not an artifact of a simpler label set. Furthermore, German and Danish, both Germanic languages, show no systematic advantage when either is the target: German achieves 96.8% accuracy, Danish 95.6%. This suggests that our adversarial language regularization and cross-fusion reduces reliance on family-specific elements rather than memorization of Germanic features.

Tab. 5. Performance gap (Δ) between K4F-Net and the ResNet-34 baseline. Positive numbers indicate an improvement in favour of K4F-Net

Target language	Δ Accuracy	Δ Precision	Δ Recall	Δ F1-score
English	+4.6	+4.1	+4.3	+4.2
German	+4.4	+4.0	+3.9	+4.0
Polish	+6.1	+5.7	+5.9	+5.8
Danish	+4.1	+3.6	+3.8	+3.5
Mean	**+4.8**	**+4.3**	**+4.5**	**+4.2**

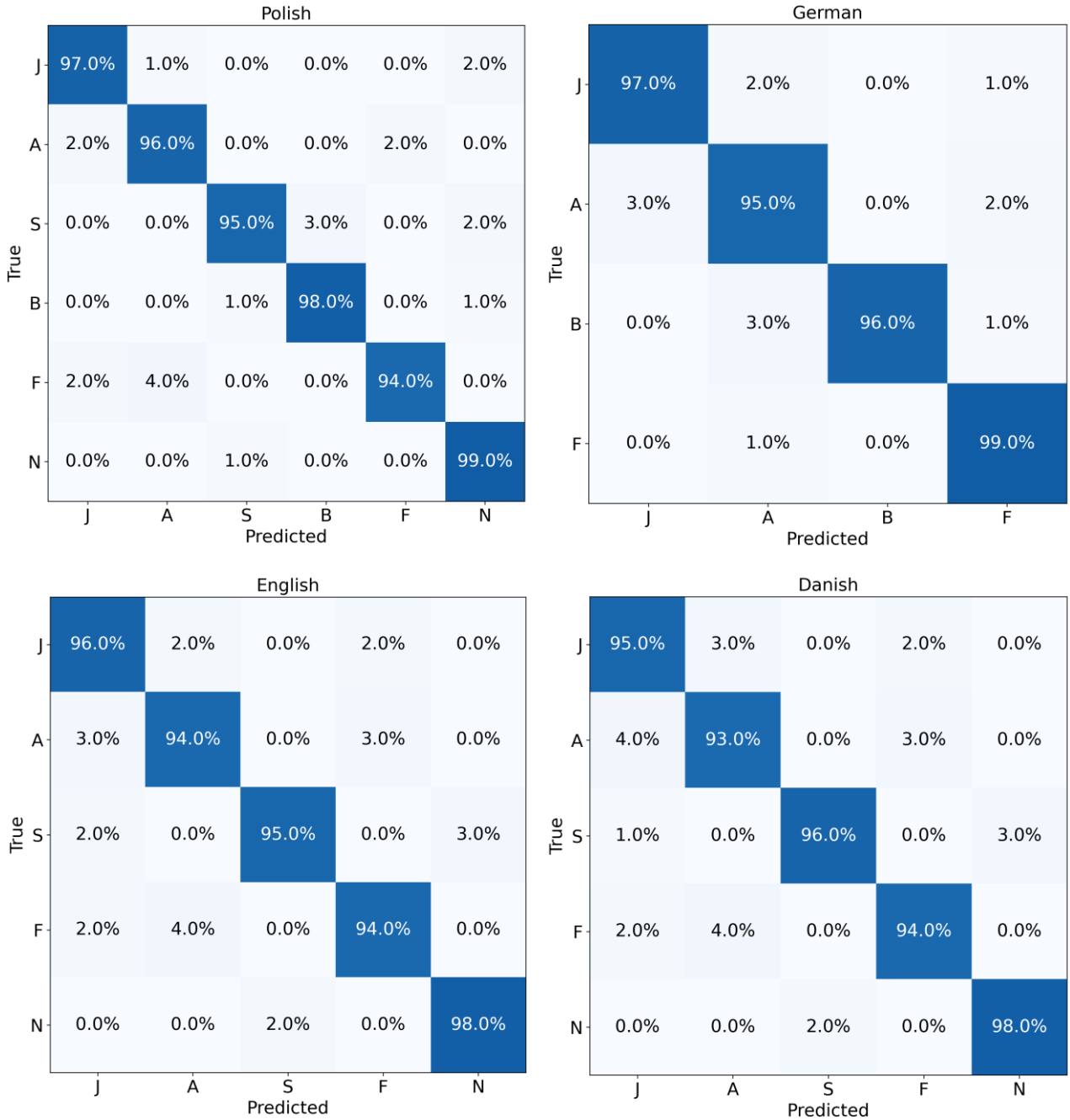

Fig. 2. The confusion matrices for proposed classifier for 4 databases.
J – joy, A – anger, S – sadness, F – fear, B - boredom N – neutral

Inspection of the confusion matrices in Fig. 2 shows that the pattern of residual errors is strongly language dependent, but broadly consistent with the acoustic similarity between anger (A) and fear (F). For Polish, English and Danish, between 3% and 4% of fear tokens are mistaken for anger, while the reciprocal error never exceeds 4%. The German fold is less affected, with only a1% of fear being mistaken for anger. Aside from

this pair, the most notable confusions are boredom (B) misread as anger in German 3% and sadness (S) drifting toward neutral in Danish 3%.

4.2. Ablation study

To quantify the importance of each component, a series of ablation experiments are performed on the polished fold. The results are shown in Table 6. Replacing each Kronecker convolution with a standard 3×3kernel of the matched receptive field, but larger parameter counts reduce the weighted accuracy from 96.3% to 90.5%. Removing the fuzzy transform branch while keeping the other three modalities yields 91.4% WA ($\Delta = -4.7\%$). Disabling cross-fusion KC attention reduces performance to 92.3%, confirming that intermodal interactions cannot be mimicked by simple concatenation. Removing the scalogram layer decreases weighted accuracy by 3.2% and increases language discriminator accuracy to 83%, demonstrating that the adversary's head is indeed erasing language-specific cues. Finally, replacing the Kronecker convolutions with standard convolutions of the same receptive span increases the parameter budget by 47%, but still lags behind the WA by 1.9%, highlighting the superior efficiency of the Kronecker operations.

Tab. 6. Ablation study on the Polish-held-out fold. WA – weighted accuracy; Δ – absolute change with respect to the full K4F-Net; #P – number of trainable parameters.

Variant	#P [M]	WA [%]	Δ [%]
K4F-Net (full)	5.1	97.1	-
with standard 3×3 convolution	7.5	91.5	-5.6
without F-transform branch	4.8	92.4	-4.7
without scalogram branch	4.8	93.1	-4.0
without mel-spectrogram branch	4.8	91.2	-5.9
without spectrogram branch	4.8	92.7	-4.4
without KC cross-attention	5.0	92.9	-4.2

The comparative evaluation in Table 7 assesses how well the proposed lightweight multi-view approach performs relative to a broad set of speech-emotion recognition baselines. Most of the baselines exceed 80% accuracy, but their effectiveness varies significantly with the corpus, the chosen signal representation, and the underlying network topology. ViT models benefit from global self-attention, which captures long-range time-frequency relationships, resulting in strong overall performance. Self-attention is particularly valuable for SER because it can explicitly model the spatial variations that encode subtle affective cues.

The proposed K4F-Net achieves a weighted accuracy of over 96%, outperforming all competing methods listed in Table 6. Its advantage stems from the complementary fusion of four orthogonal feature domains: fuzzy transform energy maps, discrete wavelet scalograms, complex STFT spectrograms, and Mel cepstral coefficients, combined with parameter-efficient Kronecker convolutions and cross-modal self-attention. Within the current state of the art, K4F-Net thus offers one of the most effective and computationally economical solutions for robust, language-independent speech emotion recognition.

The studies in Table 6 use different datasets (different emotion sets, languages, and recording conditions), so a direct comparison of absolute values may not be meaningful. When evaluated on a common dataset and protocol, attention-based architectures (e.g., ViT/BEiT variants) reliably exploit long-range temporal-spectral dependencies, whereas conventional CNNs often underperform on categories characterized by subtle or low-saliency spectral cues. Our results complement this trend by showing that multi-view fusion (spectrogram, MFCC, wavelet, fuzzy) plus Kronecker cross-fusion closes the gap without inflating parameters, and that the gains are largest when the target language belongs to the group where accent plays an important role in conveying information (e.g., Polish). We emphasize that Table 6 should be read as evidence for families of methods rather than as a table of corpora results, and we therefore provide a corpus-controlled comparison with a parameter-matched four-branch net (Table 4), where K4F-Net averages +4.8 pp of accuracy.

Tab. 7. Comparison with the-state-of-the-art

Model	Signal transform	Dataset	Accuracy [%]	Reference
ViT	log-Mel spectrogram	CREMA-D	39.02	(Kim & Lee, 2025)
	Mel spectrogram	RAVDESS	97.49	(Mishra et al., 2025)
		CREMA-D	72.06	
		ESD	95.84	
		MELD	49.83	
	Mel frequency	GTZAN, FMA	56.85	(Khasgiwala & Tailor, 2021)
BEiT	Mel spectrogram	RAVDESS	94.62	(Mishra et al., 2025)
		CREMA-D	71.85	
		ESD	96.25	
		MELD	43.32	
l-ViT	Mel spectrogram	EMODB	91.03	(Akinpelu et al., 2024)
		TESS	98.00	
Wav2.0	feature extractor	RAVDESS	98.05	(Luna-Jiménez et al., 2021)
SepTr	spectrogram	CREMA-D	70.47	(Ristea et al., 2022)
		SCV2	98.51	
		ESC-50	91.13	
CvT	linear	Emo-DB	96.99	(Echim et al., 2024)
		Emo-IIT	97.75	
	Mel spectrogram	Emo-DB	97.38	
		Emo-IIT	96.57	
	CQT	Emo-DB	97.08	
		Emo-IIT	97.63	
	MFCC	Emo-DB	96.18	
		Emo-IIT	96.43	
CoordViT	spectrogram	CREMA-D	82.96	(Kim & Lee, 2023)
CCT	spectrogram	Emo-DB	55.84	(Arezzo & Berretti, 2022)
		EMOVO	37.36	
		SAVEE	29.47	
MLP, k-NN, Decision Trees, Naive Bayes, Random Forest, Probabilistic Neural Network, Fuzzy Rule Classifier,	scalograms	EMO-DB, DES, Polish corpus, English corpus	62-94	(Powroźnik et al., 2021)
Our	4 features	EMO-DB, DES, Polish corpus, English corpus	96.30	

In machine learning approaches, four main types of fusions are widely used. The first is data fusion, which combines different types of data from different modalities as input to a single model (e.g., images and text). Model fusion is the second type where different models are combined to improve accuracy and generalization, usually by applying ensemble learning techniques. In this case, the training of this model is longer than that of the single model. Feature fusion is the third type of fusion, which aims to improve the ability to learn complex patterns using the same or different types of input data. Decision fusion is the final type where the outputs of different models are combined to make a final classification/decision. This method involves various types of averaging or voting mechanisms.

A lightweight, attention-based interaction layer is placed after each modality branch (spectrogram, MFCC, scalogram, and fuzzy) and before the common trunk. The four branch tensors are concatenated and projected (via 1×1 Kronecker conv) into Q/K/V, then a multi-head KC attention mixes information across modalities so that features from one view can content-adaptively enhance/suppress features in another. Unlike data fusion,

cross-fusion does not merge raw inputs. It operates on learned feature maps according to per-modality encoders, where the representations are cleaner and more aligned in size, so the interactions are more meaningful. In this case, we cannot talk about model fusion. There is one model, not many. No voting/averaging of separate predictors. Cross-fusion learns intermodal communication within a single network, keeping training/inference compact. Unlike standard feature fusion, cross-fusion is dynamic: attention weights depend on the current signal, capturing pairwise and higher-order relationships between modalities. In contrast to decision fusion, cross-fusion is applied before classification, forming a single, richer representation. There is no late voting. The classifier sees a fused tensor influenced by content-aware cross-modal interactions.

5. CONCLUSIONS

This paper presents K4F-Net, a compact and language-robust framework for speech-emotion recognition that processes: STFT spectrograms, Mel-frequency cepstral maps, wavelet scalograms, and fuzzy transform images in four parallel branches. Kronecker convolutions extend the receptive field of standard kernels at zero additional parameter cost, while a cross-fusion self-attention module merges complementary cues before a lightweight Transformer encoder captures long-range context. A gradient reversal head further regularizes the feature space towards language independence.

Experiments on four publicly available corpora covering Polish, English, German and Danish show that K4F-Net achieves a mean weighted accuracy of 96.3% under a fourfold leave-one-language-out protocol, outperforming a size-matched ResNet-34.

Ablation results confirm the importance of the proposed design choices: eliminating Kronecker kernels, the fuzzy transform spectrum, or the cross-fusion block reduces weighted accuracy by 4-6%. The language-adversarial loss reduces discriminator accuracy to chance level, suggesting an effective removal of language-specific artifacts without compromising emotion recognition.

Future work will explore self-supervised pre-training on unlabeled multilingual corpora to further improve cross-domain robustness. It can also be extended to spontaneous and noisy conversational speech. Multimodal fusion with facial and linguistic cues for richer affective understanding can also be explored. The results are very encouraging and suggest that parameter-efficient Kronecker convolutions together with multi-view representations open a very promising direction for building real-time, language-agnostic SER systems.

Conflicts of interest

The authors declare no conflict of interest.

REFERENCES

Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(10), 1533–1545. https://doi.org/10.1109/TASLP.2014.2339736

Abdul, Z. Kh., & Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, *10*, 122136–122158. https://doi.org/10.1109/ACCESS.2022.3223444

Ahn, C.-S., Rana, R., Busso, C., & Rajapakse, J. C. (2025). Multitask transformer for cross-corpus speech emotion recognition. *IEEE Transactions on Affective Computing*, *16*(3), 1581-1591. https://doi.org/10.1109/TAFFC.2025.3526592

Akinpelu, S., Viriri, S., & Adegun, A. (2024). An enhanced speech emotion recognition using vision transformer. *Scientific Reports*, *14*, 13126. https://doi.org/10.1038/s41598-024-63776-4

Allen, J. B., & Rabiner, L. R. (1977). A unified approach to short-time Fourier analysis and synthesis. *Proceedings of the IEEE*, *65*(11), 1558–1564. https://doi.org/10.1109/PROC.1977.10770

Arezzo, A., & Berretti, S. (2022). SPEAKER VGG CCT: Cross-corpus speech emotion recognition with speaker embedding and vision transformers. *4th ACM International Conference on Multimedia in Asia* (pp. 1-7). Association for Computing Machinery. https://doi.org/10.1145/3551626.3564937

Avots, E., Sapiński, T., Bachmann, M., & Kamińska, D. (2019). Audiovisual emotion recognition in wild. *Machine Vision and Applications*, *30*, 975–985. https://doi.org/10.1007/s00138-018-0960-9

Chowdhury, J. H., Ramanna, S., & Kotecha, K. (2025). Speech emotion recognition with light weight deep neural ensemble model using hand crafted features. *Scientific Reports*, *15*, 11824. https://doi.org/10.1038/s41598-025-95734-z

Chumachenko, K., Iosifidis, A., & Gabbouj, M. (2022). Self-attention fusion for audiovisual emotion recognition with incomplete data. *26th International Conference on Pattern Recognition (ICPR)* (pp. 2822–2828). IEEE. https://doi.org/10.1109/ICPR56361.2022.9956592

Chwaleba, K., & Wach, W. (2024). Polish dance music classification based on mel spectrogram decomposition. *Advances in Science and Technology Research Journal*, *19*(2), 95–113. https://doi.org/10.12913/22998624/195506

Czerwinski, D., & Powroźnik, P. (2018). Human emotions recognition with the use of speech signal of polish language. *Conference on Electrotechnology: Processes, Models, Control and Computer Science (EPMCCS)* (pp. 1–6). IEEE. https://doi.org/10.1109/EPMCCS.2018.8596404

Echim, S.-V., Smădu, R.-A., & Cercel, D.-C. (2024). Benchmarking adversarial robustness in speech emotion recognition: insights into low-resource romanian and german languages. In U. Endriss, F. S. Melo, K. Bach, A. Bugarín-Diz, J. M. Alonso-Moral, S. Barro, & F. Heintz (Eds.), *Frontiers in Artificial Intelligence and Applications* (pp. 2468–2475). IOS Press. https://doi.org/10.3233/FAIA240774

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572–587. https://doi.org/10.1016/j.patcog.2010.09.020

Engberg, I. S., & Hansen, A. V. (1996). *Documentation of the Emotional Speech Data Base, DES*. Aalborg Universitetsforlag.

Ezzameli, K., & Mahersia, H. (2023). Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, *99*, 101847. https://doi.org/10.1016/j.inffus.2023.101847

Ganin, Y., & Lempitsky, V. (2014). Unsupervised domain adaptation by backpropagation (Version 2). *ArXiv, abs/1409.7495*. https://doi.org/10.48550/ARXIV.1409.7495

George, S. M., & Ilyas, P. M. (2024). A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. *Neurocomputing*, *568*, 127015. https://doi.org/10.1016/j.neucom.2023.127015

Gong, Y., Chung, Y.-A., & Glass, J. (2021). AST: Audio spectrogram transformer (Version 3). *ArXiv, abs/2104.01778*. https://doi.org/10.48550/ARXIV.2104.01778

Hareli, S., & Hess, U. (2012). The social signal value of emotions. *Cognition and Emotion*, *26*(3), 385–389. https://doi.org/10.1080/02699931.2012.665029

Hashemi, S., & Asgari, M. (2023). Vision transformer and parallel convolutional neural network for speech emotion recognition. *31st International Conference on Electrical Engineering (ICEE)* (pp. 888–892). https://doi.org/10.1109/ICEE59167.2023.10334797

Ibrahim, A., Shehata, S., Kulkarni, A., Mohamed, M., & Abdul-Mageed, M. (2024). What does it take to generalize SER model across datasets? A comprehensive benchmark (Version 1). *ArXiv, abs/2406.09933*. https://doi.org/10.48550/ARXIV.2406.09933

Jin, Q., Li, C., Chen, S., & Wu, H. (2015). Speech emotion recognition with acoustic and lexical features. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4749–4753). IEEE. https://doi.org/10.1109/ICASSP.2015.7178872

Johanson, D. L., Ahn, H. S., & Broadbent, E. (2021). Improving interactions with healthcare robots: A review of communication behaviours in social and healthcare contexts. *International Journal of Social Robotics*, *13*(8), 1835–1850. https://doi.org/10.1007/s12369-020-00719-9

Kakuba, S., & Han, D. S. (2022). Speech emotion recognition using context-aware dilated convolution network. *27th Asia Pacific Conference on Communications (APCC)* (pp. 601–604). IEEE. https://doi.org/10.1109/APCC55198.2022.9943771

Kamaruddin, N., Wahab, A., & Quek, C. (2012). Cultural dependency analysis for understanding speech emotion. *Expert Systems with Applications*, *39*(5), 5115–5133. https://doi.org/10.1016/j.eswa.2011.11.028

Kaminska, D., Sapinski, T., & Pelikant, A. (2013). Recognition of emotional states in natural speech. *2013 Signal Processing Symposium (SPS)* (pp. 1–4). IEEE. https://doi.org/10.1109/SPS.2013.6623599

Khasgiwala, Y., & Tailor, J. (2021). Vision transformer for music genre classification using mel-frequency cepstrum coefficient. *IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)* (pp. 1–5). https://doi.org/10.1109/GUCON50781.2021.9573568

Kim, J.-Y., & Lee, S.-H. (2023). CoordViT: A novel method of improve vision transformer-based speech emotion recognition using coordinate information concatenate. *International Conference on Electronics, Information, and Communication (ICEIC)* (pp. 1–4). IEEE. https://doi.org/10.1109/ICEIC57457.2023.10049941

Kim, J.-Y., & Lee, S.-H. (2024). Accuracy enhancement method for speech emotion recognition from spectrogram using temporal frequency correlation and positional information learning through knowledge transfer. *IEEE Access*, *12*, 128039–128048. https://doi.org/10.1109/ACCESS.2024.3447770

Kim, J.-Y., & Lee, S.-H. (2025). A method for improving the accuracy of speech emotion recognition using implicitly filtered image generation in extreme noisy environments. *International Conference on Electronics, Information, and Communication (ICEIC)* (pp. 1–3). IEEE. https://doi.org/10.1109/ICEIC64972.2025.10879629

Kozieł, G., Harasim, D., Dziuba-Kozieł, M., & Kisała, P. (2024). Fourier transform usage to analyse data of polarisation plane rotation measurement with a TFBG sensor. *Metrology and Measurement Systems*, *31*(2), 2. https://doi.org/10.24425/mms.2024.149698

Latif, S., Rana, R., Khalifa, S., Jurdak, R., & Schuller, B. W. (2020). Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition (Version 3). *ArXiv, abs/2005.08453*. https://doi.org/10.48550/ARXIV.2005.08453

Liu, J., Ang, M. C., Chaw, J. K., Ng, K. W., & Kor, A.-L. (2025). Personalized emotion analysis based on fuzzy multi-modal transformer model. *Applied Intelligence*, *55*(3), 227. https://doi.org/10.1007/s10489-024-05954-5

Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., Montero, J. M., & Fernández-Martínez, F. (2021). A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset. *Applied Sciences*, *12*(1), 327. https://doi.org/10.3390/app12010327

Madanian, S., Adeleye, O., Templeton, J. M., Chen, T., Poellabauer, C., Zhang, E., & Schneider, S. L. (2025). A multi-dilated convolution network for speech emotion recognition. *Scientific Reports*, *15*, 8254. https://doi.org/10.1038/s41598-025-92640-2

Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning - A systematic review. *Intelligent Systems with Applications*, *20*, 200266. https://doi.org/10.1016/j.iswa.2023.200266

Mallat, S. (2009). *A wavelet tour of signal processing*. Elsevier.

Mishra, R., Frye, A., Rayguru, M. M., & Popa, D. O. (2025). Personalized speech emotion recognition in human-robot interaction using vision transformers. *IEEE Robotics and Automation Letters*, *10*(5), 4890–4897. https://doi.org/10.1109/LRA.2025.3554949

Mohamed, E. A., Koura, A., & Kayed, M. (2024). Speech emotion recognition in multimodal environments with transformer: arabic and english audio datasets. *International Journal of Advanced Computer Science and Applications*, *15*(3). https://doi.org/10.14569/IJACSA.2024.0150359

Motamed, S., Setayeshi, S., & Rabiee, A. (2017). Speech emotion recognition based on a modified brain emotional learning model. *Biologically Inspired Cognitive Architectures*, *19*, 32–38. https://doi.org/10.1016/j.bica.2016.12.002

Ntalampiras, S., Potamitis, I., & Fakotakis, N. (2009). An adaptive framework for acoustic monitoring of potential hazards. *EURASIP Journal on Audio, Speech, and Music Processing*, *2009*, 1–15. https://doi.org/10.1155/2009/594103

Ong, K. L., Lee, C. P., Lim, H. S., Lim, K. M., & Alqahtani, A. (2024). MaxMViT-MLP: Multiaxis and multiscale vision transformers fusion network for speech emotion recognition. *IEEE Access*, *12*, 18237–18250. https://doi.org/10.1109/ACCESS.2024.3360483

Patro, K. K., Allam, J. P., Neelapu, B. C., Tadeusiewicz, R., Acharya, U. R., Hammad, M., Yildirim, O., & Pławiak, P. (2023). Application of Kronecker convolutions in deep learning technique for automated detection of kidney stones with coronal CT images. *Information Sciences*, *640*, 119005. https://doi.org/10.1016/j.ins.2023.119005

Perfilieva, I. (2006). Fuzzy transforms: Theory and applications. *Fuzzy Sets and Systems*, *157*(8), 993–1023. https://doi.org/10.1016/j.fss.2005.11.012

Petrushin, V. (2000). Emotion in speech: Recognition and application to call centers. *Artificial Neural Networks in Engineering*, *710*, 22.

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv, abs/2010.16061*. https://doi.org/10.48550/ARXIV.2010.16061

Powroźnik, P. (2014). Polish emotional speech recognition using artificial neural network. *Advances in Science and Technology Research Journal*, *8*(24), 24–27. https://doi.org/10.12913/22998624/562

Powroźnik, P., & Czerwiński, D. (2016). Spectral methods in polish emotional speech recognition. *Advances in Science and Technology Research Journal*, *10*(32), 73–81. https://doi.org/10.12913/22998624/65138

Powroźnik, P., Wojcicki, P., & Przylucki, S. W. (2021). Scalogram as a representation of emotional Speech. *IEEE Access*, *9*, 154044–154057. https://doi.org/10.1109/ACCESS.2021.3127581

Prasomphan, S. (2015). Improvement of speech emotion recognition with neural network classifier by using speech spectrogram. *International Conference on Systems, Signals and Image Processing (IWSSIP)* (pp. 73–76). IEEE. https://doi.org/10.1109/IWSSIP.2015.7314180

Ristea, N.-C., Ionescu, R. T., & Khan, F. S. (2022). SepTr: Separable transformer for audio spectrogram processing (Version 3). *ArXiv, abs/2203.09581*. https://doi.org/10.48550/ARXIV.2203.09581

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Song, P., Zheng, W., Ou, S., Zhang, X., Jin, Y., Liu, J., & Yu, Y. (2016). Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. *Speech Communication*, *83*, 34–41. https://doi.org/10.1016/j.specom.2016.07.010

Tang, X., Huang, J., Lin, Y., Dang, T., & Cheng, J. (2025). Speech emotion recognition via CNN-transformer and multidimensional attention mechanism. *Speech Communication*, *171*, 103242. https://doi.org/10.1016/j.specom.2025.103242

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5200–5204). IEEE. https://doi.org/10.1109/ICASSP.2016.7472669

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*(9), 1162–1181. https://doi.org/10.1016/j.specom.2006.04.003

Wu, T., Tang, S., Zhang, R., Cao, J., & Li, J. (2019). Tree-structured Kronecker convolutional network for semantic segmentation. *IEEE International Conference on Multimedia and Expo (ICME)* (pp. 940–945). IEEE. https://doi.org/10.1109/ICME.2019.00166

Xia, W., Huang, J., & Hansen, J. H. L. (2019). Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation. *ArXiv, abs/1908.01447*. https://doi.org/10.48550/ARXIV.1908.01447

Zhao, X., Zhang, S., & Lei, B. (2014). Robust emotion recognition in noisy speech via sparse representation. *Neural Computing and Applications*, *24*, 1539–1553. https://doi.org/10.1007/s00521-013-1377-z

Zheng, W. Q., Yu, J. S., & Zou, Y. X. (2015). An experimental study of speech emotion recognition based on deep convolutional neural networks. *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 827–831). IEEE. https://doi.org/10.1109/ACII.2015.7344669