

**THE VOICE AS A NEUROPHYSIOLOGICAL
MARKER: APPLICATIONS IN IMPLICIT
EMOTION PERCEPTION FOR NEUROCASH®
AND CASH DESIGN**

2026

BANCO DE ESPAÑA
Eurosistema

**Documentos Ocasionales
N.º 2618**

Lucía Canalejo Guijosa, Fernando León Martínez
and José María Sánchez Echave

**THE VOICE AS A NEUROPHYSIOLOGICAL MARKER: APPLICATIONS IN IMPLICIT EMOTION PERCEPTION
FOR NEUROCASH® AND CASH DESIGN**

THE VOICE AS A NEUROPHYSIOLOGICAL MARKER: APPLICATIONS IN IMPLICIT EMOTION PERCEPTION FOR NEUROCASH® AND CASH DESIGN (*)

Lucía Canalejo Guijosa

BANCO DE ESPAÑA

Fernando León Martínez

BANCO DE ESPAÑA

José María Sánchez Echave

BANCO DE ESPAÑA

(*) Los autores agradecen al Consejo Editorial del Banco de España, a las personas que han participado en el proceso de revisión del documento y a los equipos de Publicaciones y Maquetación sus comentarios, sugerencias y contribuciones, así como el trabajo realizado para la edición final de este documento ocasional. Los errores y las omisiones que pueda contener este trabajo son de la exclusiva responsabilidad de los autores.

Asimismo, la autora agradece a los miembros de la Unidad de Innovación y Desarrollo del Departamento de Efectivo su apoyo, orientación y acompañamiento durante su período de prácticas y durante la elaboración del documento, con especial reconocimiento a Fernando León Martínez, José María Sánchez Echave y Miguel López Soblechero por su dedicación y por los conocimientos y experiencia compartidos.

Documentos Ocasionales. N.º 2618

October 2026

<https://doi.org/10.53479/44665>

The Occasional Paper Series seeks to disseminate work conducted at the Banco de España, in the performance of its functions, that may be of general interest.

The opinions and analyses in the Occasional Paper Series are the responsibility of the authors and, therefore, do not necessarily coincide with those of the Banco de España or the Eurosystem.

The Banco de España disseminates its main reports and most of its publications via the Internet on its website at: <http://www.bde.es>.

Reproduction for educational and non-commercial purposes is permitted provided that the source is acknowledged.

© BANCO DE ESPAÑA, Madrid, 2026

ISSN: 1696-2230 (online edition)

Abstract

Situated within the Neurocash® framework for measuring implicit emotional and cognitive responses, this review examines the human voice as a non-invasive psychophysiological marker. It outlines the physiological mechanisms of voice production and explains how emotional states influence breathing, vocal-fold vibration and articulation, shaping acoustic features such as pitch and intensity. The review then examines Speech Emotion Recognition as the main approach for extracting emotional information from speech, describing the methodological pipeline from emotional speech databases and preprocessing to classification using traditional machine-learning algorithms and modern deep-learning and self-supervised models. The analysis concludes that no single computational approach is universally optimal, as performance depends on language, emotional categories, data quality and environmental conditions. Overall, this review provides an interdisciplinary framework linking psychophysiology, vocal acoustics and artificial intelligence to support Neurocash® studies on cognitive workload, perceived security and implicit trust in cash design.

Keywords: acoustic feature extraction, affective computing, implicit emotion perception, multimodal emotion analysis, psychophysiology, self-supervised representation learning, Speech Emotion Recognition, Neurocash®, cash design.

JEL classification: C55, C80, D91.

Resumen

A partir de la metodología Neurocash® para el estudio de las respuestas emocionales y cognitivas implícitas, en este documento se examina la voz humana como un marcador psicofisiológico no invasivo. En primer lugar, se describen los mecanismos fisiológicos de producción de la voz y se explica cómo los estados emocionales modifican la respiración, la vibración de las cuerdas vocales y la articulación, lo que altera características acústicas como el tono y la intensidad. A continuación, se analiza el sistema de reconocimiento de emociones en el habla (SER, por sus siglas en inglés) como la principal herramienta para extraer información emocional a partir de la voz, describiendo su proceso metodológico desde la selección de bases de datos y el preprocesamiento hasta la clasificación mediante algoritmos de aprendizaje automático clásico y modelos de aprendizaje profundo y autosupervisado. El análisis concluye que no existe un modelo universalmente óptimo y proporciona un marco interdisciplinar para apoyar estudios Neurocash® sobre carga cognitiva, seguridad percibida y confianza implícita en el diseño del efectivo.

Palabras clave: extracción de características acústicas, computación afectiva, percepción implícita de la emoción, análisis multimodal de emociones, psicofisiología, aprendizaje de representaciones autosupervisado, reconocimiento de emociones en el habla, Neurocash®, diseño del efectivo.

Códigos JEL: C55, C80, D91.

Contents

1	Introduction	9
1.1	Justification of the review	10
2	Methods	11
2.1	Search Strategy and Sources	11
2.2	Inclusion Criteria	11
2.3	Exclusion Criteria	12
2.4	Selection and Synthesis Process	12
3	Definition and Applications of SER	14
3.1	Definition	14
3.2	Applications	14
4	Voice Physiology and Acoustic Correlates of Emotion	15
5	Emotions and their Influence on Acoustics	17
6	Methodological Pipeline SER	18
6.1	Emotional Speech Databases	18
6.2	Preprocessing	19
6.3	Feature extraction	20
6.3.1	Temporal and fluency features	20
6.3.2	Prosodic features	20
6.3.3	Spectral features	20
6.3.4	Cepstral features	20
6.3.5	Energy-based features	20
6.3.6	Frequency- and pitch-related indicators.	21
6.3.7	Predefined standardized feature sets	21
6.4	Classifier Approaches: Conventional (Classical Machine Learning) vs. Deep Learning	22
6.4.1	Conventional methods (classical machine learning)	22
6.4.2	Deep learning methods	23
6.4.3	Is there an optimal model?	25
7	Complementary Approach to SER	26
8	Available Tools for SER Implementation	27
8.1	Commercial tools	28
8.2	Non-academic tools with PTM and external classifiers	28
8.3	Non-academic tools with PTM and NO external classifiers	29
8.4	Create a model from scratch	30
8.5	MixedEmotions Toolbox	31

9 Discussion 32

- 9.1 Overview and Interpretation of the Synthesis 32
- 9.2 Interdisciplinary Integration and Added Value 32
- 9.3 Methodological Heterogeneity in SER 32
- 9.4 Implications for Research and Practice 32
- 9.5 Potential Application to Cash Perception Research 33
- 9.6 Limitations of the Present Review 33
- 9.7 Future Directions 33

10 Conclusion 35

References 36

Annex: List of Acronyms 39

1 Introduction

Cash design is not only a matter of visual aesthetics and security features, but also of how people implicitly perceive, interpret, and emotionally respond to cash during daily interactions. In this context, voice analysis offers a promising bridge between cash perception research and psychophysiological measurement, as vocal signals can reflect subtle changes in emotional arousal, cognitive load, confidence, and perceived security. Understanding these implicit human reactions is fundamental in cash design to optimize currency usability and the effectiveness of anti-counterfeiting measures.

The aim of this review is to offer a broad and connected perspective on voice analysis, encompassing the physiological bases of vocal production, its relationship with emotional expression, and the technological tools currently available for extracting acoustic features in perception and affective-computing research. In doing so, the work synthesizes the present state of Speech Emotion Recognition (SER) technologies while preserving a conceptual and accessible perspective grounded in psychology. Rather than offering an exhaustive technical review of every computational component, the goal is to present a coherent, interdisciplinary overview that bridges psychophysiology, acoustics, and machine-learning approaches to emotional voice analysis.

SER is situated within the broader domain of Automatic Emotion Recognition (AER), a field that integrates multiple biosignals, such as EEG (electroencephalography), facial expressions, gestures, or speech, to infer affective states. Within this ecosystem, the human voice stands out as a unique psychophysiological signal: it arises from the coordination of breathing, vocal-fold movement, and articulation, all of which change with emotional, cognitive, and autonomic states (De Lope & Graña, 2023). Understanding these biological mechanisms is therefore essential for interpreting how emotions shape the acoustic signal. For this reason, the document begins by outlining the physiological and anatomical bases of vocal production, following authors who argue that acoustic analyses must be grounded in vocal biomechanics.

Building on this theoretical foundation, the second section examines how emotional states influence key acoustic features such as fundamental frequency, intensity, spectral balance, and temporal dynamics. Recent reviews show that emotions reliably shape vocal acoustics, allowing speech to convey meaningful psychophysiological information (Van Puyvelde et al., 2018; Schewski et al., 2025). These relationships are central both to psychological research and to the development of computational models that aim to decode emotions from speech.

The study then presents the complete Speech Emotion Recognition (SER) methodological pipeline, outlining the process of transforming raw audio into emotional labels. These stages include audio-source selection, preprocessing, acoustic feature extraction, and classification using either classical machine-learning algorithms or modern deep-learning architectures.

Finally, after describing the pipeline, the document provides a structured overview of the most relevant technological tools for implementing SER, ranging from commercial to open-source options. This includes commercial approaches (e.g., audeERING, iMotions), pretrained model pipelines with external or internal classifiers, custom SER systems built from scratch, and modular multimodal frameworks such as the MixedEmotions Toolbox.

By comparing these alternatives, practical considerations that researchers must evaluate when selecting tools for academic or applied contexts are highlighted, including model interpretability, data requirements, and computational costs.

Although this review focuses on the human voice as a neurophysiological marker for implicit emotion, it is motivated by the broader research context of neurodesign for cash. Within the Neurocash® framework, the Innovation & Development unit of the Banco de España aims to establish an objective, multidimensional methodology to evaluate the public's perception of banknotes. By integrating neuroscientific tools and implicit behavioral measures, this project seeks to overcome the traditional limitations of subjective surveys, capturing the subconscious emotional and cognitive reactions of users when interacting with currency (Banco de España, n.d.; Ortuño et al., 2020).

Specifically applied to cash perception research, the Neurocash® methodology provides cash designers with a robust, quantitative tool to optimize the design process. By capturing these acoustic fluctuations during exposure to or evaluation of currency features, the human voice can complement other biometric and behavioral measures within this framework. This allows for the precise evaluation of how different visual and tactile elements capture attention, how cognitive workload is modulated during the verification of security features, and how these factors influence the user's implicit trust and perceived security (Banco de España, n.d.; Ortuño et al., 2020). By systematically analyzing these biometric, neurophysiological, and vocal responses, Neurocash® acts as a scientific bridge that enhances the aesthetics, intuitive usability, and anti-counterfeiting resilience of banknotes throughout their entire lifecycle.

1.1 Justification of the review

Despite the rapid growth of Speech Emotion Recognition (SER), existing reviews remain fragmented and often focus on only one part of the field. Most surveys highlight technical advances, especially, new machine-learning models and self-supervised speech methods, while giving much less attention to the psychophysiological and acoustic principles that support emotional expression in the voice. As a result, many overviews adopt a mainly computational viewpoint, with limited connection to the biological mechanisms of vocal production or to the well-established acoustic markers of emotion.

This limitation in the review literature reflects a broader problem in the field: a lack of integration across disciplines. Psychophysiology, vocal acoustics, and deep-learning-based SER systems are often treated as separate research areas, even though they naturally depend on one another. Because of this separation, key concepts, such as subglottal pressure changes, vocal-fold behaviour, prosodic shifts, and spectral patterns, rarely appear together in computational surveys. This reduces the interpretability of SER models and makes it harder to understand the psychophysiological meaning of the features these systems extract.

A further challenge is that research on how emotions influence vocal acoustics is spread across different fields such as clinical voice science, neuroscience, phonetics, and machine learning. Because these findings are rarely brought together, it is difficult to build a clear, unified understanding of how emotional processes shape the voice and how these cues are used in SER systems.

For these reasons, an integrative review that unifies physiological mechanisms, acoustic correlates, psychological theory, and contemporary SER methodologies is necessary. The present work aims to consolidate these domains into a coherent conceptual framework, offering the scientific community a comprehensive, interdisciplinary account of how emotions are encoded and decoded through the human voice.

2 Methods

This article was written as a narrative integrative review, a method that brings together different types of evidence from fields such as psychophysiology, acoustics, speech science, affective computing, and machine learning. This type of review makes it possible to combine empirical studies, theoretical work, technical documentation, and methodological papers to create a broad understanding of a research area. Because Speech Emotion Recognition (SER) involves many disciplines, this approach is well suited to summarizing its physiological, acoustic, and computational aspects within a single framework.

2.1 Search Strategy and Sources

Between October 2025 and February 2026, a broad literature search was performed across the following databases and platforms:

- Google Scholar
- PubMed (physiology, voice science)
- IEEE Xplore (machine learning, signal processing)
- ACM Digital Library [Human Computer Interaction (HCI), affective computing]
- arXiv (deep learning and transformer-based SER studies)
- Publisher websites for technical documentation (e.g., audEERING, iMotions, Affectiva, Microsoft, and Amazon Web Services)
- GitHub repositories for open-source SER toolkits.

The search combined keywords and Boolean operators such as:

“speech emotion recognition” OR “SER” OR “emotional prosody” OR “acoustic correlates of emotion” OR “self-supervised speech models” OR “wav2vec2” OR “HuBERT” OR “WavLM” OR “affective computing” OR “voice physiology” OR “emotion acoustics”.

2.2 Inclusion Criteria

Sources were included if they met at least one of the following criteria:

- Conceptual relevance to the physiological, psychological, or computational foundations of emotional voice production (e.g., Liebenthal et al., 2016).
- Empirical or methodological contribution to acoustic correlates of emotion, vocal stress, or prosody (e.g., Van Puyvelde et al., 2018; Schewski et al., 2025).

- Technical relevance to SER pipelines, preprocessing, feature sets, or databases (e.g., openSMILE, GeMAPS, EmoDB, RAVDESS, IEMOCAP).
- Machine learning or deep learning significance (e.g., de Lope & Graña, 2023; Kaur & Singh, 2023; Latif et al., 2023; Baevski et al., 2020; Hsu et al., 2021).
- Applied frameworks in HCI, IoT, or affective technologies (e.g., Chang et al., 2024; Wang et al., 2022).
- Commercial and tool documentation, when necessary to describe real world SER implementations (e.g., audEERING, iMotions, Affectiva).

2.3 Exclusion Criteria

- Studies unrelated to emotional voice analysis or without a clear methodological contribution.
- Papers focused exclusively on Automatic Speech Recognition (ASR) without relevance to emotion or prosody.
- Non scholarly sources lacking technical or conceptual value.
- Duplicated versions of the same paper (e.g., arXiv version and published journal version; only one was retained).

2.4 Selection and Synthesis Process

Rather than filtering the literature in a strict, linear way, the review built its themes gradually through several cycles of reading and analysis.

1. Initial screening of titles and abstracts to identify relevance.
2. Full-text assessment of a large corpus of documents (approx. 80–100) dated between 1980–2026
3. Thematic clustering into core domains:
 - Vocal physiology and how the voice is produced (breathing, vocal folds, and articulation).
 - How emotions shape the acoustic signal (changes in pitch/F0, loudness, voice quality, and timing).
 - The SER pipeline (audio collection, preprocessing, feature extraction, and classification methods).
 - Emotional speech datasets used for training and evaluation (e.g., EmoDB, RAVDESS, IEMOCAP).
 - Types of acoustic features (prosodic, spectral, cepstral, energy related, temporal/fluency features, and standard sets like eGeMAPS).

- Modelling approaches in SER (classical machine learning, deep learning models, and self supervised/pretrained speech models).
 - Implementation options and tools (commercial platforms, pretrained models with or without external classifiers, full custom pipelines, and multimodal frameworks like MixedEmotions).
 - Complementary NLP methods (ASR to text combined with sentiment or affect analysis).
4. Cross disciplinary integration, linking psychophysiology with computational modelling to construct a coherent conceptual narrative.

Because integrative reviews focus on conceptual synthesis rather than performing statistical analyses, no formal quality-assessment tools (such as PRISMA or GRADE) were used. Instead, priority was given to selecting recent, rigorous, influential sources that showed consistency with findings from other studies.

3 Definition and Applications of SER

3.1 Definition

Within the broader field of Automatic Emotion Recognition (AER), which includes analyses based on EEG, gestures, facial expressions, and other biometric signals, we find Automatic Speech Recognition (ASR), whose primary function is to convert audio into text. This process relies on signal processing and machine learning methods that are shared with emotion related systems (De Lope & Graña, 2023).

One of ASR's key subfields is Speech Emotion Recognition (SER), which uses similar signal processing tools such as speech rate, pitch, and prosody to identify and classify emotions present in speech. Recent work highlights the importance of these acoustic cues in predicting emotional states accurately (Van Puyvelde et al., 2018; Schewski et al., 2025).

3.2 Applications

SER is used in many areas, including psychology, healthcare, and marketing. However, for the purposes of this work, in the context of implicit perception measurement, we focus on its role within the Internet of Things (IoT). In IoT environments, devices interact continuously with users, and emotional information can help these systems adjust their behaviour, personalise their responses, and create more natural interactions (Chang et al., 2024). The IoT ecosystem includes several interaction focused fields, such as Human-Computer Interaction (HCI), Human-Machine Interaction (HMI), Human-Robot Interaction (HRI), and Affective Computing.

Within this ecosystem, SER becomes especially relevant for Affective Computing, a field whose goal is to detect and interpret human emotions so that systems can adapt affectively. This “affective layer” is what distinguishes Affective Computing from other interaction oriented disciplines:

- HCI, HMI, and HRI focus on the mechanics and design of interaction, such as interfaces, usability, ergonomics, and task performance, without necessarily analysing emotional states.
- Affective Computing, in contrast, explicitly incorporates emotion. It studies how machines can understand and respond to the user's affective signals (Wang et al., 2022), making SER a key component for enabling emotionally adaptive behaviour.

4 Voice Physiology and Acoustic Correlates of Emotion

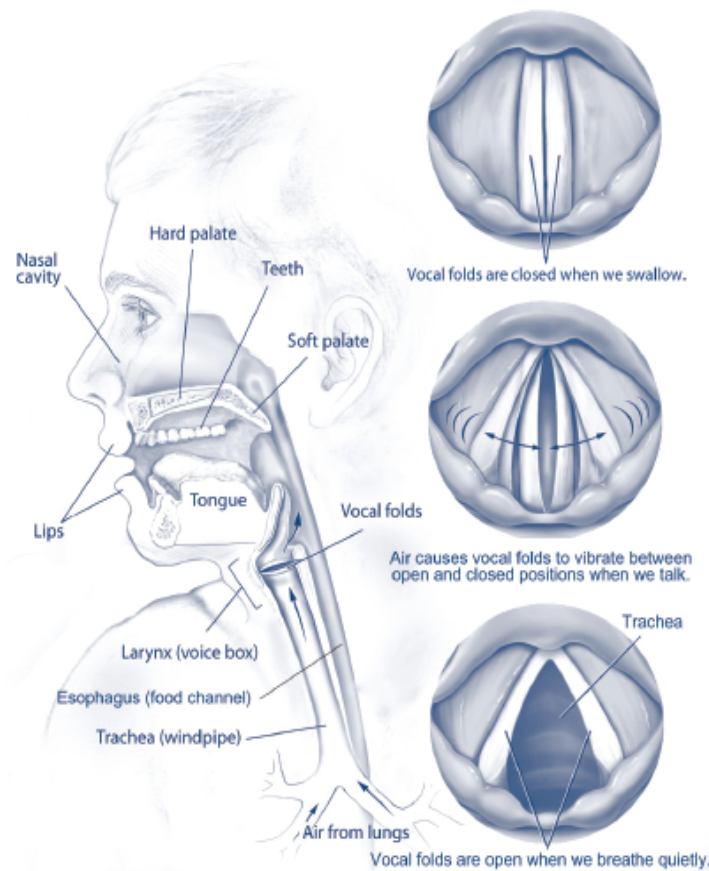
As illustrated in Figure 1, voice production begins with airflow originating in the lungs. The vocal folds, two bands of tissue located in the larynx between the base of the tongue and the top of the trachea, remain open during breathing and close during phonation (National Institute on Deafness and Other Communication Disorders [NIDCD, n.d.]). When speaking, the diaphragm contracts to expand the thoracic cavity, allowing the lungs to fill with air before releasing it during exhalation, which initiates the phonatory process (Ladefoged, 2001). Sound is produced when exhaled air passes through the closed vocal folds, increasing subglottal pressure until they separate, allowing air to flow through the glottis. The glottis widens during breathing and narrows during voice production (NIDCD, n.d.; Zhang, 2016). Rapid airflow creates pressure changes that bring the vocal folds back together, causing them to vibrate and generate sound. Typical speech requires normal vibratory behaviour of both the larynx and the vocal folds (Morrison, Rammage, Nichol, Pullan, & May, 1994).

Once phonation is understood, further components of speech production can be defined, namely articulation and resonance. Vocal production relies on the coordinated action of several subsystems:

- The respiratory phase: during which air exits the lungs (Morrison, Rammage, Nichol, Pullan, & May, 1994).

Figure 1

Structures involved in speech and voice production



SOURCE: From Vocal fold paralysis, by the National Institute on Deafness and Other Communication Disorders [NIDCD], n. d., Department of Health & Human Services, <https://www.nidcd.nih.gov/health/vocal-fold-paralysis>.

- Phonation: in which airflow sets the vocal folds into vibration (Morrison, Rammage, Nichol, Pullan, & May, 1994).
- Articulation, involving lips, tongue, teeth, and hard palate.
- Resonance, produced by the oral and nasal cavities that amplify and shape the voice (NIDCD, n.d.).

Physiological factors such as pitch, influenced by vocal-fold length and tension; intensity, determined by airflow force and vibratory amplitude; timbre, shaped by the vocal tract's configuration; and duration, defined by the length of vocal-fold vibration, all contribute to the acoustic characteristics of the human voice (Morrison, Rammage, Nichol, Pullan, & May, 1994 & Zhang, 2016).

5 Emotions and their Influence on Acoustics

Emotions influence the way our voice sounds because they affect how we breathe, how the larynx is positioned, and how the vocal tract moves. In psychology, emotions are understood as complex reactions that include thoughts, bodily changes, and expressive behaviours, all of which can shape our vocal output (American Psychological Association [APA], n.d.). Research shows that emotional states modify key acoustic features, such as fundamental frequency (F0), loudness, spectral balance or voice quality, and timing, through physiological changes in subglottal pressure, vocal fold vibration, and the configuration of the vocal tract (Liebenthal et al., 2016; Van Puyvelde et al., 2018). For example, positive emotions often occur alongside facial and postural changes like smiling, which alter the shape of the vocal tract and tend to produce a higher, brighter-sounding voice. More generally, high-arousal emotions such as joy or anger are associated with louder voices and greater variation in intensity than low-arousal states like sadness or tenderness (Liebenthal et al., 2016).

Emotional and cognitive demands also influence vocal behaviour. When people are under cognitive–emotional load, their voices show measurable shifts in F0, loudness, and perturbation-based voice measures, reflecting the combined effect of attention, mental effort, and affective state on vocal production (Van Mersbergen & Payne, 2021; Van Puyvelde et al., 2018). Clinical reviews further show that increased tension in the larynx can produce rougher voice quality and reduced vocal power, illustrating how physiological stress and emotional strain affect the sound of the voice (Eastwood et al., 2015). Taken together, these findings support the idea that emotions influence speech through coordinated changes in bodily, cognitive, and expressive systems, consistent with the three-component model of emotion.

A growing body of research demonstrates that emotions also shape prosody, patterns of pitch, rhythm, and intonation, in reliable ways. A large review of speech produced under negative emotions, stress, and cognitive load shows that anger, fear, anxiety, and stress frequently involve increases in F0 and, in some cases, noticeable changes in speech rate or pausing, although results vary across studies (Schewski et al., 2025). Evidence from acted emotional speech in Mexican Spanish similarly reveals distinct prosodic profiles: *emocionado* [excited] and *enojado* [angry] show high F0 values, wide pitch movements, and rising bitonal accents, while *tranquilo* [calm] and *triste* [sad] display flatter, more even pitch contours with smaller differences between syllables. Neutral utterances follow the expected declarative pattern typical of Central Mexican Spanish (González Franco et al., 2023).

6 Methodological Pipeline SER

Once we understand how the voice is produced and how emotions can shape it, we can look at how Speech Emotion Recognition (SER) systems go from a simple audio recording to identifying the emotion behind it. This process typically involves four stages:

- **Audio source selection:** Studies may employ spontaneous speech recordings or standardized emotional speech databases. Widely used corpora, such as EmoDB (Burkhardt et al., 2005), RAVDESS (Livingstone & Russo, 2018), and IEMOCAP (Busso et al., 2008), provide controlled, annotated emotional material essential for SER benchmarking.
- **Data preprocessing:** Includes denoising, normalization, segmentation, and amplitude or sampling-rate standardization to ensure signal quality and consistency across recordings (De Lope & Graña, 2023).
- **Feature extraction:** Extracted acoustic features typically involve pitch, prosodic dynamics, spectral energy, speech rate, and voice quality. In many studies, researchers rely on standardized sets like eGeMAPS or openSMILE so that acoustic features remain comparable across different projects (Eyben et al., 2016).
- **Classification approaches:** SER tasks may be addressed using traditional machine-learning models (e.g., SVMs, Random Forests) as well as deep learning architectures (e.g., CNN-, RNN-, or transformer-based systems) trained on emotional speech databases (De Lope & Graña, 2023).

6.1 Emotional Speech Databases

Table 1 presents a comparison of emotional speech databases according to the number of target emotions, language, availability of visual and acoustic modalities, whether speech is acted or natural, speaker demographics, and total number of utterances (De Lope & Graña, 2023). These databases can be used both to train deep learning models for emotion recognition and to benchmark different SER approaches.

Table 1
Database summary table

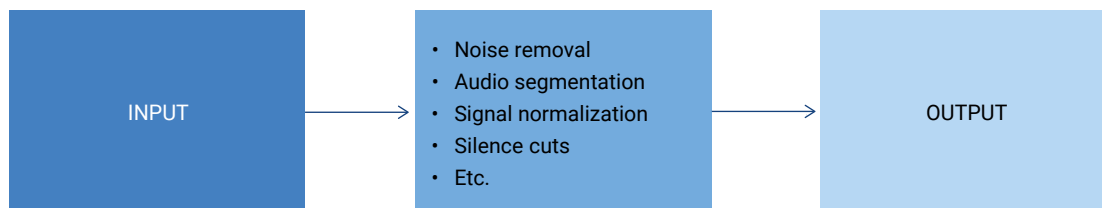
Database	Characteristics						
	Number of emotions displayed	Language	Visual & acoustic data	Natural vs Actors	Gender M: Men W: Women	Number of phrases	Number of times they have been referenced until 2023
DES	5	Danish	No	Actors	2 M 2 W	2 isolated words 9 short phrases and 2 passages	6
EMODB	7	German	No	Actors	5 M 5 W	800	51
eINTERFACE	6	English	Yes	Natural	34 M 8 W	6 successive short stories	6
IEMOCAP	10	English	Yes	Actors	5 M 5 W	Scripted/improvised sessions 12h	25
SAVEE	7	English	Yes	Actors	4 M	480	10
Thai DB	6	Thai	Yes	Actors	3 M 3 W	5.832	1
INTER1SP	7	Spanish	No	Actors	1 M 1 W	6.040	2
TESS	7	English	No	Actors	2 W	2.800	4
RAVDESS	8	English	Yes	Actors	12 M 12 W	1.440	22

SOURCE: Author's own elaboration adapted from De Lope and Graña (2023). An ongoing review of speech emotion recognition, *Neurocomputing*, 528, 1-11. <https://doi.org/10.1016/j.neucom.2023.01.002>

6.2 Preprocessing

Before extracting acoustic features, the audio signal must undergo a dedicated preprocessing stage. This phase typically includes noise reduction, signal normalization, segmentation, and the conversion of the waveform into formats suitable for analysis or model training. Figure 2 summarizes the main steps involved in the preprocessing workflow. Preprocessing is essential because improperly cleaned or poorly normalized audio can degrade the performance of classifiers and distort the resulting emotional inferences, as emphasized in recent reviews of speech emotion recognition pipelines (De Lope & Graña, 2023). Furthermore, detailed methodological work on audio-based affective systems highlights that preprocessing steps, such as padding, spectrogram generation, normalization, and

Figure 2
Preprocessing workflow schema (noise removal, audio segmentation, signal normalization, silence trimming)



SOURCE: Author's own elaboration adapted from De Lope and Graña (2023) and Katti and Sumana (2023).

silence trimming, constitute a foundational component of any machine-learning workflow involving speech signals (Katti & Sumana, 2023). Together, these studies demonstrate that preprocessing is not an optional step but a critical requirement for ensuring reliability, reproducibility, and model robustness in SER research.

6.3 Feature extraction

A speech recording contains thousands of measurable acoustic properties, many of which are useful for Speech Emotion Recognition (SER). Recent reviews group these properties into several main categories, temporal, prosodic, spectral, cepstral, and energy-related features, and describe standardized feature sets developed for paralinguistic analysis (De Lope & Graña, 2023; Bhangale & Kothandaraman, 2023).

6.3.1 Temporal and fluency features

Temporal features describe how speech unfolds over time. They include overall duration, speech rate, syllable timing, rhythm, and the presence or length of pauses, as well as distinctions between vocalic and non-vocalic segments. These patterns often change with emotional arousal or cognitive load and are frequently included in both traditional feature sets and deep-learning SER models (Byun & Lee, 2021; De Lope & Graña, 2023).

6.3.2 Prosodic features

Prosodic features relate to pitch and loudness patterns, such as fundamental frequency (F0), intonation contours, intensity, accentuation, and overall variability. Because they reflect adjustments in phonation and respiration, prosodic cues are reliably associated with emotional expression in speech (Bhangale & Kothandaraman, 2023; De Lope & Graña, 2023).

6.3.3 Spectral features

Spectral features capture how energy is distributed across frequencies. They include formants, spectral centroid, roll-off, kurtosis, and various energy bands, as well as spectrogram-based representations such as Mel-spectrograms. These features reflect vocal-tract configuration and are essential in both classical machine-learning and deep learning-based SER systems (Bhangale & Kothandaraman, 2023).

6.3.4 Cepstral features

Widely used in speech and emotion recognition, cepstral coefficients such as MFCCs, delta and delta-delta coefficients, GFCCs, and LPCCs capture perceptually and physiologically meaningful aspects of the signal. Many SER studies adopt them as core inputs for both machine-learning and neural architectures (Byun & Lee, 2021; Bhangale & Kothandaraman, 2023).

6.3.5 Energy-based features

Energy features measure the overall power of the signal and its variation over time. Metrics such as root mean square (RMS) energy can help differentiate high-arousal from low-arousal emotional states (Bhangale & Kothandaraman, 2023).

6.3.6 Frequency- and pitch-related indicators

Low-level features such as zero-crossing rate (ZCR) and line spectral pairs (LSP) describe signal oscillation patterns and spectral-envelope properties. They remain useful complementary indicators in multi-feature SER models (Bhangale & Kothandaraman, 2023).

6.3.7 Predefined standardized feature sets

Standard feature sets like eGeMAPS, GeMAPS, and ComParE (available in the openSMILE toolkit) combine many of the audio features mentioned earlier into well-tested, ready-to-use configurations. These sets are widely used in affective computing and other speech-related research areas (audEERING, n.d.; De Lope & Graña, 2023).

Finally, current SER research often compares these handcrafted features with representations learned directly from data, such as embeddings derived from self-supervised learning (SSL) models or automatic speech recognition (ASR) systems. Even so, traditional feature categories remain essential for benchmarking and interpreting these newer approaches (Gao et al., 2024).

For reference, these categories are summarized in Figure 3, with the features highlighted in bold corresponding to those repeatedly shown to carry strong emotional information.

Figure 3
Overview of principal acoustic feature categories used in Speech Emotion Recognition (SER)

CHARACTERISTICS				
Temporal and Fluency	Prosodic	Spectral	Cepstrals	Energy Based
• Total duration (speech speed) • Speech rate • Timelines (temporal representation) • Vowel and non-vowel segments • Rhythm and pauses • Syllable duration	• Fundamental Frequency (F0) • Tone / intonation • Intensity • Accentuation • Vocal cord tension (P08) • Glottal force (D09)	• Formants • Spectralbands • Conventional spectrograms • Mel spectrogram • Morlet waves	• MFCC (Mel-Frequency Cepstral Coefficients) • Delta and Delta Deltas • GFCC (Gammatone Frequency Cepstral Coefficients) • LPCC (Linear Predictive Cepstral Coefficients)	• RMS (Root Mean Square) • Derived variability indices
				Frequency and Pitch
				• LSP (Line Spectral Pairs) • ZCR (Zero Crossing Rate)
				Predefined Sets
				• eGeMAPS (Packages that group several features)

SOURCE: Author's own elaboration synthesizing concepts discussed in De Lope and Graña (2023), Bhangale and Kothandaraman (2023), Byun and Lee (2021) and Gao et al. (2024).

6.4 Classifier Approaches: Conventional (Classical Machine Learning) vs. Deep Learning

Speech Emotion Recognition (SER) systems are commonly divided into two major methodological families. The first relies on conventional machine-learning models built on hand-crafted acoustic features, while the second uses deep learning techniques that learn informative representations directly from the speech signal or from its time–frequency transformations. This section provides an overview of the modelling assumptions, typical input features, representative algorithms, and the main advantages and limitations associated with each approach. A comprehensive survey of the field discusses these developments in detail and motivates the comparison presented here (Kaur & Singh, 2023).

6.4.1 Conventional methods (classical machine learning)

- Modelling pipeline: Conventional Speech Emotion Recognition (SER) methods begin by extracting a set of hand-crafted features from the audio signal. These features describe core aspects of the voice, such as pitch, loudness, rhythmic patterns, and widely used acoustic descriptors like Mel-Frequency Cepstral Coefficients (MFCCs), which summarise how energy is distributed across frequency bands. After these features are obtained, they are passed to a traditional machine-learning classifier, such as Support Vector Machines, k-Nearest Neighbours, Naïve Bayes, decision trees or Random Forests, Gaussian Mixture Models, or Hidden Markov Models, which learns to distinguish between emotional categories. Because these features are manually designed rather than learned automatically, this approach tends to work reliably even when only limited training data are available. It is also generally considered more transparent and easier to interpret than modern deep-learning systems. MFCCs, in particular, have a long history in speech processing research, originating in the influential work of Davis and Mermelstein (1980).
- Representative classifiers:
 - Support Vector Machines (SVMs): SVMs try to find the best boundary that separates different emotion classes in the feature space, keeping the classes as far apart as possible. They were widely used in early SER systems (Kaur & Singh, 2023).
 - k-Nearest Neighbours (k-NN): k-NN classifies a new speech sample by checking which labelled samples it most closely resembles and assigning it the majority emotion among those neighbours (Kaur & Singh, 2023).
 - Naïve Bayes: A simple probabilistic model that estimates which emotion is most likely based on the features, assuming each feature contributes independently (Kaur & Singh, 2023).
 - Decision Trees / Random Forests: hierarchical partitioning; Decision trees split the data step by step based on feature values until they reach an emotion label. Random Forests combine many such trees, which makes them more stable and often more accurate, for example, outperforming SVMs on some SER datasets (Yan et al., 2020).
 - Gaussian Mixture Models (GMMs): GMMs represent each emotion as a combination of several Gaussian (bell-shaped) distributions. The model then estimates which emotion's distribution best fits the input speech (Bishop, 2006).

- Hidden Markov Models (HMMs): HMMs model speech as a sequence of short frames and capture how acoustic patterns change over time. They are especially useful for time-dependent signals like speech (Rabiner, 1989).
- Strengths and limitations: Conventional pipelines are computationally light, perform reasonably well with small, labelled datasets, and facilitate diagnostics (e.g., feature importance). However, performance strongly depends on the quality of the hand-crafted features, and classical models may face challenges when conditions change (e.g., speaker variation, noise, or language differences); limited multimodal fusion support is another constraint (Lian et al., 2023, Entropy).

6.4.2 Deep learning methods

- Modelling pipeline: Deep learning models tend to discover acoustic patterns on their own, whether they work directly with the waveform or with spectral representations. In doing so, they reduce the need for manually engineered features and instead discover informative representations automatically. A variety of neural encoder architectures are commonly used for this purpose. Convolutional Neural Networks (CNNs) are effective for analysing the two-dimensional structure of spectrograms, capturing local spectro-temporal patterns. Recurrent Neural Networks (RNNs), including LSTMs and GRUs, are suited to modelling how emotional cues evolve over time in the speech signal. Many SER systems combine these strengths through hybrid CNN–RNN models, which first extract local acoustic patterns and then learn their temporal dependencies. More recently, Transformer-based encoders have become prominent, using multi-head self-attention to capture long-range relationships across the speech sequence (Latif et al., 2023).
- Representative deep learning model families:
 - Convolutional Neural Networks (CNNs): CNNs extract local patterns from spectrograms (e.g., pitch, formants, energy) and are very effective when SER is formulated as an image-like problem. They were among the first deep models to outperform classical MFCC+SVM pipelines (Neumann & Vu, 2017; Satt et al., 2017).
 - Recurrent Neural Networks (RNNs): RNNs, especially LSTMs and GRUs, model the temporal dynamics of speech, capturing how emotions evolve across frames. They became a standard baseline for sequence modelling in SER before Transformers gained traction (Trigeorgis et al., 2016; Satt et al., 2017).
 - CNN–RNN hybrids: Hybrid models combine CNNs for spectral feature extraction with RNNs for temporal modelling. This architecture is common in SER because it combines the strengths of both approaches: they capture local spectrotemporal details while also modelling how the signal evolves over time (Trigeorgis et al., 2016; Abdelhamid et al., 2022).
 - Variational Autoencoders (VAEs): learn compressed emotional representations without labels. They are used for feature learning, noise reduction, or generating synthetic emotional speech (Latif et al., 2018).

- Transformer-based models: Transformers use self-attention to model long-range dependencies and have become the most powerful architecture in SER. Pretrained variants like wav2vec 2.0 and HuBERT dominate current benchmarks (Pepino et al., 2021; Latif et al., 2023).
- Supervised vs. Self-supervised learning:
 - Supervised learning requires labelled emotional audio and was the traditional setup for CNN/RNN models. Its performance depends strongly on dataset size and annotation quality (Trigeorgis et al., 2016).
 - Self-supervised learning (SSL) learns general speech features from unlabelled audio (e.g., wav2vec 2.0, HuBERT), achieving far better generalisation in low-resource emotional datasets (Hsu et al., 2021; Pepino et al., 2021).
- Novelties:
 - Pretrained models: One of the biggest recent advances is the use of models like wav2vec 2.0 and HuBERT. They are first trained on huge amounts of unlabelled audio and then fine-tuned for emotion recognition. This helps a lot when emotional datasets are small and currently give some of the strongest results, especially for valence (Pepino et al., 2021; Hsu et al., 2021; Wagner et al., 2023).
 - Multimodal models: Deep learning also makes it easy to combine multiple signals at once, for example, speech together with facial expressions or text. Different “fusion” strategies (early, late or intermediate) are used to mix these modalities, and recent surveys summarise the main techniques and datasets in multimodal emotion recognition (Lian et al., 2023).
 - Data augmentation: Since emotional datasets are often small, data augmentation is increasingly used to improve model robustness. Typical techniques include pitch shifting, time-stretching, adding noise, SpecAugment, vocal tract length perturbation and synthetic data generation. These methods improve generalisation, especially for deep models working in low-resource conditions (Ko et al., 2015; Park et al., 2019; Atmaja & Sasou, 2022; Chatziagapi et al., 2019; Qu et al., 2022).
- Strengths and limitations. Deep learning eliminates manual feature design, learns non-linear hierarchical patterns, scales with data, and integrates multimodal cues. However, these advantages come with trade-offs: deep models usually require large amounts of labelled data (although this can be reduced through self-supervised or transfer learning), they are harder to interpret, and they demand more computational resources during training and deployment (Latif et al., 2023; Wagner et al., 2023).

The main differences between conventional and deep learning approaches are summarized in Figure 4, including their respective architectures, advantages, and limitations.

Figure 4

Conventional vs. Deep Learning Approaches

Conventional learning				
Manual feature extraction	Most Relevant Classifiers	Advantages	Disadvantages	
	• SVM (Support Vector Machines) • k-NNY Naive Bayes • Decision Trees and Random Forests • HMM (Hidden Markov Models) • Gaussian Mixture Models (GMM)	• They work well with small datasets • They are more interpretable and easier to train • They are computationally lightweight	• They depend heavily on the quality of the features designed • Difficulty generalizing to noisy conditions, speaker variations, or different languages • Limited multimodal fusion (voice + physiology + context)	
Deep learning				
Direct data input:	Riesgo de mercado	Novelties	Advantages	Disadvantages
• Raw waveform • Spectrogra ms • Mel spectrogra ms	• CNN (Convolutional Neural Networks) • RNN (LSTM, GRU) • CNN+RNN híbridas • Transformers • Autoencodersand self-supervised models	• Data augmentation: add noise, pitch shifting, time stretching • Transfer learning: use pre-trained models in voice or music recognition tasks • Multimodal fusion: integrate voice with facial expressions, physiology, or text	• They eliminate the need to manually design features • They capture complex, non-linear relationships • They scale well with large databases	• They require a lot of labeled data • Lower interpretability • Higher computational cost

SOURCE: Author's own elaboration adapted from De Lope & Graña (2023).

6.4.3 Is there an optimal model?

A recurrent theme in the literature is that there is no single optimal model for Speech Emotion Recognition (SER). As highlighted by De Lope and Graña (2023), SER encompasses a wide spectrum of classical machine learning methods and deep learning architectures, each with different strengths depending on the characteristics of the data and the task at hand. Their review shows that performance varies substantially across datasets, languages, emotional taxonomies, and feature extraction strategies, and that both classical and deep learning approaches continue to coexist in SER research. This diversity of methods reflects that model choice is inherently problemdependent rather than universal.

In practical terms, this means that the “best” model depends on what the researcher seeks to achieve, for example, maximising accuracy, ensuring interpretability, working with limited training data, or integrating multimodal inputs. As De Lope and Graña (2023) emphasise, SER systems are highly sensitive to dataset properties and experimental conditions, and no approach consistently outperforms all others across benchmarks. Thus, the field does not converge toward a single optimal classifier, but rather toward a set of techniques that must be matched to the specific requirements of each application.

7 Complementary Approach to SER

A complementary approach to Speech Emotion Recognition (SER) is Natural Language Processing (NLP), which focuses on analysing the verbal content of speech rather than its acoustic properties. NLP requires converting the audio signal into text using an Automatic Speech Recognition (ASR) system. Modern ASR pipelines, covered extensively in standard NLP textbooks, transform the speech waveform into linguistic units that can be processed computationally (Jurafsky & Martin, 2026).

Once the speech is transcribed, NLP techniques can be applied to perform lexical, syntactic, and semantic analyses of the text. One widely used NLP application is Sentiment Analysis, which identifies explicit or implicit emotional content expressed through words, considering both word frequency and contextual meaning. Sentiment analysis is recognised as a core task within NLP, enabling the detection of attitudes, affect, and evaluative language in written or transcribed text (Ramdani, 2025; Jurafsky & Martin, 2026).

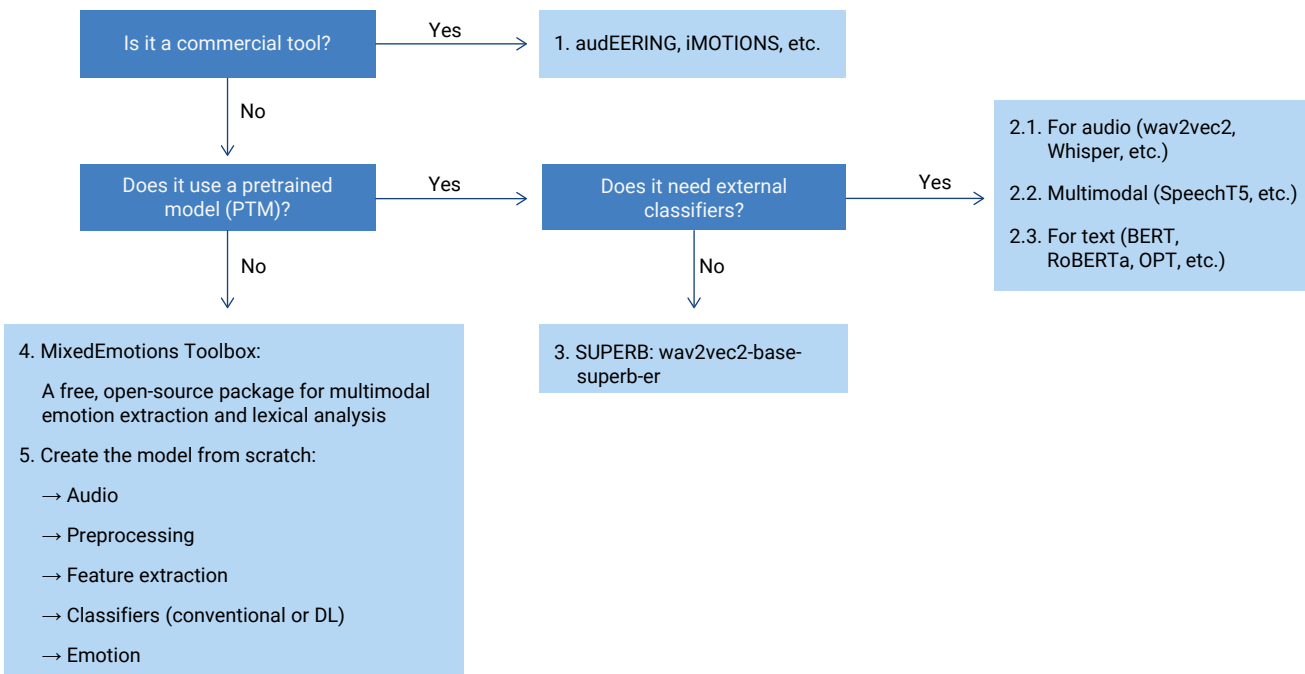
8 Available Tools for SER Implementation

This section reviews the tools that are most commonly used today to build SER systems. By summarizing the most relevant tools and computational approaches, it provides a clear view of the technical options and considerations that guide the design of contemporary SER systems.

To provide an integrated view of the technological landscape for Speech Emotion Recognition (SER), Figure 5 presents a structured decision pipeline that organizes current implementation pathways into a sequence of choices. The process begins by determining whether the system will rely on a commercial tool or an opensource alternative. If a commercial option is selected, platforms such as audEERING or iMotions offer readymade, endtoend solutions (Step 1). For noncommercial approaches, the next decision concerns the use of a pretrained model (PTM), which may require additional classifiers depending on the target task (Steps 2 and 3). If PTMs are not used, practitioners may either construct the entire SER pipeline from scratch—covering audio preprocessing, feature extraction, model design, and emotion inference (Step 4), or employ modular research frameworks such as the MixedEmotions Toolbox, which supports multimodal analysis across audio, text, and video (Step 5). In the following sections, each of these numbered pathways is examined in detail.

Figure 5 provides a visual representation of the decision pathway described above.

Figure 5
Decision map of Speech Emotion Recognition (SER) implementation options



SOURCE: Author's own elaboration adapted from audEERING (2025); Butelaar et al. (2018); iMotions (2025); and Ravanelli et al. (2021).

8.1 Commercial tools

Table 2 summarises the most relevant commercial tools currently available for voice-based and multimodal emotion analysis. The listed services include widely used platforms such as iMotions and Affectiva/Smart Eye (Affectiva, n.d.; iMotions, 2023, 2025), Microsoft Azure Cognitive Services/Foundry Speech (Microsoft, n.d.), Amazon Web Services—Amazon Chime SDK Voice Analytics/Tone Analysis (Amazon Web Services, 2023, 2024), Google Cloud Natural Language for text-based sentiment analysis (Google Cloud, 2026), Beyond Verbal (Beyond Verbal, n.d.), audEERING (audEERING, n.d.), and Imentiv (Imentiv, n.d.).

These products differ in the modalities they support (voice, text, facial expression, or multimodal bio-signal integration) and are generally neither open-source nor free to use. A notable exception is audEERING, which provides an academic-use-only Free Research Edition of its technology through the openSMILE-based toolkit (audEERING, n.d.).

Overall, Table 2 provides a structured comparison of the commercial ecosystem in SER and multimodal emotion analytics. The compilation is based on publicly available documentation from Affectiva, Amazon Web Services, audEERING, Beyond Verbal, iMotions, Google Cloud, Imentiv, and Microsoft (Affectiva, n.d.; Amazon Web Services, 2023, 2024; audEERING, n.d.; Beyond Verbal, n.d.; Google Cloud, 2026; iMotions, 2023, 2025; Imentiv, n.d.; Microsoft, n.d.).

Table 2
Commercial tools

Tools	Modality	Category	Free	Open Source
Microsoft Azure Cognitive Services	Text + Voice (SER + NLP)	Commercial	No	No
iMOTIONS	Multimodal (EEG, GSR, facial, video)	Commercial	No	No
Affectiva Emotion API NOW SMARTEYE	FACIAL + Voice (limited)	Commercial	NO (Offers demo/Free trial)	No
AudEERING devAlce®	Voice	Commercial	Free Research Edition (openSMILE)	Yes
Google Cloud Natural Language + Speech	Text + Voice (SER + NLP)	Commercial	Commercial with limited testing	No
Amazon Chime SDK Voice Tone Analysis	Text + Voice	Commercial	Commercial with limited testing	No
Innerity AI API	Multimodal (audio, text, image)	Commercial	No	No
Beyond Verbal	Voice	Commercial	No	No

SOURCE: Author’s own elaboration. Tools were included based on publicly available documentation from Affectiva (n.d.); Amazon Web Services (2023, 2024); audEERING (n.d.); Beyond Verbal (n.d.); Google Cloud (2026); iMotions (2023, 2025); Innerity (n.d.); and Microsoft Azure Cognitive Services (n.d.).

8.2 Non-academic tools with PTM and external classifiers

In the second branch, we focus on open-source approaches that are common in research and usually free to use. A practical strategy is to start with pretrained speech models (PTMs), models that are first trained on large audio datasets for general speech tasks such as ASR, rather than building an SER system from scratch (e.g.,

wav2vec 2.0, HuBERT, WavLM) (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022). These PTMs produce speech embeddings that can then be (a) fine-tuned with a small task head, or (b) used as features for a simple external classifier such as KNN, SVM, or MLP. This “embeddings + lightweight classifier” workflow is also how the SUPERB benchmark evaluates models: a frozen encoder with a small prediction head per task (Yang et al., 2021). In practice, many studies extract PTM embeddings and train conventional classifiers, or fine-tune a shallow head, to achieve competitive SER results (Chen & Rudnicky, 2023; Chen et al., 2024; Matt & Mani Sekhar, 2026).

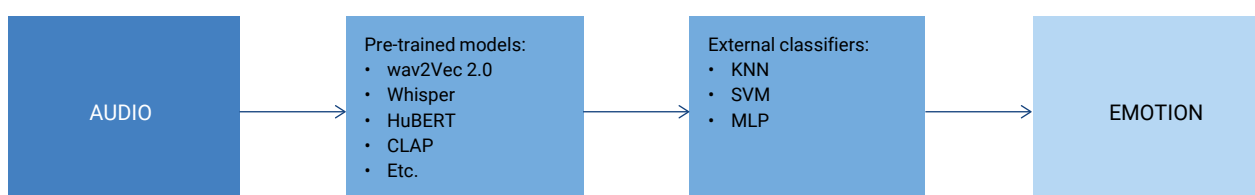
Figure 6 summarizes the general pipeline for using pre-trained speech models alongside external classifiers for speech emotion recognition.

Advantages. PTMs let researchers reuse large models without running expensive pre-training, which reduces compute cost while keeping good accuracy and flexibility across tasks (Yang et al., 2021; Chen et al., 2022).

Limitations: PTMs are not end-to-end SER systems. They usually require an external classifier or a small task-specific head, plus some integration steps (e.g., choosing layers, pooling embeddings, fusing features), which adds pipeline complexity (Yang et al., 2021; Baevski et al., 2020).

Figure 6

Non-academic tools with PTM and external classifiers



SOURCE: Author’s own elaboration, based on information from wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022), the SUPERB benchmark describing PTM-based feature extraction pipelines (Yang et al., 2021), and recent SER studies employing PTM embeddings with external classifiers such as SVM or MLP (Chen & Rudnicky, 2023; Chen et al., 2024; Matt & Mani Sekhar, 2026).

8.3 Non-academic tools with PTM and NO external classifiers

A second category of pretrained models includes end-to-end SER systems that already incorporate an internal classifier. Unlike feature-extraction models that require an external classifier, these systems provide a single-model solution: they process the audio, extract representations, and output an emotion label directly. Their simplicity comes at a cost, as their performance is constrained by the emotion labels and datasets used during fine-tuning, most commonly the IEMOCAP corpus in its four-class configuration defined in the SUPERB benchmark (Yang et al., 2021).

Several ready-to-use models of this type are available on HuggingFace, including the SUPERB ports `superb/wav2vec2-base-superb-er` and `superb/wav2vec2-large-superb-er`, both fine-tuned on IEMOCAP for emotion classification. Additionally, the SpeechBrain toolkit publishes its own end-to-end SER model, `speechbrain/emotion-recognition-wav2vec2-IEMOCAP`, which is likewise fine-tuned on the IEMOCAP dataset. These models can be used directly through HuggingFace pipelines without the need to train an external classifier (see Table 3).

Table 3

Non-academic tools with PTM and NO external classifiers

PTMs that do not require an external classifier	Base	Data Base
superb/wav2vec2-base-superb-er	superb/wav2vec2-base-superb-er	SUPERB - Emotion Recognition (IEMOCAP, 4 clases)
superb/wav2vec2-base-superb-er	superb/wav2vec2-base-superb-er	SUPERB - Emotion Recognition (IEMOCAP, 4 clases)
speechbrain/emotion-recognition-wav2vec2-IEMOCAP	brain/emotion-recognition-wav2vec2-IE	IEMOCAP pre trained model for SER

SOURCE: Author's own elaboration based on the SUPERB benchmark (Yang et al., 2021) and HuggingFace model documentation for superb/wav2vec2-base-superb-er, superb/wav2vec2-large-superb-er and speechbrain/emotion-recognition-wav2vec2-IEMOCAP.

8.4 Create a model from scratch

Developing a custom SER system involves manually designing and integrating each component of the pipeline. This approach offers full control over how audio is processed, which features are extracted, and how the classifier is built, but it also introduces greater technical complexity.

- The workflow typically follows the previously reviewed five stages:
 1. **Audio acquisition:** The process begins with selecting an audio source, either from existing emotional speech datasets such as EmoDB (Burkhardt et al., 2005), RAVDESS (Livingstone and Russo, 2018), or IEMOCAP (Busso et al., 2008), or from self-recorded material.
 2. **Preprocessing:** Before analysis, the audio must be cleaned and normalized to ensure consistent signal quality. Standard preprocessing steps include noise reduction, segmentation, silence trimming, and amplitude or sampling-rate normalization (De Lope & Graña, 2023; Katti & Sumana, 2023). These operations are essential because inadequate preprocessing negatively affects feature extraction and classifier performance.
 3. **Feature extraction:** Once the signal is prepared, relevant acoustic properties are extracted. Emotional information is typically encoded in temporal, prosodic, spectral, and cepstral features, such as pitch, duration, energy, MFCCs, and formants—commonly used in traditional SER systems (De Lope & Graña, 2023; Bhangale & Kothandaraman, 2023; Byun & Lee, 2021).
 4. **Classifier design:** After feature extraction, a classifier is trained to map the acoustic descriptors to emotional categories. Conventional machine-learning methods include SVMs, k-NN, and Random Forests (Kaur & Singh, 2023; Yan et al., 2020), whereas deep-learning approaches such as CNNs, RNNs, and hybrid CNN-RNN architectures have become increasingly common for modelling both spectral and temporal patterns (Latif et al., 2023; Trigeorgis et al., 2016; Pepino et al., 2021).
 5. **Emotion prediction:** Finally, the system outputs emotion labels or dimensional affect scores based on the trained model.

- **Advantages:** This approach provides complete control over the system, allowing researchers to tailor the dataset, features, and classifier architecture to the specific goals of the study.
- **Limitations:** Its main drawback is complexity. Designing, integrating, and validating all pipeline components requires more technical expertise and computational effort than using pretrained models or end-to-end systems.

8.5 MixedEmotions Toolbox

In 2018, the MixedEmotions project released an opensource and free platform designed for multimodal emotion analysis, integrating modules for text, audio, video, and linkeddata processing (Buitelaar et al., 2018). The platform was developed through a collaboration between several European institutions, including the National University of Ireland, Galway (Ireland), the University of Passau (Germany), the Universidad Politécnica de Madrid (Spain), and the Brno University of Technology (Czech Republic).

The MixedEmotions Toolbox provides a unified framework that enables automatic extraction of emotions across modalities and supports lexical and semantic analysis. For Speech Emotion Recognition (SER), the toolbox includes the “Emotion Recognition Audio (m6)” module, which implements a complete SER pipeline—from acoustic feature extraction to modelbased emotion prediction—returning continuous affect estimates such as arousal and valence (Buitelaar et al., 2018).

The platform can be downloaded from GitHub and executed using Docker containers, as each module is packaged as an independent microservice that can be deployed without dependency issues (Buitelaar et al., 2018).

9 Discussion

9.1 Overview and Interpretation of the Synthesis

This review integrates evidence from psychophysiology, vocal acoustics, and contemporary Speech Emotion Recognition (SER) to clarify how emotional information is encoded and decoded through the human voice. Vocal expression arises from coordinated respiratory, laryngeal, and articulatory mechanisms that systematically shape acoustic features such as pitch, intensity, spectral balance, and temporal dynamics. From this perspective, SER is not only a classification task, but a modelling process grounded in how the vocal system functions physiologically and how these mechanisms give rise to measurable acoustic markers of affect.

9.2 Interdisciplinary Integration and Added Value

A major contribution of this review is the integration of domains that are often treated separately. By linking physiological mechanisms with acoustic correlates, psychological interpretation, and machinelearning methodologies, the review shows how internal states are translated into features that SER systems can analyse. This interdisciplinary perspective improves interpretability, supports more informed feature selection, and provides a stronger foundation for applications in HCI, HRI, and IoT, where systems must adapt to users' emotional states.

9.3 Methodological Heterogeneity in SER

The field exhibits substantial heterogeneity at multiple levels:

- Data: emotional classifications, acted versus naturalistic speech, language coverage, recording conditions, and speaker diversity vary widely across databases; these factors directly affect benchmark comparability and model generalizability.
- Features and representations: The field ranges from using handcrafted features like prosody and MFCCs to relying on deep-learned and self-supervised representations, and this choice affects both model performance and the interpretability of the results.
- Modelling approaches: Classical machine-learning models still work well when data are limited and are easier to interpret, while deep learning, especially self-supervised models, delivers stronger performance and better transfer, but at the cost of more computation and lower interpretability. Given these sources of variation, there is no single optimal model for SER. Instead, model choice should be matched to the task, data regime, constraints on interpretability, and deployment context.

9.4 Implications for Research and Practice

Connecting computational methods to what we know about how the voice works offers two clear benefits:

First, it helps researchers make better decisions when choosing acoustic features and analysing errors. If a model misclassifies an emotion, the mistake can often be linked to specific physiological or articulatory factors—such as changes in vocalfold vibration or airflow—rather than being treated as an unexplained blackbox outcome.

Second, incorporating knowledge about vocal behaviour into the model design allows developers to include constraints that reflect how real speech is produced. These constraints, whether introduced through architecture or the training process, make the system more robust when dealing with background noise, different speakers, or natural variation in emotional expression.

In applied fields such as HCI, HRI, and IoT, this approach leads to systems that react more appropriately to users' emotions while remaining interpretable for developers and practitioners. In studies related to cash perception, these methods can help analyse emotional and cognitive reactions to cash design and perceptions of security. This aligns with the goals of the Neurocash® framework, which focuses on understanding emotional impact, cognitive processing, visual impact, and the public's sense of safety when interacting with cash (Banco de España, n.d.; Ortuño et al., 2020).

9.5 Potential Application to Cash Perception Research

Situated within the Neurocash® methodology, SER can act as an additional, noninvasive bio-signal to complement gaze, tactility and other physiological measures when evaluating cash prototypes, both in laboratory and virtual neurodesign settings. By capturing vocal features associated with arousal, valence and cognitive load during task-oriented interactions with cash stimuli, SER could help predict acceptance, clarity and perceived security of design elements. Moreover, integrating SER with current Neurocash® multimodal studies, such as EEG, gesture analysis and other biometric indicators, would provide a more holistic and robust view of implicit perception, enriching the project's ability to identify how users emotionally and cognitively experience cash-related designs.

9.6 Limitations of the Present Review

This work is a narrative integrative review rather than a systematic review or meta-analysis. As such, the goal was conceptual synthesis across heterogeneous literatures rather than quantitative aggregation. Because the sources span more than four decades of research (approximately 1980–2026), methodological conventions and technological capabilities vary considerably, which may limit direct comparability across studies. Although the search spanned multiple scholarly databases, preprint servers, and technical documentation, coverage cannot be exhaustive. In addition, the rapid pace of advances in deep learning and self-supervised speech modelling implies that specific tools and benchmarks may evolve quickly. Finally, while this review emphasizes physiological grounding and interdisciplinary integration, it does not resolve ongoing challenges in cross-lingual generalization, naturalistic emotion capture, and fine-grained affect modelling (e.g., mixed emotions, appraisal dimensions, context dependence).

9.7 Future Directions

Building on the perspective developed in this review, several areas deserve attention:

- More natural and cross-language data.
- Create larger and more diverse emotion-speech datasets that:
 - Use clear and consistent ways of labelling emotions.

- Include detailed information about speakers and situations.
- Capture emotions in real, everyday scenarios, not only acted speech.

This would help systems work better across different people, contexts, and languages.

- Models guided by how the voice works physically: Design models that consider what we know about voice production, for example, how pitch changes, how air pressure works, or how the tongue and lips move when emotions shift. Combining these understandable features with modern deep-learning methods may make systems stronger and easier to interpret.
- Understanding implicit emotion and multimodality: Develop studies and models that focus on implicit emotional signals, not only on explicit emotion labels. Additionally, evaluate multimodal systems that combine different sources of information, such as speech, text, facial expressions, and physiological signals, and systematically remove one source at a time to clearly identify the contribution of each modality.
- Better evaluation and more transparency: Develop testing standards that consider:
 - Uncertainty in emotion labels.
 - Unbalanced datasets.
 - Changes between training and real-world conditions.
 - Pair performance scores with tools that show what parts of the sound the system paid attention to, and whether those parts match known acoustic patterns.
- Responsible and fair deployment: For real-world applications (HCI, HRI, IoT), make sure systems handle privacy, consent, and fairness properly. This means documenting how data were collected, who is represented in the dataset, and how the system was checked for bias, not just for technical accuracy.

10 Conclusion

This review has brought together different research domains to show that the human voice carries meaningful traces of emotional and cognitive activity. Rather than treating speech as a purely linguistic signal, the evidence reviewed here highlights how changes in respiration, vocal-fold behaviour, and articulation leave measurable acoustic marks. These patterns —shifts in pitch, loudness, spectral balance, and timing— are not random; they reflect underlying psychophysiological states in ways that both psychologists and engineers can make use of.

Throughout the review, it has been emphasized that effective Speech Emotion Recognition (SER) depends as much on understanding how the voice works as on choosing the right computational tools. Clean preprocessing, thoughtful feature design or representation learning, and careful modelling choices are all important, but they make the most sense when grounded in vocal physiology and established acoustic correlations of emotion. Classical machine-learning models are still relevant, especially when transparency or limited data are concerned, whereas deep and self-supervised approaches have opened the door to stronger generalization. Even so, the field has not identified a single optimal modelling “answer”, as it is unlikely that one model will satisfy all research needs.

For applications in Human Computer Interaction (HCI), Human Robot Interaction (HRI), Internet of Things (IoT), and related domains, including emerging frameworks such as Neurocash®, SER can serve as an additional layer of information about users’ implicit reactions. Combining vocal cues via SER with NLP or with other biometric modalities seems promising, particularly when the goal is to capture subtle affective or cognitive shifts rather than explicit emotional displays. The success of these approaches, however, will likely depend on more diverse and naturalistic datasets, as well as clearer evaluation practices that consider label uncertainty and the mismatch between controlled laboratory data and real-world settings. This becomes especially critical as perception studies for cash design have become increasingly important in recent years, demanding that the data obtained are sufficiently relevant to support reliable conclusions. In this context, modifying certain procedures—such as replacing on-screen response formats with the recording of participants’ vocal responses to questions presented on a screen—allows the direct application of SER to extract emotional information from speech, yielding richer data for cash design within the Neurocash® methodology.

Finally, advances in SER must be accompanied by responsible treatment of privacy, consent, and fairness. Voice is a personal biometric, and emotion inference can raise concerns beyond those encountered in traditional speech processing. Clear documentation, careful dataset construction, and transparent model reporting will be essential if voice-based affective technologies are to be used in a way that is both scientifically grounded and socially acceptable.

References

- Abdelhamid, Abdelaziz A., El-Sayed M. El-Kenawy, Bandar Alotaibi, Ghada M. Amer, Mahmoud Y. Abdelkader and Abdelhameed Ibrahim. (2022). "Robust speech emotion recognition using CNN+LSTM based on stochastic fractal search optimization algorithm". *IEEE Access*, 10, pp. 49265-49284. <https://doi.org/10.1109/ACCESS.2022.3172954>
- Affectiva. (n. d. a). *Human Behavioral Research With iMotions*. <https://www.affectiva.com/product/imotions-behavioral-research>
- Affectiva. (n. d. b). *The Science Behind Emotion AI*. <https://www.affectiva.com/science-overview/>
- Alsabhan, Waleed. (2023). "Human-Computer Interaction with a Realtime Speech Emotion Recognition with Ensembling Techniques 1D Convolution Neural Network and Attention". *Sensors*, 23(1386). <https://doi.org/10.3390/s23031386>
- American Psychological Association. (n. d.). "Emotion". In *APA Dictionary of Psychology*. Retrieved February 9, 2026, from <https://dictionary.apa.org/emotion>
- Atmaja, Bagus Tris, and Akira Sasou. (2022). "Effects of Data Augmentations on Speech Emotion Recognition". *Sensors*, 22(5941). <https://doi.org/10.3390/s22165941>
- audEERING. (2025). <https://www.audeering.com/>
- audEERING. (n. d. a). *iMotions Collaborating With audEERING*. <https://www.audeering.com/company/imotions-collaboration/>
- audEERING. (n. d. b). *OpenSMILE: The Munich Open-Source Large-Scale Multimedia Feature Extractor*. <https://github.com/audeering/opensmile>
- Baeviski, Alexei, Henry Zhou, Abdelrahman Mohamed and Michael Auli. (2020). "Wav2vec 2.0: A Framework for self-Supervised Learning of Speech Representations". *Advances in Neural Information Processing Systems*, 33, pp. 12449-12460. https://proceedings.neurips.cc/paper_files/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html
- Banco de España. (n. d.). *El Banco de España y el LabLENI (Universidad Politécnica de Valencia) han desarrollado Neurocash®, una nueva tecnología basada en la neurociencia para diseñar el efectivo del futuro*. <https://www.bde.es/wbe/es/areas-actuacion/billetes-monedas/billetes-euros/contenido/investigacion-desarrollo/neurocash/neurocash--una-nueva-forma-de-disenar-el-efectivo.html>
- Bhangale, Kishor, and Mohanaprasad Kothandaraman. (2023). "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network". *Electronics*, 12-4(839). <https://doi.org/10.3390/electronics12040839>
- Bishop, Christopher. (2006). *Pattern Recognition And Machine Learning. Information Science And Statistics*. Springer.
- Buitelaar, Paul, Ian D. Wood, Sapna Negi, Mihael Arcan, John P. McCrae, Andrejs Abele, Cécile Robin, Vladimir Andryushechkin, Housam Ziad, Hesam Sagha, Maximilian Schmitt, Björn W. Schuller, J. Fernando Sánchez-Rada, Carlos A. Iglesias, Carlos Navarro, Andreas Giefer, Nicolaus Heise, Vincenzo Masucci, Francesco A. Danza, ... Giovanni Tummarello. (2018). "MixedEmotions: An Open Source Toolbox for Multimodal Emotion Analysis". *IEEE Transactions on Multimedia*, 20(9), pp. 2454-2465. <https://doi.org/10.1109/TMM.2018.2798287>
- Burkhardt, Felix, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier and Benjamin Weiss. (2005). "A database of German emotional speech". *Proceedings of Interspeech 2005*, pp. 1517-1520 [Datasets]. <https://doi.org/10.21437/Interspeech.2005-446>
- Busso, Carlos, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee and Shrikanth Narayanan. (2008). "IEMOCAP: Interactive emotional dyadic motion capture database". *Language Resources and Evaluation*, 42(4), pp. 335-359. <https://doi.org/10.1007/s10579-008-9076-6>
- Byun, Sung-Woo, and Seok-Pil Lee. (2021). "A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using deep Learning Algorithms". *Applied Sciences*, 11-4(1890). <https://doi.org/10.3390/app11041890>
- Chang, Yi, Zhao Ren, Zhonghao Zhao, Thanh Tam Nguyen, Kun Qian, Tanja Schultz and Björn W. Schuller. (2024). "Breaking Resource Barriers in Speech Emotion Recognition via Data Distillation". *Computer Science - Sound*. arXiv preprint. <https://arxiv.org/abs/2406.15119>
- Chatziagapi, Aggelina, Georgios Paraskevopoulos, Dimitris Sgouropoulos, Georgios Pantazopoulos, Malvina Nikandrou, Theodoros Giannakopoulos, Athanasios Katsamanis, Alexandros Potamianos and Shrikanth Narayanan. (2019). "Data Augmentation Using GANs for Speech Emotion Recognition". *Proceedings of Interspeech 2019*, pp. 171-175. <https://doi.org/10.21437/Interspeech.2019-2561>
- Chen, Li-Wei, and Alexander Rudnicky. (2023). "Exploring Wav2vec 2.0 Fine Tuning for Improved Speech Emotion Recognition". *ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing*. <https://doi.org/10.1109/ICASSP49357.2023.10095036>
- Chen, Sanyuan, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu and Furu Wei. (2022). "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing". *IEEE Journal of Selected Topics in Signal Processing*, 16(6), pp. 1505-1518. <https://doi.org/10.1109/JSTSP.2022.3188113>
- Chen, Weidong, Xiaofen Xing, Peihao Chen and Xiangmin Xu. (2024). "Vesper: A Compact and Effective Pretrained Model for Speech Emotion Recognition". *IEEE Transactions on Affective Computing*, 15(3), pp. 1711-1724. <https://doi.org/10.1109/TAFFC.2024.3369726>

- Davis, Steven, and Paul Mermelstein. (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), pp. 357-366. <https://doi.org/10.1109/TASSP.1980.1163420>
- De Lope, Javier, and Manuel Graña. (2023). "An ongoing review of speech emotion recognition". *Neurocomputing*, 528, pp. 1-11. <https://doi.org/10.1016/j.neucom.2023.01.002>
- Eastwood, Clare, Catherine Madill and Patricia McCabe. (2015). "The behavioural treatment of muscle tension voice disorders: A systematic review". *International Journal of Speech-Language Pathology*, 17(3), pp. 287-303. <https://doi.org/10.3109/17549507.2015.1024169>
- Eyben, Florian, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan and Khiet P. Truong. (2016). "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing". *IEEE Transactions on Affective Computing*, 7(2), pp. 190-202. <https://doi.org/10.1109/TAFFC.2015.2457417>
- Gao, Yuan, Hao Shi, Chenhui Chu and Tatsuya Kawahara. (2024). "Speech Emotion Recognition with Multi-level Acoustic and Semantic Information Extraction and Interaction". *Proceedings of Interspeech 2024*, pp. 1060-1064. <https://doi.org/10.21437/Interspeech.2024-2385>
- Golbaghi, Aryan, and Shuo Zhou. (2025). "Enhancing Speech Emotion Recognition Via Fine-Tuning Pre-Trained Models and Hyper-Parameter Optimisation". *Computer Science – Machine Learning*. arXiv preprint. <https://doi.org/10.48550/arXiv.2510.07052>
- González Franco, Ofelia Elizabeth, Elia Haydée Carrasco Ortiz and Eva Patricia Velásquez Upegui. (2023). "Descripción prosódica de las emociones actuadas en español mexicano". *Lingüística y Literatura*, 44(83), pp. 43-64. <https://doi.org/10.17533/udea.lyl.n83a02>
- Hamilton, Nick J. I. (2022). "The life-cycle and restoration of the human vocal fold". *Laryngoscope Investigative Otolaryngology*, 8(1), pp. 168-176. <https://doi.org/10.1002/lio2.993>
- Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov and Abdelrahman Mohamed. (2021). "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units". *Computer Science – Computation and Language*. arXiv preprint. <https://doi.org/10.48550/arXiv.2106.07447>
- iMotions. (2023). *audEERING & iMotions Partnership Announcement*. <https://imotions.com/blog/learning/product-news/audEERING-imotions-partnership-announcement/>
- iMotions. (2025). *Voice Analysis Module*. <https://imotions.com/support/document-library/voice-analysis-module/>
- Jurafsky, Dan, and James H. Martin. (2026). *Speech and Language Processing: An Introduction To NLP, Computational Linguistics, And Speech Recognition*. Online manuscript. January 2026 release. <https://web.stanford.edu/~jurafsky/slp3/>
- Katti, Anantshesh, and Mahadevaiah Sumana . (2023). "Pipeline for Pre-processing of Audio Data". *IOT with Smart Systems. Smart Innovation, Systems and Technologies*, 312, pp. 191-205. https://doi.org/10.1007/978-981-19-3575-6_21
- Kaur, Kamaldeep, and Parminder Singh. (2023). "Trends in speech emotion recognition: a comprehensive survey". *Multimedia Tools and Applications*, 82, pp. 29307-29351. <https://doi.org/10.1007/s11042-023-14656-y>
- Ko, Tom, Vijayaditya Peddinti, Daniel Povey and Sanjeev Khudanpur. (2015). "Audio augmentation for speech recognition". *Proceedings of Interspeech 2015*, pp. 3586-3589. <https://doi.org/10.21437/Interspeech.2015-711>
- Ladefoged, P. (2001). *A course in phonetics* (4th ed.). Heinle & Heinle.
- Latif, Siddique, Rajib Rana, Junaid Qadir and Julien Epps. (2018). "Variational Autoencoders for Learning Latent Representations of Speech Emotion: A Preliminary Study". *Proceedings of Interspeech 2018*, pp. 3097-3101. <https://doi.org/10.21437/Interspeech.2018-1568>
- Latif, Siddique, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir and Björn W. Schuller. (2023a). "Survey of Deep Representation Learning for Speech Emotion Recognition". *IEEE Transactions on Affective Computing*, 14(2), pp. 1634-1654. <https://doi.org/10.1109/TAFFC.2021.3114365>
- Latif, Siddique, Aun Zaidi, Heriberto Cuayáhuil, Fahad Shamshad, Moazzam Shoukat, Muhammad Usama and Junaid Qadir. (2023b). "Transformers In Speech Processing: A Survey". *Computer Science – Computation and Language*. arXiv preprint. <https://doi.org/10.48550/arXiv.2303.11607>
- Lian, Hailun, Cheng Lu, Sunan Li, Yand Zhao, Chuangao Tang and Yuan Zong. (2023). "A Survey Of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, And Face". *Entropy*, 25-10(1440). <https://doi.org/10.3390/e25101440>
- Liebenthal, Einat, David A. Silbersweig and Emily Stern. (2016). "The Language, Tone and Prosody of Emotions: Neural Substrates and Dynamics of Spoken-Word Emotion Perception". *Frontiers in Neuroscience*, 10(506). <https://doi.org/10.3389/fnins.2016.00506>
- Livingstone, Steven R., and Frank A. Russo. (2018). "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English". *PLoS One*, 13-5(e0196391). <https://doi.org/10.1371/journal.pone.0196391>
- Matt, Siddesh Gaddadevara, and S. Mani Sekhar. (2026). "Speech emotion recognition with Wav2vec and deep gated recurrent units". *Multimedia Tools and Applications*, 85(176). <https://doi.org/10.1007/s11042-026-21280-z>

- Mares, Alex, Gerardo Díaz-Arango, Jorge Pérez-Jácome-Friscione, Hector Vázquez-Leal, Luis Hernández-Martínez, Jesús Huerta-Chua, Andrés Jaramillo-Alvarado and Alfonso Domínguez-Chávez. (2025). "Advancing Spanish Speech Emotion Recognition: A Comprehensive Benchmark of Pre-Trained Models". *Applied Sciences*, 15-8(4340). <https://doi.org/10.3390/app15084340>
- Morrison, Murray, Linda Rammage, Hamish Nichol, Bruce Pullan, Phillip May and Lesley Salkeld. (1994). "Chapter 10 - Anatomy and physiology of voice production". In *The Management Of Voice Disorders*. Springer, pp.161-200. https://doi.org/10.1007/978-1-4899-2903-7_10
- National Institute on Deafness and Other Communication Disorders (NIDCD). (n. d. a). *Taking Care Of Your Voice*. <https://www.nidcd.nih.gov/health/taking-care-your-voice>
- National Institute on Deafness and Other Communication Disorders (NIDCD). (n. d. b). *Vocal Fold Paralysis*. <https://www.nidcd.nih.gov/health/vocal-fold-paralysis>
- Neumann, Michael, and Ngoc Thang Vu. (2017). "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech". *Proceedings of Interspeech 2017*, pp. 3672 -3676. <https://doi.org/10.21437/interspeech.2017-917>
- Ortuño, Rubén, José M. Sánchez, Diego Álvarez, Miguel López and Fernando León. (2020). "Neurometrics applied to banknote and security features design". Documentos Ocasionales, 2008, Banco de España. <https://repositorio.bde.es/handle/123456789/10447>
- Park, Daniel S., William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk and Quoc V. Le. (2019). "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition". *Proceedings of Interspeech 2019*, pp. 2613-2617. <https://doi.org/10.21437/Interspeech.2019-2680>
- Pepino, Leonardo, Pablo Riera and Luciana Ferrer. (2021). "Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings". *Computer Science – Sound*. arXiv preprint. <https://doi.org/10.48550/arXiv.2104.03502>
- Phukan, O., A. Buduru and R. Sharma. (2023). "Transforming the embeddings: A lightweight technique for speech emotion recognition tasks". *Proceedings of Interspeech 2023*, pp. 1903-1907. <https://doi.org/10.21437/Interspeech.2023-2561>
- Qu, Leyuan, Wei Wang, Cornelius Weber, Pengcheng Yue, Taihao Li and Stefan Wermter. (2022). "Improving speech emotion recognition with unsupervised speaking style transfer". arXiv:2211.08843. <https://doi.org/10.48550/arxiv.2211.08843>
- Rabiner, Lawrence. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". *Proceedings of the IEEE*, 77(2), pp. 257-286. <https://doi.org/10.1109/5.18626>
- Ramdani, Fatwa. (2025). "Natural Language Processing And Sentiment Analysis". In *Data Science: Foundations And Hands-On Experience*. Springer, pp. 359-394. https://doi.org/10.1007/978-981-96-4683-8_9
- Satt, Aharon, Shai Rozenberg and Ron Hoory. (2017). "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms". *Proceedings of Interspeech 2017*, pp. 1089-1093. <https://doi.org/10.21437/Interspeech.2017-200>
- Schewski, Lilien, Mathew Magimai Doss, Guido Beldi and Sandra Keller. (2025). "Measuring negative emotions and stress through acoustic correlates in speech: A systematic review". *PLoS One*, 20-7(e0328833). <https://doi.org/10.1371/journal.pone.0328833>
- Shevhuk, Ihor, and Iryna Dumyn. (2024). "Emotion-Based Voice Control for IoT: Enhancing Smart Device Interaction with Speech Emotion Classification". *CEUR Workshop Proceedings*, 3974. <https://ceur-ws.org/Vol-3974/short04.pdf>
- Trigeorgis, George, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou and Björn Schuller. (2016). "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network". *IEEE ICASSP 2016*. <https://doi.org/10.1109/ICASSP.2016.7472669>
- Van Mersbergen, Miriam, and Alexis E. Payne. (2021). "Cognitive, Emotional, and Social Influences on Voice Production Elicited by Three Different Stroop Tasks". *Folia Phoniatrica et Logopaedica*, 73(4), pp. 326-334. <https://doi.org/10.1159/000508572>
- Van Puyvelde, Martine, Xavier Neyt, Francis McGlone and Nathalie Pattyn. (2018). "Voice Stress Analysis: A New Framework for Voice and Effort in Human Performance". *Frontiers in Psychology*, 9(1994). <https://doi.org/10.3389/fpsyg.2018.01994>
- Wagner, Johannes, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben and Björn W. Schuller. (2023). "Dawn Of The Transformer Era In Speech Emotion Recognition: Closing The Valence Gap". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9), pp. 10745-10759. <https://doi.org/10.1109/TPAMI.2023.3263585>
- Wang, Yan, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge and Wei Zhang. (2022). "A Systematic Review On Affective Computing: Emotion Models, Databases, And Recent Advances". *Computer Science – Multimedia*. arXiv preprint. <https://doi.org/10.48550/arXiv.2203.06935>
- Yan, Susu, Liang Ye, Shuai Han, Tian Han, Yue Li and Esko Alasaarela. (2020). "Speech Interactive Emotion Recognition System Based on Random Forest". *2020 International Wireless Communications and Mobile Computing Conference*, pp. 1194-1198. <https://doi.org/10.1109/IWCMC48107.2020.9148117>
- Yang, Shu-wen, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed and Hung-yi Lee. (2021). "SUPERB: Speech Processing Universal PERFORMANCE Benchmark". *Proceedings of Interspeech 2021*, pp. 1194-1198. <https://doi.org/10.21437/Interspeech.2021-1775>
- Zhang, Zhaoyan. (2016). "Mechanics of human voice production and control". *The Journal of the Acoustical Society of America*, 140(4), pp. 2614-2635. <https://doi.org/10.1121/1.4964509>

Annex: List of Acronyms

AER	Automatic Emotion Recognition	MFCCs	Mel-Frequency Cepstral Coefficients
AI	Artificial Intelligence	MLP	Multilayer Perceptron
APA	American Psychological Association	NIDCD	National Institute on Deafness and Other Communication Disorders
ASR	Automatic Speech Recognition	NLP	Natural Language Processing
CNN / CNNs	Convolutional Neural Network(s)	openSMILE	Toolkit name; widely known as "open-source Speech and Music Interpretation by Large-space Extraction"
ComParE	(Computational Paralinguistics "COMPARE") feature set used with openSMILE*	PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses (quality-assessment/reporting framework)
EEG	Electroencephalography	PTM / PTMs	Pretrained Model(s)
EmoDB	Berlin Database of Emotional Speech	RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
eGeMAPS	extended Geneva Minimalistic Acoustic Parameter Set	RMS	Root Mean Square (energy)
F0	Fundamental frequency	RNN / RNNs	Recurrent Neural Network(s)
GFCCs	Gammatone Frequency Cepstral Coefficients	SDK	Software Development Kit (as in "Amazon Chime SDK ...")
GMMs	Gaussian Mixture Models	SER	Speech Emotion Recognition
GRADE	Grading of Recommendations Assessment, Development and Evaluation (quality-assessment framework)	SSL	Self-supervised learning
GRU	Gated Recurrent Unit	SVM / SVMs	Support Vector Machine(s)
HCI	Human Computer Interaction	SUPERB	Benchmark name for speech tasks
HMI	Human Machine Interaction	UCLA	University of California, Los Angeles (appears as "UCLA Linguistics")
HRI	Human Robot Interaction	U.S.	United States (in "U.S. Department of Health & Human Services")
HMMs	Hidden Markov Models	VAE / VAEs	Variational Autoencoder(s)
HuBERT	Self-supervised speech model name; commonly "Hidden-Unit BERT"	WavLM	Self-supervised speech model name
IEMOCAP	Interactive Emotional Dyadic Motion Capture (dataset)	wav2vec 2.0	Self-supervised speech model name
IoT	Internet of Things	ZCR	Zero-Crossing Rate
k-NN / KNN	k-Nearest Neighbours		
LPCCs	Linear Predictive Cepstral Coefficients		
LSP	Line Spectral Pairs		
LSTM	Long Short-Term Memory		

OCCASIONAL PAPERS

- 2430 JOSE GONZÁLEZ MÍNGUEZ: The Letta report: a set of proposals for revitalising the European economy. (There is a Spanish version of this edition with the same number).
- 2431 MARIYA MELNYCHUK and JAVIER MENCÍA: A taxonomy of macro-financial risks and policies to address them.
- 2432 DMITRY KHAMETSHIN, DAVID LÓPEZ RODRÍGUEZ and LUIS PÉREZ GARCÍA: El mercado del alquiler de vivienda residencial en España: evolución reciente, determinantes e indicadores de esfuerzo.
- 2433 ANDRÉS LAJER BARON, DAVID LÓPEZ RODRÍGUEZ and LUCIO SAN JUAN: El mercado de la vivienda residencial en España: evolución reciente y comparación internacional.
- 2434 CARLOS GONZÁLEZ PEDRAZ, ADRIAN VAN RIXTEL and ROBERTO PASCUAL GONZÁLEZ: Navigating the boom and bust of global SPACs.
- 2435 PATROCINIO TELLO-CASAS: El papel de China como acreedor financiero internacional.
- 2436 JOSÉ RAMÓN MARTÍNEZ RESANO: CBDCs, banknotes and bank deposits: the financial stability nexus.
- 2501 PEDRO DEL RÍO, PAULA SÁNCHEZ, MARÍA MÉNDEZ, ANTONIO MILLARUELO, SUSANA MORENO, MANUEL ROJO, JACOPO TIMINI and FRANCESCA VIANI: La ampliación de la Unión Europea hacia el este: situación e implicaciones para la economía española y la Unión Europea.
- 2502 BANCO DE ESPAÑA: In-person access to banking services in Spain. 2024 monitoring report. (There is a Spanish version of this edition with the same number).
- 2503 ANDRÁS BORSOS, ADRIAN CARRO, ALDO GLIELMO, MARC HINTERSCHWEIGER, JAGODA KASZOWSKA-MOJSA and ARZU ULUC: Agent-based modeling at central banks: recent developments and new challenges.
- 2504 ANDRES ALONSO-ROBISCO, ANDRES AZQUETA-GAVALDON, JOSE MANUEL CARBO, JOSE LUIS GONZALEZ, ANA ISABEL HERNAEZ, JOSE LUIS HERRERA, JORGE QUINTANA and JAVIER TARANCON: Empowering financial supervision: a SupTech experiment using machine learning in an early warning system.
- 2505 JÉSSICA GUEDES, DIEGO TORRES, PAULINO SÁNCHEZ-ESCRIBANO and JOSÉ BOYANO: Incertidumbre en el mercado de bonos: una propuesta para identificar sus narrativas con GDELT.
- 2506 LAURA JIMENA GONZÁLEZ GÓMEZ, FERNANDO LEÓN, JAIME GUIXERES PROVINCIALE, JOSÉ M. SÁNCHEZ and MARIANO ALCANIZ: Evolución de la investigación neurocientífica del efectivo: revisión y perspectivas actuales.
- 2507 LUIS FERNÁNDEZ LAFUERZA, IRENE ROIBÁS and RAQUEL VEGAS SÁNCHEZ: Indicadores de desequilibrios de precios del mercado inmobiliario comercial.
- 2508 PANA ALVES and OLIVIER HUBERT: ¿Influye la eficiencia energética en el precio de la vivienda en España?.
- 2509 ALEJANDRO FERRER and ANA MOLINA: The interaction of liquidity risk and bank solvency via asset monetisation mechanisms. (There is a Spanish version of this edition with the same number).
- 2510 ISABEL ALCALDE and PATRICIA STUPARIU: Financial education at an early age. (There is a Spanish version of this edition with the same number).
- 2511 ALEJANDRO GONZÁLEZ FRAGA, AITOR LACUESTA GABARAIN, JOSÉ MARÍA LABEAGA AZCONA, MARÍA DE LOS LLANOS MATEA ROSA, SOLEDAD ROBLES ROMERO, MARÍA VALKOV LORENZO and SERGIO VELA ORTIZ: Estructura del mercado de electrolineras.
- 2512 FERNANDO ARRANZ GOZALO, CLARA I. GONZÁLEZ MARTÍNEZ and MERCEDES DE LUIS LÓPEZ: Sovereign assets and sustainable and responsible investment: the importance of climate metrics. (There is a Spanish version of this edition with the same number).
- 2513 IRMA ALONSO-ÁLVAREZ and DANIEL SANTABÁRBARA: Decoding Structural Shocks in the Global Oil Market.
- 2514 JOSÉ MANUEL CARBÓ, CLAUDIA TOLEDO and ÁNGEL IVÁN MORENO: Hacia un diccionario panhispánico de sentimiento de la estabilidad financiera.
- 2515 IGNACIO FÉLEZ DE TORRES, CLARA I. GONZÁLEZ MARTÍNEZ and ELENA TRIEBSKORN: The puzzle of forward-looking climate transition risk metrics.
- 2516 MIGUEL GARCÍA-POSADA: Un análisis de los efectos de la introducción del procedimiento especial de insolvencias para microempresas sobre la propensión a concursar.
- 2517 ADRIAN VAN RIXTEL: Whatever it takes? Economic policymaking in China in the context of a possible deflationary spiral.
- 2518 PATRICIA STUPARIU and JUAN RAFAEL RUIZ: Adding up the benefits: education, numeracy and financial literacy. (There is a Spanish version of this edition with the same number).
- 2519 DAVID CUBERES, AITOR LACUESTA, MARÍA DE LOS LLANOS MATEA and DANIEL OTO-PERALÍAS: El efecto de la regulación sobre el tamaño de las plantas fotovoltaicas.
- 2520 MARINA GARCÍA GIL and DMITRY KHAMETSHIN: Ayudas directas de la línea COVID-19: los determinantes de la asignación y el efecto sobre el crédito bancario.
- 2521 LUCÍA CUADRO-SÁEZ, CORINNA GHIRELLI, MAXIMILIANO MORENO-LÓPEZ and JAVIER J. PÉREZ: Monitoring and forecasting food prices in the euro area.

- 2522 LAURA HOSPIDO, JÚLIA MARTÍ LLOBET and CARLOS SANZ: ¿Qué políticas son efectivas para reducir la exclusión social? Evaluación de cinco proyectos piloto de inclusión social a través de ensayos aleatorizados.
- 2523 MARIO ALLOZA, MARÍA ELENA CRISTÓBAL RODRÍGUEZ, JULIA GARCÍA-ROYO DÍAZ, BEATRIZ GONZÁLEZ, ALBERTO MARTÍN DEL CAMPO SOLA, ANE MARTÍN UGARTE, ENRIQUE MORAL-BENITO and IRENE PINILLA MELGAREJO: Diagnóstico y consecuencias económicas del grado de competencia en las licitaciones públicas.
- 2524 ADRIÁN CARRO, JORGE E. GALÁN, ENRIC MARTORELL and RAQUEL VEGAS: A literature review on ex-ante and ex-post analysis of the implications of borrower-based macroprudential measures.
- 2525 PAULA SEMPERE and RAQUEL PANTOJA: Efectivo y *neuromarketing*: Concepto, funciones y aplicación.
- 2601 RODOLFO G. CAMPOS, JACOPO TIMINI, FRANCESCA VIANI and ELENA VIDAL: The EU-Mercosur agreement: analysis of its characteristics from a sectoral perspective. (There is a Spanish version of this edition with the same number).
- 2602 PABLO AGUILAR, CORINNA GHIRELLI and SAMUEL HURTADO: MTBE v2025: new version of the Quarterly Model of the Banco de España.
- 2603 IRMA ALONSO-ALVAREZ, EKATERINA BUKINA, MARINA DIAKONOVA, NINO KHITARISHVILI, JAVIER J. PÉREZ and PEDRO PIQUERAS: Geopolitical risk: a database of general and bilateral indices.
- 2604 RUBÉN DOMÍNGUEZ-DÍAZ, MARTA GARCÍA-RODRÍGUEZ, JAVIER QUINTANA and RUBÉN VEIGA-DUARTE: Estimación del crecimiento potencial de la economía española: una revisión metodológica.
- 2605 CARLES MANERA, FERRAN NAVINÉS, JAVIER FRANCONETTI, MIQUEL QUETGLAS and JOSÉ A. PÉREZ-MONTIEL: Regional outlook on labor productivity in Spain, 2000–2022.
- 2606 RICARDO BARAHONA: Current trends in performance and flows in the Spanish pension funds industry.
- 2607 FRANCISCO GONZÁLEZ RODRÍGUEZ, ALBERTO ORTOS TORRES, JOSÉ MARÍA SERENA GARRALDA and MIQUEL TARÍ SÁNCHEZ: Issuance yield of MREL-eligible bank debt: a benchmarking model.
- 2608 MÓNICA CORREA-LÓPEZ, MAR DELGADO-TÉLLEZ and MARTA SUÁREZ-VALERA: Europa en transición energética: respuestas y desafíos tras dos crisis consecutivas.
- 2609 AITOR LACUESTA, MARÍA VALKOV and MICOLE DE VERA: Un análisis de los mecanismos de protección de los operadores económicos establecidos por la Ley de garantía de la unidad de mercado.
- 2610 Encuesta Financiera de las Familias (EFF) 2024: métodos, resultados y cambios desde 2022.
- 2611 ALFONSO CAMBA, MATÍAS COVARRUBIAS, JOSÉ LUÍS RODRÍGUEZ and RAQUEL TÁRREGA: El coste en bienestar de los márgenes en mercados de productos en España: un análisis con empresas heterogéneas.
- 2612 MICAELA ARIAS, LUCÍA CRIADO, JAVIER GARCÍA-VERDUGO, EDUARDO GUTIÉRREZ, ALMUDENA KESSLER, JOSÉ MANUEL MONTERO, SARA PEREIRA and PAU ROLDÁN-BLANCO: Evolución de la estructura de mercado y de los indicadores de competencia en España en las dos últimas décadas.
- 2613 ANDRÉS ALONSO-ROBISCO, JOSÉ MANUEL CARBÓ, CARLOS JOSÉ GARCÍA, JORGE QUINTANA and JAVIER TARANCÓN: Retriever or reasoner? Decomposing retrieval-augmented generation performance in external audit supervision.
- 2614 RUBÉN VEIGA-DUARTE, EDUARDO GUTIÉRREZ, ALBERTO MARCOS-FERNÁNDEZ, ENRIQUE MORAL-BENITO and JAVIER QUINTANA: Las fuentes del crecimiento en España y las principales economías europeas: productividad total de los factores y heterogeneidad sectorial.
- 2615 MATEO CACHO UZAL, M. DANIELA FILIP and DIANA POSADA RESTREPO: Regional convergence dynamics in the main european economies.
- 2616 IRMA ALONSO-ALVAREZ, IVÁN KATARYNIUK, DAVID LÓPEZ-SALIDO and JAVIER J. PÉREZ: Geopolitical Oil-Supply Scenarios and Macroeconomic Risks: A Bayesian Toolkit for Policy Analysis.
- 2617 MIHAI-VASILE CÎRJĂ and JORDI ROMEU GRANADOS: Transparency and accountability as drivers of cultural change. A comparative analysis of central banks and national competent authorities.
- 2618 LUCÍA CANALEJO GUIJOSA, FERNANDO LEÓN MARTÍNEZ and JOSÉ MARÍA SÁNCHEZ ECHAVE: The voice as a neurophysiological marker: applications in implicit emotion perception for Neurocash® and cash design.