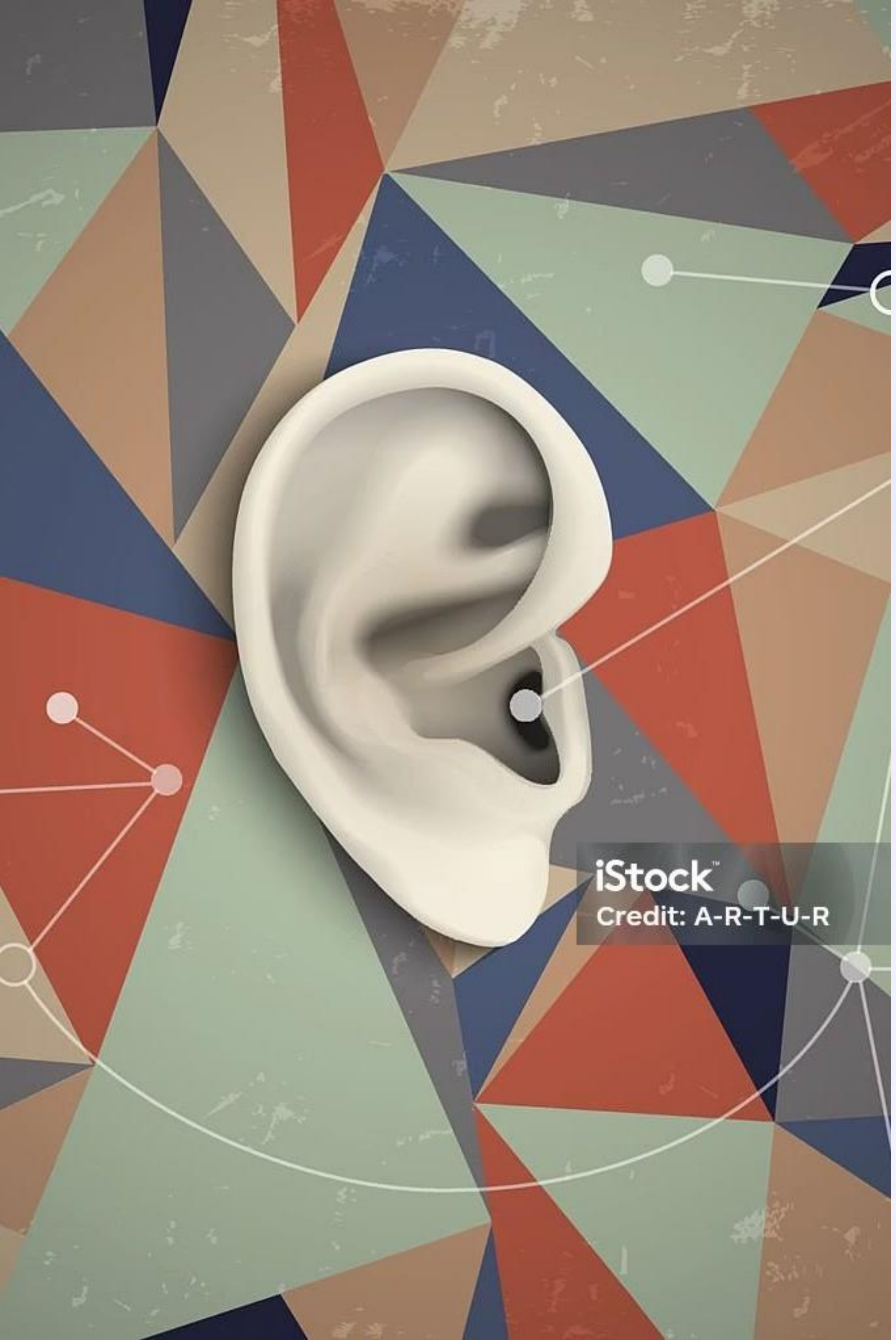


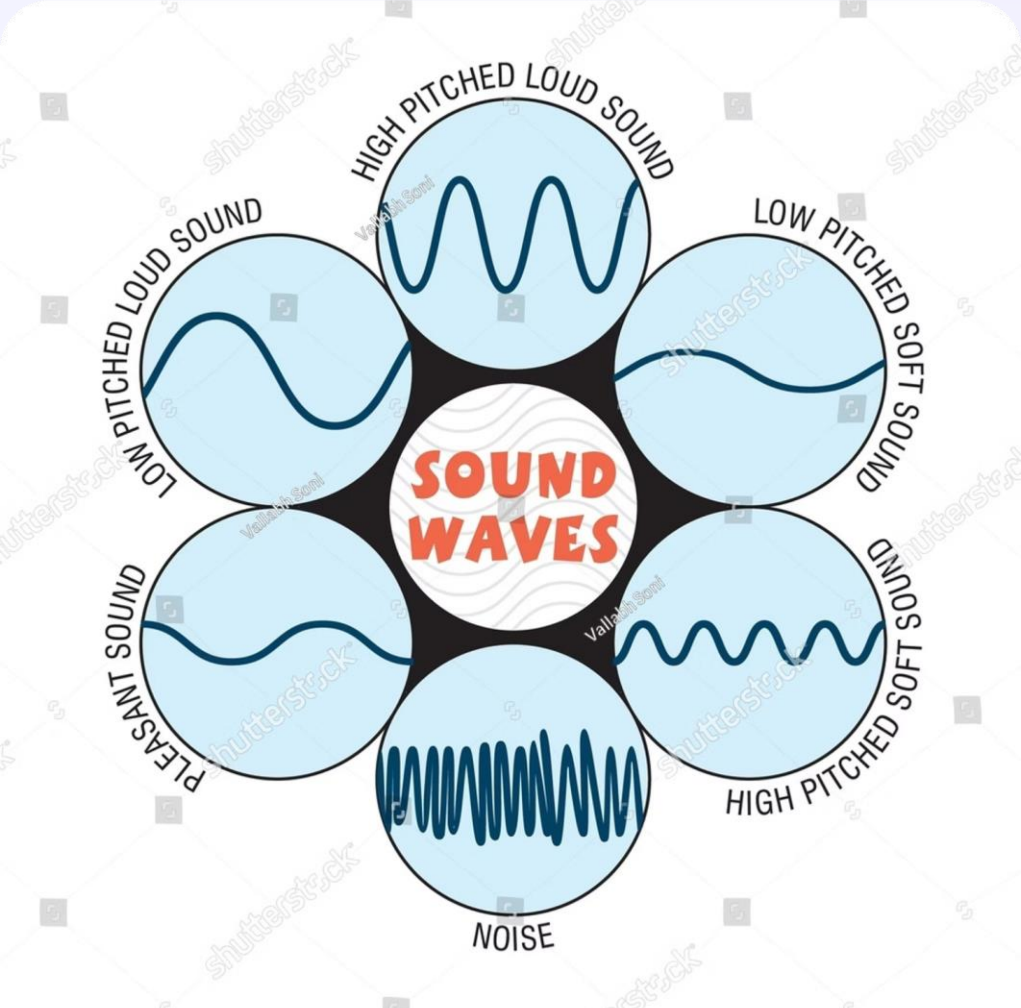
Speech Emotion Recognition System

Unlocking the power of AI to understand human emotions through speech. This presentation explores the fascinating field of Speech Emotion Recognition (SER) and its transformative potential across various sectors, from human-computer interaction to healthcare.

What is Speech Emotion Recognition (SER)?

Speech Emotion Recognition (SER) is a cutting-edge technology that identifies and predicts human emotions from spoken language. By analyzing various acoustic properties of speech, SER systems can infer emotional states such as happiness, sadness, anger, and neutrality.

- Predicts emotions like happy, sad, angry from voice Uses
- audio signal processing and AI classification
- Emotions conveyed via tone, pitch, energy, speech patterns



Key Datasets for SER Training

Effective SER models rely on high-quality, diverse datasets. These datasets contain recorded speech labeled with specific emotions, allowing AI models to learn the intricate acoustic patterns associated with different human feelings.

1

RAVDESS

Ryerson Audio-Visual Database of Emotional Speech and Song, featuring 24 professional actors with a North American accent.

2

SAVEE

Surrey Audio-Visual Expressed Emotion, developed with 4 male actors, capturing various emotions.

3

TESS

Toronto Emotional Speech Set, includes 200 words spoken by two actresses, covering a range of emotions.

4

EMO-DB

German Emotional Speech Database, recorded in an anechoic chamber for high-fidelity audio.

Feature Extraction Techniques

Feature extraction is the process of converting raw audio signals into meaningful numerical representations that machine learning models can understand.These features capture the essence of emotional expression in speech.

MFCCs

Mel-Frequency Cepstral Coefficients capture the spectral envelope of the sound,crucial for distinguishing vocal timbre.

Chromagram

Represents the distribution of energy across different pitch classes,similar to musical notes.

Prosodic Features

Includes pitch (F0),energy,and formants,which relate to the vocal tract's resonance and influence perceived emotion.

Machine Learning & Deep Learning Models

The evolution of SER has seen a shift from traditional machine learning approaches to sophisticated deep learning architectures, significantly enhancing accuracy and robustness.

Traditional Models

- Support Vector Machines (SVM)
- Gaussian Mixture Models (GMM)
- Hidden Markov Models (HMM)

These models rely on pre-defined statistical features and are effective for smaller datasets.

Deep Learning Models

- Convolutional Neural Networks (CNN)
- Recurrent Neural Networks (RNN) Long
- Short-Term Memory (LSTM) Hybrid CNN-
- LSTM Architectures

Deep learning models automatically learn features and excel with large datasets, offering superior performance.

Recent Advances: Hybrid Features & OpenAI Whisper

The SER field is rapidly advancing with innovative techniques. Hybrid feature combinations and the fine-tuning of large pre- trained models are driving significant performance improvements, pushing accuracy to new heights.

Hybrid Features

Combining MFCCs with time-domain features, such as signal energy and zero-crossing rate, has yielded up to 97% accuracy with CNN models.

OpenAI Whisper V3

Fine-tuning the Whisper Large V3 model for SER has achieved approximately 92% accuracy, leveraging its robust audio understanding capabilities.

Transformative Applications of SER Systems

SER technology is poised to revolutionize various industries by enabling machines to better understand and respond to human emotions.

Call Centers

Monitor customer emotions to improve service quality, prioritize urgent cases, and train agents for empathetic interactions.

Healthcare

Detect patient emotional states to enhance diagnostics, personalize care, and support mental health monitoring.

Human-Computer Interaction

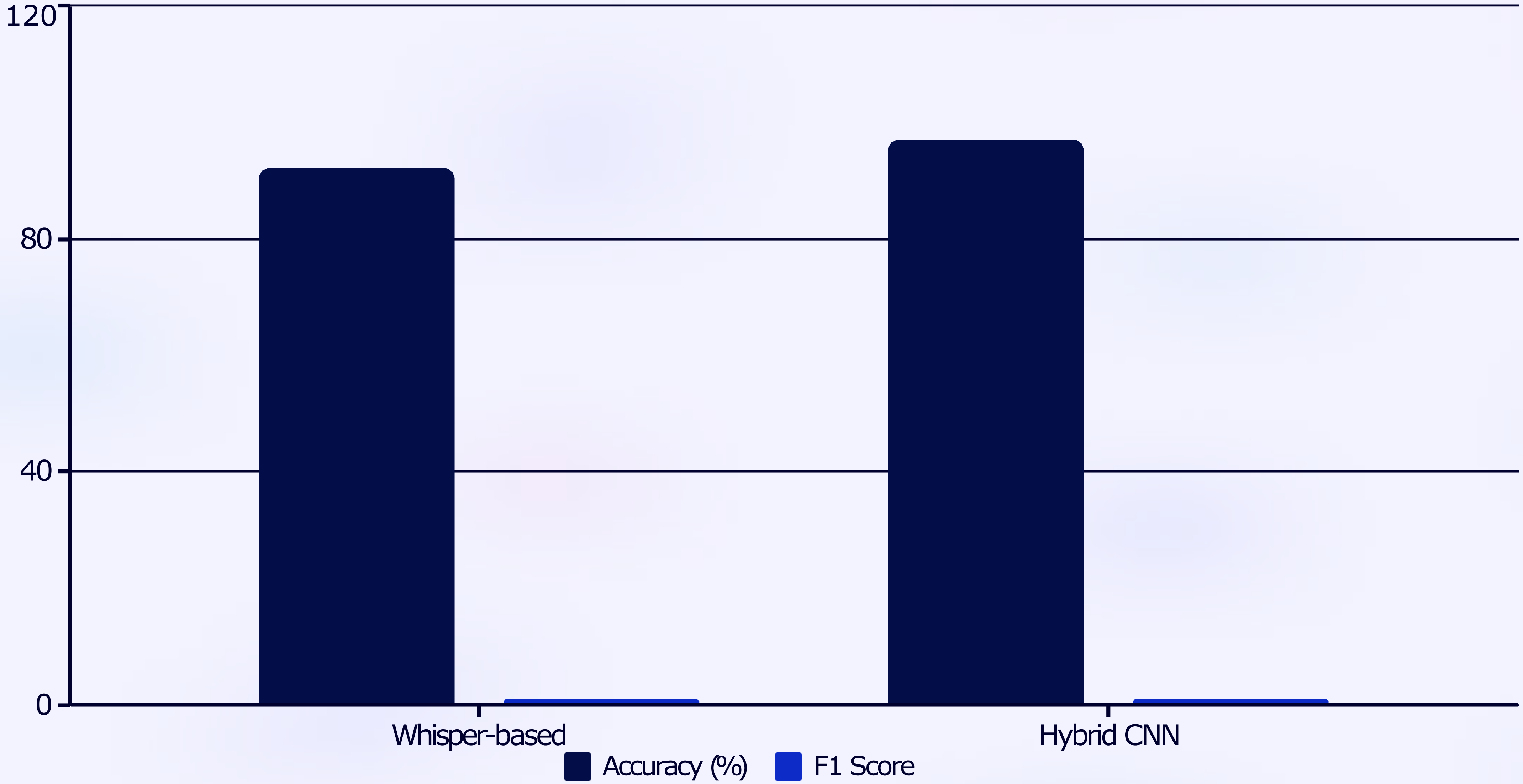
Develop more empathetic and natural virtual assistants and chatbots that can adapt responses based on user emotions.

Entertainment

Enable personalized content recommendations in music, movies, and games based on user emotional responses.

Evaluating SER Performance

The effectiveness of SER systems is measured using standard evaluation metrics, providing a quantitative assessment of their accuracy and reliability.



The chart illustrates the typical performance of advanced SER models. Performance greatly depends on the balance and quality of the training dataset, as well as the chosen feature extraction techniques.

Challenges & Future Directions

While SER has made significant strides, several challenges remain. Addressing these will pave the way for more robust and versatile emotion recognition systems.

1

Variability

Speech styles, accents, and cultural differences pose significant challenges to universal emotion detection.

2

Subtle Emotions

Handling blended and subtle emotions (e.g., sarcasm, mixed feelings) remains a complex task.

3

Data Needs

There's a continuous need for larger, more diverse, and accurately labeled datasets.

4

Real-time Processing

Optimizing models for efficient real-time processing is crucial for many practical applications.

5

Multimodal Integration

Combining SER with Natural Language Processing (NLP) for a holistic, multimodal understanding of emotion.

Summary & Outlook

Speech Emotion Recognition systems are transforming how we interact with technology, turning spoken words into actionable emotional insights.

- SER systems transform speech into emotional insights using AI.
- Feature extraction and model choice are key to success.
- Recent models like OpenAI Whisper show promising results.
- Broad applications across industries with ongoing research.