ORIGINAL ARTICLE

Syllabic-scale deep learning for detecting ineffective communication in neonatal simulation-based training

Gianpaolo Coro · Armando Cuttano · Serena Bardelli

Received: 17 March 2026 / Accepted: 12 July 2026

Abstract
Simulation-based training is widely used in neonatology to prepare healthcare professionals for high-risk clinical situations. In this context, effective communication and teamwork are critical for patient safety. However, evaluating communication quality during simulation sessions often relies on time-consuming manual review of audio and video recordings by expert trainers. Moreover, the poor acoustic quality of recordings and the presence of overlapping speech and background noise make conventional speech-recognition-based analyses unreliable. To address these limitations, this paper proposes an automated workflow that detects acoustic-prosodic anomalies associated with potentially ineffective communication in neonatal simulation sessions using syllabic-scale acoustic analysis and unsupervised deep learning. The approach extracts super-segmental features inspired by psychoacoustic principles and models them through deep neural architectures to identify anomalous dialogue segments associated with stress, uncertainty, elevated vocal effort, or inefficient interaction patterns. Unlike traditional methods relying on speech transcription or phonetic-level processing, the proposed system operates directly on syllabic-scale acoustic patterns, making it more robust to noisy and heterogeneous recording conditions. The workflow was evaluated across multiple neonatal simulation case studies and compared with a traditional cluster-analysis baseline. The proposed method achieved an average accuracy of 87.7%, with a maximum of 91.2%, representing improvements of up to 59% over the baseline approach. The system also demonstrated robustness to high noise levels and produced training-quality assessments showing strong agreement with expert evaluations. By automatically identifying acoustic-prosodic indicators associated with potentially ineffective communication, the proposed approach can support objective review and feedback processes in neonatal simulation-based training.

Keywords Artificial Intelligence · Healthcare · Neonatology · Psychoacoustics

1 Introduction

Training through simulation is a widely adopted educational methodology that recreates realistic scenarios in controlled environments, allowing trainees to practise technical and non-technical skills without risk to patients [1–4]. In medicine, simulation-based training plays a crucial role in preparing healthcare professionals to manage

Published online: 04 August 2026

complex, high-risk situations while maintaining patient safety. Simulation enables learners to translate theoretical knowledge into practice through hands-on activities and structured feedback [5–7].

Neonatology is one of the medical specialities in which simulation-based training is particularly valuable [8]. Neonatologists are responsible for the care of newborn infants, especially those with critical conditions such as prematurity, congenital anomalies, or respiratory failure. In these contexts, rapid decision-making and coordinated team actions are essential to ensure successful clinical outcomes. Because neonatal emergencies often require simultaneous interventions from multiple professionals, effective teamwork and communication are as important as individual technical competence [9–12]. In fact, evidence suggests that optimal clinical outcomes depend on procedural expertise and the quality of collaboration and communication within the medical team [8].

Simulation-based neonatal training is therefore needed to develop technical proficiency and also to improve so-called *human factors*, including communication, situational awareness, leadership, and teamwork. Human factors are non-technical skills that influence how healthcare professionals interact with equipment, clinical environment, and with each other during high-pressure situations [13–15]. Ineffective communication, such as unclear instructions, misunderstandings, or emotionally charged interactions, may compromise team coordination and delay critical interventions, such as neonatal resuscitation [16–18]. For this reason, trainers pay close attention to analysing team dynamics during simulation sessions to identify communication patterns that may hinder effective clinical performance.

Currently, the evaluation of simulation sessions most often relies on expert observation and manual analysis of audio and video recordings. Trainers review these recordings to identify communication patterns, decision-making processes, and behavioural responses to stressful scenarios. However, this approach is time-consuming and strongly dependent on expert interpretation. As simulation programmes expand and large volumes of training data become available, automated tools capable of analysing team interactions can provide valuable support to instructors by highlighting relevant segments of a session and offering objective insights into communication dynamics [19–21].

Recent advances in Artificial Intelligence (AI) provide new opportunities to analyse complex human interactions in clinical training environments. In particular, speech processing techniques and automatic speech recognisers (ASRs) can extract meaningful information from audio recordings of simulation sessions, enabling the detection of patterns associated with stress, uncertainty, or ineffective communication. Such approaches have the potential to complement expert evaluation by providing quantitative indicators of communication behaviour and facilitating the rapid exploration of ineffective communication [22–26]. However, the use of speech recognition, phonetic-scale signal analyses, and text-based processing presents several drawbacks when applied to spoken dialogues recorded during training-through-simulation sessions. A primary limitation concerns the very low audio quality typically encountered in these environments. The signal-to-noise ratio is often below 20 dB in environmental recordings contaminated by white, violet, and heterogeneous background noise from clinical equipment. Moreover, low-volume recordings are frequent, especially in historical data collections. In some cases, speech can be barely intelligible, severely degrading the performance of transcription systems. As a result, even state-of-the-art ASRs perform poorly, with full sentences transcribed incorrectly and many critical words missing from the output [27–29] (as also demonstrated in Sect. 3.1). This scenario makes any analysis relying on textual transcription inherently unreliable. These limitations arise because ASRs - both traditional and end-to-end architectures - depend heavily on the phonetic structure of speech, which becomes distorted under high-noise conditions, leading transcription-based approaches to fail in realistic medical simulation environments. Furthermore, the intrinsic characteristics of dialogue during emergency simulations further complicate transcription: speech often includes overlapping utterances, emotionally charged expressions (e.g., stress or anger), rapid speaking tempo during critical interventions, and long silences followed by short bursts of speech, all of which significantly reduce ASR accuracy [28, 30, 31]. Finally, there may be high variability across recordings, including differences in team size, gender composition, noise conditions, and interaction frequency. Developing a robust supervised model based on speech transcription would therefore require a very large annotated corpus covering

all these conditions, which is difficult and expensive to obtain. For these reasons, approaches relying solely on speech recognition or phonetic-level processing are not well-suited to analysing spoken communication in noisy and heterogeneous simulation environments.

Alternatively, psycho-acoustics-inspired speech processing has been proposed based on syllabic-scale, super-segmental, and rhythmic components, which are more robust to noise and better aligned with emotions and human factors [32, 33]. Although these techniques cannot provide a sufficiently detailed phonemic structure to allow for transcribing utterances, they have great potential for detecting speech alterations and disfluencies, which often manifest as irregular rhythms or pauses rather than segmental errors [34]. Moreover, they can improve speech recognition stability across different noise conditions by mimicking crucial underlying processing strategies of the auditory cortex, which relies on slower rhythmic structures to support perception under challenging conditions [35, 36].

In this study, we propose an automated workflow, based on super-segmental features and syllabic-scale rhythmic components of speech, designed to support simulation-based training in neonatology by identifying segments of dialogue that may contain *ineffective communication* among medical team members. Unlike earlier approaches (Sect. 2) based primarily on phonetic features and clustering techniques, or on transcriptions, the present work explores advanced modelling strategies that process syllabic-scale features using unsupervised deep learning. By processing these features and self-adapting to the audio recorded from a training session, the workflow can capture richer patterns in speech dynamics without requiring phonetic-level features, thereby addressing the challenge posed by adverse noise and recording conditions.

The workflow was conceived to analyse audio recordings from neonatal simulations (e.g., resuscitation and critical scenario management) to identify dialogue segments characterised by acoustic and conversational patterns associated with stress, misunderstanding, uncertainty, or inefficient interaction. It particularly aims to detect dialogue anomalies associated with ineffective communication and provide trainers with more informative summaries of simulation sessions. Moreover, it estimates an overall quality assessment of the training session based on the proportion of ineffective to effective verbal communication.

Notably, the proposed system does not directly model ineffective communication in a semantic or interactional sense. Instead, it detects alterations in acoustic-prosodic patterns at the syllabic scale that correlate with interactional difficulties, and interprets them as indicators of potentially ineffective communication. This operational definition reflects the constraints of the analysed scenarios - where severe noise conditions and overlapping speech make reliable transcription and semantic analysis infeasible - but improves automation.

Automating the identification of potentially problematic interactions and suggesting an overall quality assessment can (i) significantly reduce the time trainers spend analysing simulation sessions, (ii) enhance the evaluation of team communication in neonatal teams, and (iii) support trainers in improving both individual and collective performance during high-risk clinical scenarios.

To summarise, this work presents a recording-adaptive, unsupervised deep learning workflow for detecting ineffective communication in neonatal simulations under acoustic constraints. Key innovations include the use of syllabic-scale, psychoacoustically motivated super-segmental features within a two-step anomaly-detection framework that can estimate an interpretable overall training quality score. The main contribution of this work lies in the integration of established techniques into a coherent, recording-adaptive framework tailored for the analysis of communication in highly noisy simulation environments.

The paper is organised as follows: Section 2 reports on current techniques used to enhance communication in medical teams and the issues modern technology faces. Section 3 describes the case studies, the workflow, and the evaluation methodology. Section 4 reports the results of the application of our workflow to the case studies. Finally, Sect. 5 discusses the results and draws conclusions.

2 Overview

Automatic systems for detecting ineffective verbal communication are increasingly discussed in the context of optimising training through simulation with data-driven feedback. Alongside high-fidelity manikins, multi-camera recording, and structured debriefing, recent work has explored AI-assisted debriefing and multimodal learning analytics. These techniques help instructors rapidly identify crucial moments in a training session (e.g., teamwork and communication events) and support debriefing through summaries or dashboards that aid manual review [21, 23, 37–40].

In this context, *ineffective verbal communication* can be operationally defined in terms of its prosodic and acoustic profile and its behavioural correlates [28, 41, 42]. One common signature is high syllabic-scale (100–250 ms) speech energy, indicating increased speech intensity: louder-than-average speech, stronger energy peaks, abrupt energy bursts, and increased vocal effort are often consistent with anger, stress, fear, agitation, or over-activation during emergencies. A second signature is high syllabic-scale pitch (elevated fundamental frequency and more extreme intonation contours): higher average F_0, sharper tone, increased pitch variability, and a raised voice quality are often associated with urgency, anxiety, irritation, insistence, and emotional tension. When high energy and high pitch co-occur, the result is an intensified prosodic pattern that frequently accompanies escalating stress, misunderstandings, cognitive overload, timeline pressure, and breakdowns in controlled, assertive coordination such as medical training. Evidence across speech and emotion research supports the relation between negative affect or stress and overall increases in intensity, volume and pitch [42–47]. Although ineffective communication is broader than negative emotion, in simulation practice the two often intersect (e.g., during stress escalation and misunderstandings), making prosody, and generally rhythmic and *super-segmental* features (i.e, speech attributes that extend over units larger than a phonetic-scale) a natural and often more noise-tolerant target than lexical content. By contrast, *effective verbal communication* in simulation settings is typically characterised as assertive, controlled, clear, and intonation-stable. Therefore, exaggerated prosody and intensity can serve as practical indicators of potentially ineffective speech.

Modern automated approaches to analysing ineffective verbal communication in spoken dialogues usually combine several layers of signal and human interaction modelling. One family of methods emphasises prosodic and paralinguistic analyses (energy, pitch contours, rhythm, speech rate, turn-taking dynamics), sometimes explicitly at syllabic or sub-utterance time scales to better capture short bursts, tension, and rapid interaction shifts [42, 48–50]. A second family uses machine learning for detecting emotional, stress, and cognitive load, where prosodic features often play a central role and can be integrated with phonetic-scale (10–20 ms) spectral features or learned representations [51–54]. A third family utilises multimodal systems that combine audio with video, proximity, positioning, or interaction traces to produce relevant indicators for the trainers (e.g., participation balance and temporal patterns of collaboration) and to support debriefing with structured summaries [19, 23, 55].

A persistent challenge is that text-based processing and modern ASRs often fail in the conditions where ineffective communication matters most: noisy, heterogeneous, multi-speaker, emotionally charged emergency dialogue. ASR performance drops with background noise, overlapping speech, rapid tempo, and domain-specific jargon, especially when the system relies on phonetic structure that noise and overlapping speech distort [33, 56–58]. Studies benchmarking ASRs in simulated emergency settings have highlighted clinically relevant limitations under realistic acoustic conditions [59, 60]. For this reason, many of these systems have restricted transcription to better-quality segments, added noise-robust adaptation and re-training strategies, or reduced dependence on transcription by focusing on prosody and voice activity patterns.

Above all, psychoacoustic theories can inform automatic systems why super-segmental cues remain useful in detecting ineffective verbal communication in noisy environments and how to engineer robust detectors. Psychoacoustics emphasises human sensitivity to temporal modulations, loudness, and pitch cues that remain perceivable even when fine phonetic details are masked. ASR and speech processing systems based on psychoacoustic

theories have often adopted design choices that operate at syllabic time scales and prioritise energy and pitch dynamics [33, 34, 61, 62]. These outcomes have been particularly used for noise-robust modelling and for producing alternative representations for recognition and analysis in the presence of additive noise, reinforcing the use of prosody-first or syllable-scale pipelines when transcription is unreliable [32–34].

A significant technological help in this context has been recently provided by self-supervised and representation learning approaches, sometimes reused from different application domains [63–65]. Models such as Wav2Vec 2.0 [66, 67], HuBERT [68, 69], and related embedding-based architectures [70, 71] can learn high-level representations of speech directly from large-scale unlabelled corpora, enabling robust downstream tasks including speech recognition, speaker identification, and paralinguistic analysis. These representations have also been employed in anomaly detection settings [72, 73], where deviations are identified in the learned embedding space rather than in handcrafted acoustic features. Modern anomaly detection methods for audio have explored a wide range of techniques, including deep autoencoders, contrastive learning frameworks, density estimation in latent spaces, and hybrid architectures combining temporal modelling with deep generative models [74–76]. These approaches often rely on large and diverse training datasets to learn generalisable representations of normal behaviour. However, the applicability of such methods to the detection of ineffective verbal communication is constrained by several factors. First, the available data consist of a limited number of highly heterogeneous recordings acquired under severe noise conditions, which differ substantially from the clean or moderately noisy corpora typically used to train self-supervised models [77, 78]. Second, large annotated or pretraining datasets are unavailable, especially for low-resource languages [79]. Therefore, the methods should be able to adapt directly to the acoustic characteristics of each session. For these reasons, this paper proposes a methodology that adopts a recording-adaptive unsupervised strategy based on syllabic-scale acoustic features and generative modelling, which provides robustness under extreme acoustic variability while remaining applicable in data-poor scenarios.

2.1 Main novelties of this work

The present work builds on previous efforts to incorporate psychoacoustic principles into speech processing. It addresses the challenge of detecting ineffective verbal communication in recordings from medical training through simulation under particularly complex conditions. These conditions include highly variable noise environments, very low signal-to-noise ratios (down to approximately 4 dB), and multiple speakers with overlapping speech. The case studies were derived from real-world hospital simulations and belonged to a valuable historical collection routinely used by medical trainers to support the optimisation of learning outcomes (Sect. 3.1). In these scenarios, approaches based on text processing or phonetic-scale signal processing were not applicable because the phonetic structure of speech was heavily masked by noise and environmental disturbances (Sect. 3.1.2). In contrast, supra-segmental cues remained relatively robust and therefore provided a viable basis for analysis.

Within this context, the proposed approach adopts syllabic-scale acoustic representations, including Modulation Spectrogram features and super-segmental energy–pitch descriptors, which are more resilient to noise and better aligned with perceptual mechanisms related to speech dynamics. Since no annotated training set was available, the workflow was designed as a recording-adaptive, unsupervised system that self-adjusts to the acoustic characteristics of each session. In addition, the method produces an interpretable Training Quality Score, providing trainers with a concise summary of communication effectiveness within a simulation.

Overall, the contributions of the proposed workflow can be grouped into two main categories: (i) methodological contributions, and (ii) task-specific integration for neonatal simulation analysis. From a methodological perspective, the main contribution lies in the design of a recording-adaptive, unsupervised anomaly-detection workflow operating on syllable-scale acoustic features. This process combines modulation-based representations and energy–pitch features within a two-stage deep learning-based architecture that first operates noise removal and then detects communication-related anomalies. From an application perspective, the novelty lies in integrating these components into a coherent workflow - tailored to highly noisy medical simulations - that includes

tone-unit segmentation, anomaly aggregation, and the definition of an interpretable Training Quality Score to assess overall training effectiveness.

In summary, the contribution of the proposed system is not the introduction of a new deep learning architecture *per se*, but rather the principled combination of psychoacoustically motivated features and adaptive anomaly detection techniques within a domain-specific framework designed for challenging real-world conditions.

3 Methods

This section outlines the details of our workflow. A general conceptual schema is illustrated in Fig. 1, which specifies the two main phases: noise removal and ineffective verbal communication detection. Figures 2 and 3 provide detailed representations of the workflow components within each phase.

The following sections first describe the data for which the workflow was designed (Sect. 3.1). Next, they outline the acoustic units (Sect. 3.2) and features (Sect. 3.3) utilised. Subsequently, Sects. 3.4 and 3.5 describe the two main phases of the workflow. Finally, Sect. 3.6 explains the evaluation methodology used to assess the workflow performance.

3.1 Data

The foundation for the development of our workflow was the historical audio-recording archive of the Centro di Simulazione e Formazione Neonatale (Centro NINA) of the Santa Chiara Hospital in Pisa, Italy. This public research centre is dedicated to training neonatologists through simulation. The archive serves as a valuable resource for developing new communication protocols and conducting comprehensive analyses of prevalent communication errors and learning obstacles among trainees. The recordings used in this study were collected during simulation activities conducted within the Dipartimento Materno Infantile of the Azienda Ospedaliero-Universitaria Pisana for neonatology speciality training. The simulations were conducted in a dedicated room at the Santa Chiara Hospital of Pisa. This room was equipped with a realistic neonatal island, a delivery room, and clinical tools designed to enhance participants' emotional engagement and provide a lifelike context for the procedures being practised. The simulated scenarios focused primarily on neonatal resuscitation, specifically addressing cases involving a newborn experiencing respiratory failure or a hypotonic newborn. Trainees interacted with a neonatal simulation manikin (Laerdal's SimNewB [80]), which replicated identical critical clinical scenarios for all participants. The management of SimNewB was conducted via Laerdal's LLEAP simulation software, which simulated clinical states and was equipped with sensors for recording the effectiveness of clinical interventions, subsequently providing feedback to the medical team. The feedback and interactions were entirely a consequence of the trainees' interventions and the simulator's automatic feedback mechanism. Recording of the sessions was facilitated by Laerdal's SimView software and used one environmental microphone (Behringer B2 PRO) positioned 1.5 meters above the simulation manikin, centrally located in the room, with no noise filtering applied.

The training sessions were organised into three primary phases. The initial phase involved trainees assisting an expert trainer who managed a simulated emergency scenario. The expert trainers were certified by NRP-AHA (Neonatal Resuscitation Program of the American Heart Association) and SIN (Societá Italiana Neonatologia). During the second phase, the trainees were tasked with managing the simulated situation independently, without the trainer's support. Following this session, the trainer conducted a debriefing session to analyse and evaluate the simulation outcomes. The third phase, conducted two to three months later without a prearranged date, required trainees to repeat the simulation with a trainer present among the assistants. This phase was particularly instrumental in assessing the overall level of expertise attained by the trainees, thereby making it more suitable for evaluating the performance of our workflow.

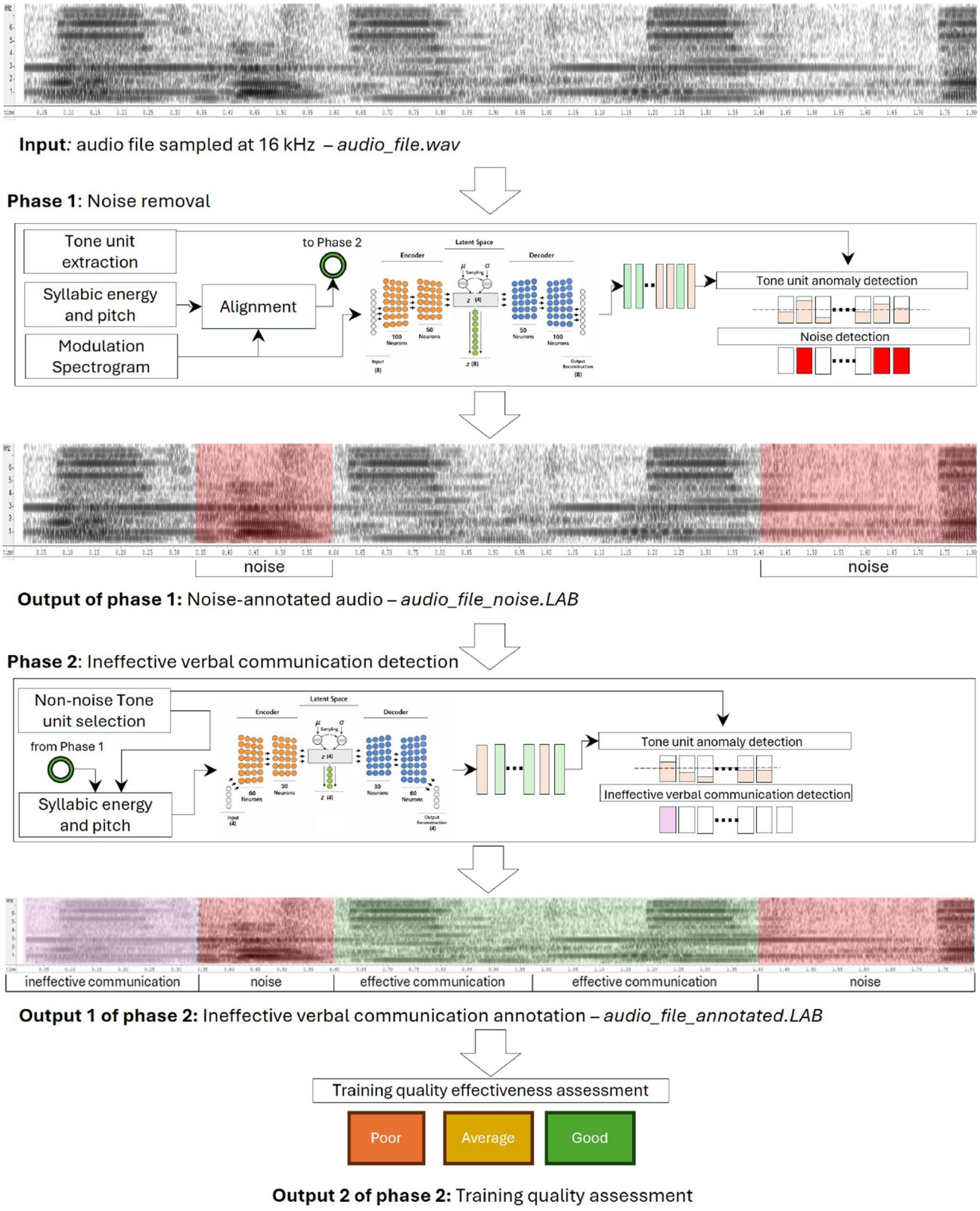

Fig. 1 Conceptual schema of the proposed workflow

https://doi.org/10.1007/s00521-026-12319-1

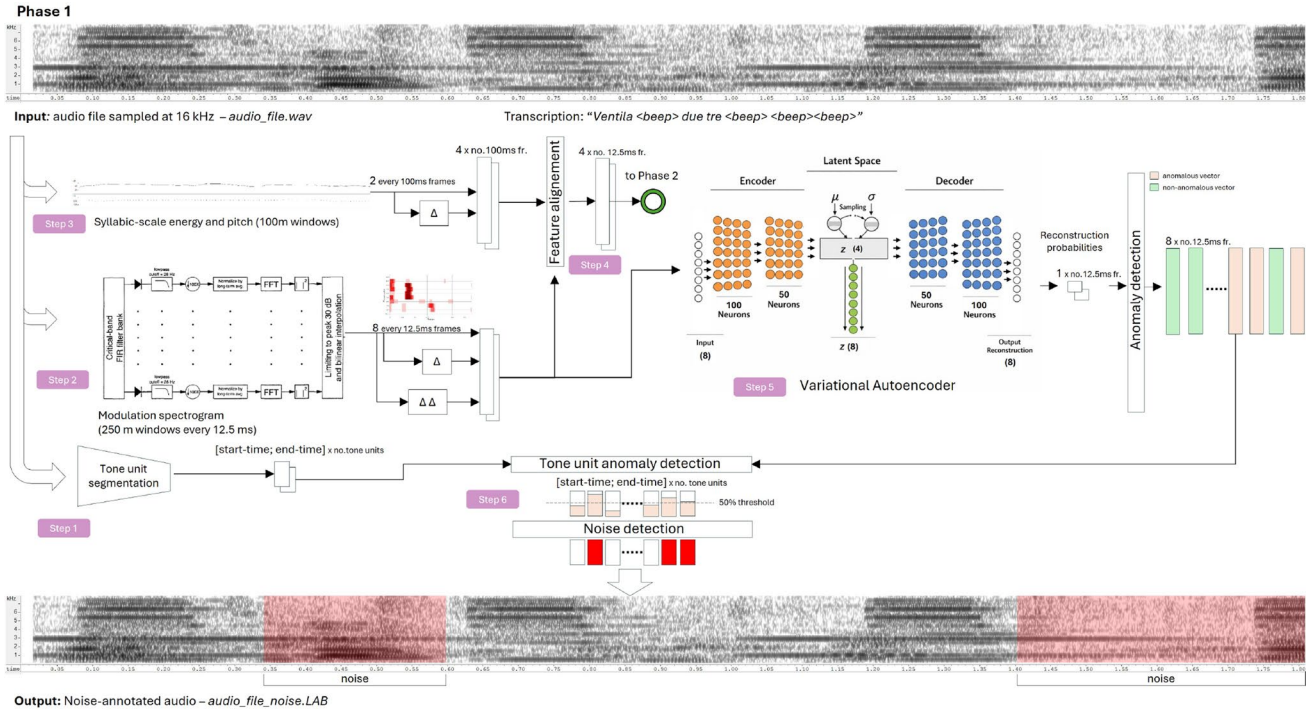

Fig. 2 First phase of our workflow, addressing noise removal

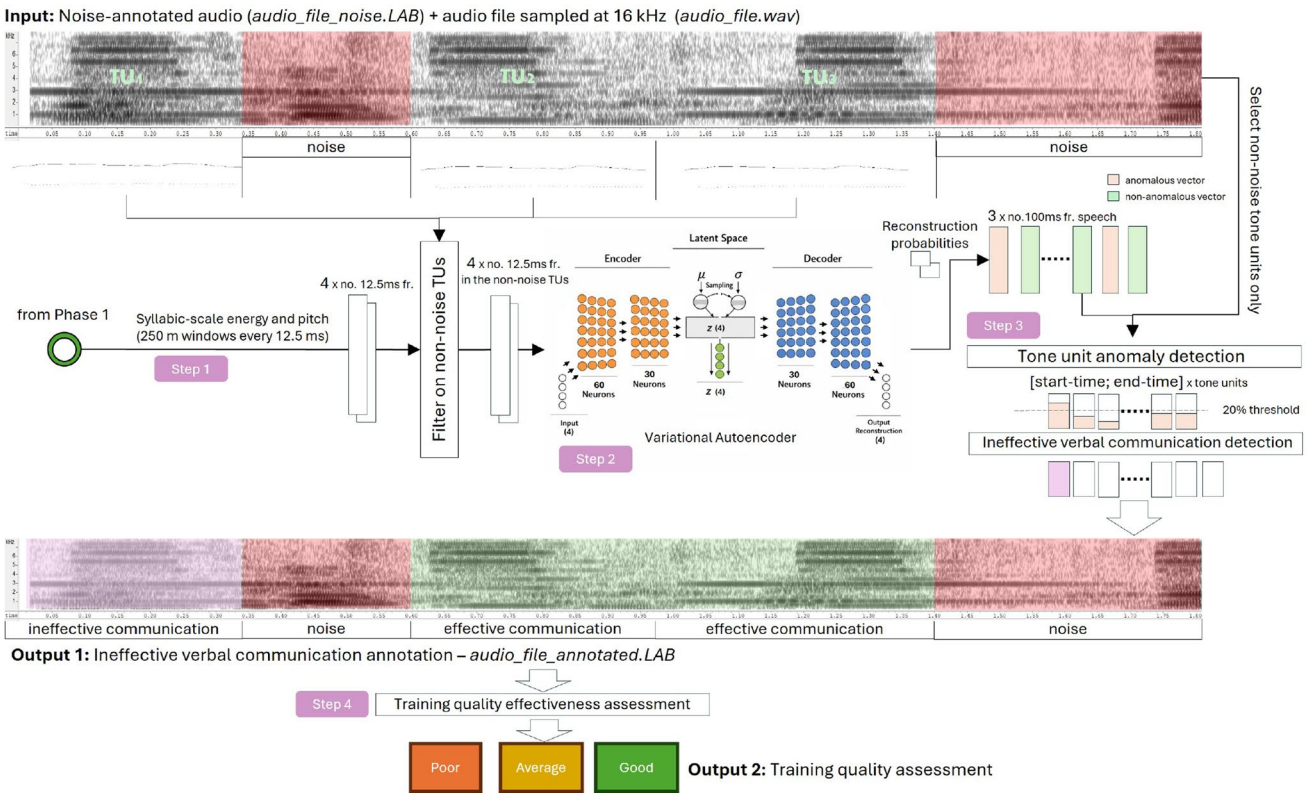

Fig. 3 Second phase of our workflow, addressing ineffective verbal communication detection

The hospital was requested to provide audio recordings of these final sessions for use as case studies, contingent on confirming that the informed consent obtained from the trainees was compliant with the study's objectives. Ultimately, a total of 10 sessions were procured (Table 1), corresponding to those also employed in [28], for a total duration of approximately 79 minutes, captured under various high-noise-level conditions (ranging from 4.63 to 14.17 SNR) with a sampling frequency of 16 kHz. The recordings featured diverse team and gender compositions (two or three team members, male and female trainees), with a total of 10 speakers across the case studies. They represented common conditions and recording settings of historical audio recordings from hospital archives.

3.1.1 Data preparation

The objective of our workflow was to identify and tag audio segments, from each case study recording, characterised by ineffective verbal communication among team members, based on the presence of at least one of the following *negative evaluation factors*:

- High emotional and psychological stress
- Communication not directed towards the appropriate team member
- Non-assertive or aggressive communication style
- Uncontrolled emotional content and verbal intonation
- Uncertainty within the communication
- Confusion about the technical terminology
- Lack of confirmation regarding actions required from team members
- Excessive use of the first person plural
- Insufficient leadership exhibited by the medic managing the operation
- Discussion of strategies during emergency situations
- Inadequate level of expertise concerning specific tasks.

Table 1 Summary of the case studies involved in our work, along with the indications of (i) the type of simulation managed by the medical team, (ii) the estimated signal-to-noise ratio (SNR), (iii) the noise conditions, (iv) the audio durations, (v) the number of speakers involved and (vi) their sexes, the (vii) average (avg.) tone unit (TU) duration, (viii) the average duration of the TUs containing speech, and (ix) the average duration of the TUs containing ineffective verbal communication

Case	Simulation type	SNR (dB)	Noise conditions	Duration (minutes)	No. speakers	Participant sexes	Avg. TU duration (s)	Avg. TU duration of speech (s)	Avg. TU duration of ineffective verbal communication (s)
T1	Newborn with respiratory failure	14.2	white (low level)	11	2	Female	1.18	3.37	4.69
T2	Hypotonic newborn	6.8	white (high level)	6.5	2	Female	1.17	3.41	4.10
T3	Newborn with respiratory failure	11.0	white and violet	8	2	Female	1.07	2.65	3.44
T4	Newborn with respiratory failure	11.1	white and blue	7	2	Female	1.08	2.64	3.10
T5	Hypotonic newborn	6.8	white and violet (high level)	9	2	Female	1.09	2.33	3.33
T6	Hypotonic newborn	11.5	violet (low level)	6.5	3	Female and Male	1.09	2.54	2.58
T7	Hypotonic newborn	12.3	violet (low level)	6.5	3	Female (2) and Male (1)	1.02	2.57	2.57
T8	Newborn with respiratory failure	4.6	white (high level)	9.5	2	Female	1.03	2.15	2.51
T9	Newborn with respiratory failure	4.8	white (high level)	7.5	3	Female (2) and Male (1)	1.09	2.51	3.24
T10	Hypotonic newborn	10.1	white (low level)	7.5	2	Male	1.02	2.05	1.94

While these factors help define ineffective communication, not all can be observed through acoustic analysis alone. For example, confusion over terminology and discussions during emergencies entail semantic information that would require reliable speech transcription. Given the operational scenario constraints, this study specifically focuses on capturing the ineffective communication that emerges in acoustic-prosodic features, such as deviations in energy, pitch, and rhythm, which indicate stress and interactional disruption.

A gold standard of annotated audio segments was established, encompassing 4,385 audio segments across the case studies, each corresponding to a tone unit (as detailed in Sect. 3.2). Each segment was manually classified into one of three categories: *ineffective communication* (at least one negative evaluation factor present), *noise* (machine or background noise and non-speech audio), or *effective communication* (otherwise). A collection of randomly selected segments totalling 5 minutes from the corpus (approximately 6%) was used as a *development set* for calibration of workflow parameters.

Furthermore, since the workflow was also designed to yield an overall quality assessment categorised as *poor*, *average*, or *good*, each case study was assessed by two experts, each from distinct areas of specialisation. Expert 1 possessed expertise in psychoacoustics and speech annotation and evaluated the cases through the lens of emotional factors, stress levels, disfluencies, uncertain speech, and acoustic cues pertinent to the aforementioned negative evaluation factors. In contrast, Expert 2 brought experience in simulation-based training in neonatology, assessing the training sessions from a content-oriented perspective, specifically regarding whether trainees met the anticipated level of expertise. Both experts formulated their evaluations (*poor*/*average*/*good* training effectiveness) by assigning scores on a 5-point Likert scale to each evaluation factor and then subjectively determining the overall training effectiveness based on the cumulative score.

3.1.2 Speech recognition performance on the case studies

The high and heterogeneous noise levels present in the case studies, together with spontaneous and overlapping speech and the complexity of the real-world simulation environment, prevented the reliable use of automatic speech transcriptions for downstream analysis based on text-derived features such as sentiment, emotion, or keyword extraction. Our tests on this corpus indicated that the state-of-the-art OpenAI Whisper v3-large model (1.6 billion parameters) achieved 42.9% word accuracy, 14.1% substitution rate, 14.1% deletion rate, and 132.9% insertion rate. Under these conditions, the probability of correctly capturing a critical keyword was only 18.4%. Although transcription performance has approximately doubled compared with five years ago - when the Google ASR system achieved 23.3% word accuracy on the same corpus [28] - these results remain insufficient to support the development of a reliable text-based classification system. These observations justify the search for alternative solutions that do not rely on phonetic-scale recognition to process the presented case studies and, more generally, historical audio recordings from hospital archives.

3.2 Tone unit segmentation

Ineffective verbal communication can be investigated within dialogue units that contain finite and meaningful interactions between speakers. An acoustic correlate of such a dialogue unit is the *tone unit* (TU), defined as a portion of speech produced under a coherent intonation contour, representing a prosodically organised segment of discourse [81, 82]. Within the prosodic hierarchy, tone units correspond to short speech segments characterised by relatively stable patterns of energy and pitch, typically concluding with a marked decrease in energy or a boundary-related prosodic cue [83, 84].

Tone units frequently contain a complete or semantically coherent utterance, although they may also correspond to structured non-speech events such as short noise bursts or pauses occurring within conversational exchanges. For this reason, they provide a convenient and linguistically meaningful representation for analysing the quality of verbal communication in conversational settings. TU-based segmentation can help isolate salient

speech segments, discard noisy or irrelevant portions of the signal, and consequently improve the robustness of downstream processing tasks of ASRs or dialogue analyses [28, 32, 34, 85].

The automatic detection of tone units can be made more robust to noisy conditions by operating at a *syllabic temporal scale*. In particular, the use of analysis windows in the range of 100–250 ms allows the capture of acoustic events associated with syllabic nuclei that remain relatively stable even when the phonetic structure of speech is partially masked by noise [28, 34, 86–88]. Such syllabic-scale processing is therefore well suited for analysing conversational speech recorded in challenging environments.

As a first step, the proposed workflow (Fig. 2-step 1) segments the acoustic signal into TUs. This segmentation is performed using a modified version of an adaptive algorithm also employed in other speech recognition and speech processing studies [28, 34, 89]. The algorithm applies an iterative adaptation procedure that starts from an initial energy-drop threshold, expressed as a percentage parameter, to identify the end of a TU. This threshold is progressively increased until a minimum expected number of TUs is detected in the audio signal. Therefore, a TU is identified as the termination of a sequence of relatively high-energy speech followed by a significant energy drop. Previous studies have shown that this approach is robust for speech segmentation, particularly under heterogeneous and noisy recording conditions [87].

In the present implementation, the method was re-adapted by defining low-energy regions as those falling below the 25th percentile of the energy distribution of the analysed signal and using a relative energy drop threshold. The algorithm pseudocode is reported in the following:

1. Load the input audio recording
2. Calculate the audio-signal energy time series over time-frame windows of *time-frame-length* duration (provided as an input parameter)
3. Calculate *low energy* as the 25th percentile of the energy distribution
4. Calculate the derivative of the energy time series
5. For each time frame:
   If
   (i) the energy derivative is negative *AND*
   (ii) energy is below *low energy AND*
   (iii) the relative energy drop from the previous time frame is higher than an *energy drop threshold* (input parameter)
   → mark the energy window start time as a tone unit end and the beginning of a new tone unit
6. If at least a *minimum-TU* (input parameter) number of TUs has been found → report all TU markers and terminate the processing
7. Otherwise, decrease the *energy drop threshold* by 10% and go to step 5.

Algorithm 1 Tone unit segmentation algorithm

Based on the analysis of the development set, our workflow used a 200 ms *time-frame-length* (to likely catch syllabic-duration silences, in line with previous applications of the algorithm [28]), a starting 60% *energy drop threshold* (to account for high background noise), and 3 *minimum TU*s to search.

Across the analysed case studies, the average TU duration of segments containing ineffective verbal communication was often longer than that of segments containing spoken dialogue (Table 1). Moreover, the overall average TU duration was generally shorter than that of TUs containing speech. This pattern suggests that noise events - such as machine-generated, coloured, or white noise - were frequently captured as individual tone units when they did not overlap with speech, as they exhibited well-defined energy profiles that satisfied the tone unit segmentation criteria.

3.3 Acoustic features

Syllables play a central role in speech organisation and production, as they are associated with respiratory muscle impulses, peaks of sonority, and rhythmic energy patterns in the speech signal. Although a precise acoustic definition of the syllable remains elusive, syllables constitute a fundamental unit in human speech perception, and speakers appear to possess an intuitive awareness of syllabic structure [90, 91].

Within a spoken utterance, a syllable is typically organised around a *nucleus*, which represents the most intense and acoustically stable component of the segment across different speech realisations. The nucleus is usually preceded by an *onset*, characterised by increasing energy, and may terminate with a descending portion known as the *coda*. Because of the inherent variability of spontaneous speech, this canonical structure is not always clearly preserved. Nevertheless, syllables that approximate this structure (*prominent* syllables) often carry important information for both human and automatic speech recognition processes [92, 93]. Prominent syllables tend to preserve the rhythmic organisation of speech. They are typically less reduced than surrounding segments and therefore contribute significantly to maintaining intelligibility along the speech chain [94]. Generally, the importance of prominent and non-prominent syllables in speech perception extends beyond their phonetic composition. Evidence suggests that there exist temporal dynamics of speech at the syllabic scale that encode information that goes beyond fine phonetic details, supporting robust speech perception even when phonetic cues are partially degraded [32, 36, 95, 96].

For these reasons, and because the phonetic structure of speech in our case studies was heavily compromised by environmental noise, our workflow adopts a syllabic-scale analysis approach spanning 250 ms frames, without relying on fine phonetic details. The next sections describe the two principal types of features used in the proposed workflow, which play distinct and complementary roles. The first type, the Modulation Spectrogram (Sect. 3.3.1), is primarily used in the first phase of the workflow to distinguish speech from non-speech segments, as it provides a robust representation of slow temporal modulations that are resilient to noise. The second type, energy and pitch (Sect. 3.3.2), together with their temporal variations, is used in the second phase to detect alterations in speech behaviour associated with potentially ineffective communication. While the Modulation Spectrogram features are well-suited for identifying speech activity, energy-pitch features are more sensitive to prosodic changes, such as intensity and intonation variations, which are indicative of stress, uncertainty, or interactional disruption. This separation of roles reflects the different nature of the two tasks addressed by the workflow: noise removal and communication analysis.

3.3.1 Modulation Spectrogram

Several studies have investigated acoustic representations of syllables that do not rely on their phonetic structure. One of the most influential approaches was proposed by Greenberg and Kingsbury through the Modulation Spectrogram (MS) representation [35]. The MS provides a noise-robust, syllable-centred representation of speech based on slow temporal modulations of the acoustic signal rather than fine spectro-temporal details. It has been used in diverse applications, including ASR performance improvement in adverse noise conditions [36, 97, 98], speech intelligibility evaluation [99], acoustic scene classification [100], and even for fake speech detection [101].

The MS emphasises the low-frequency portion of the modulation spectrum (0–8 Hz), with a peak sensitivity around 4 Hz, which approximately corresponds to the typical syllabic rate of speech. This representation captures acoustic properties consistent with articulatory dynamics (typically 2–12 Hz movement rates), auditory cortical sensitivity to slow temporal modulations (below approximately 20 Hz), and the statistical organisation of spoken language. The MS complies with the hypothesis that speech perception relies heavily on low-frequency modulation patterns and syllabic-scale temporal organisation.

One of the key assumptions underlying this approach stems from psychoacoustic experiments demonstrating that speech remains intelligible to humans even when only slow temporal modulations below approximately 6

Hz are preserved [102]. These low-frequency modulation cues contribute significantly to the robustness of human speech recognition in noisy and reverberant environments [36, 103].

The MS is computed through the following processing pipeline (Fig. 2-step 2): First, the signal is passed through an 8-channel filterbank centred on Greenwood's critical frequencies [104], approximating the frequency sensitivity of the human auditory system:

$$s_k(t) = s(t) * h_k(t), \quad k = 1, \ldots, 8,$$

where $h_k(t)$ represents the impulse response of the *k*-th band-pass filter.

The output of each filterbank channel is then processed independently. The signal envelope is extracted using half-wave rectification followed by low-pass filtering ($\mathrm{LPF}_{28\mathrm{Hz}}$) with a cutoff frequency of 28 Hz:

$$e_k(t) = \mathrm{LPF}_{28\mathrm{Hz}}(\max(s_k(t), 0)).$$

The resulting envelope is downsampled to 80 Hz and normalised by its average level. The temporal modulations of the normalised envelope are then analysed by computing the Fast Fourier Transform (FFT) over a 250-ms Hamming window with a 12.5 ms shift. Formally, for each frequency *f* of the spectrum and time frame (being *t* the central time of the window), the following squared magnitude is computed:

$$E_k(f, t) = \mid \mathcal{F}\{e_k(\tau)\}_{\tau \in [t - T/2,\, t + T/2]} \mid^2$$

where *T* denotes the duration of the window (250 ms), and $\mathcal{F}\{\cdot\}$ denotes the Fourier transform.

Finally, the squared magnitude of the 4 Hz component of the FFT (expressed in dB) is retained as the MS feature associated with that channel:

$$\mathrm{MS}_k(t) = E_k(f_m, t),$$

where $f_m \approx 4$ Hz.

The features across all channels capture the spectral components of slow speech modulations at approximately 4 Hz (250 ms), which are closely related to syllabic-scale speech dynamics. This process thus produces a time series of 8-dimensional vectors at 12.5 ms intervals, with each vector representing the MS features of a 250 ms frame. Formally, the process yields an 8-dimensional feature vector $\bar{x}(t) = [\mathrm{MS}_1(t), \ldots, \mathrm{MS}_8(t)]$ for each time frame.

MS features proved particularly suitable for our analysis because they facilitated discrimination between speech and non-speech segments in recordings characterised by very high, heterogeneous noise levels. As illustrated in Fig. 4, the MS representation of speech segments exhibited distinctive patterns compared with noisy segments. In general, speech segments showed higher MS intensity in both lower- and higher-frequency bands, often spanning multiple channels, whereas machine and background noise showed more uniform or band-limited patterns. White noise, in particular, exhibited very low modulation intensity across all channels. Across the analysed case studies, the lowest-frequency band (band 1, approximately centred on 210 Hz) showed the largest difference, with an average MS intensity increase of +93% in speech segments compared to non-speech segments. This discrepancy was followed by band 2 (+80%, approximately at 290 Hz), band 8 (+70%, approximately at 2400 Hz), band 4 (+51%, approximately at 560 Hz), band 7 (+25%, approximately at 650 Hz), and band 5 (+18%, approximately at 800 Hz). In contrast, bands 6 (approximately at 1150 Hz) and 3 (approximately at 400 Hz) did not exhibit statistically significant differences according to a Welch t-test.

These observations motivated the development of a pattern classification model capable of distinguishing speech from non-speech segments within the proposed workflow, as described in Sect. 3.4.

 https://doi.org/10.1007/s00521-026-12319-1

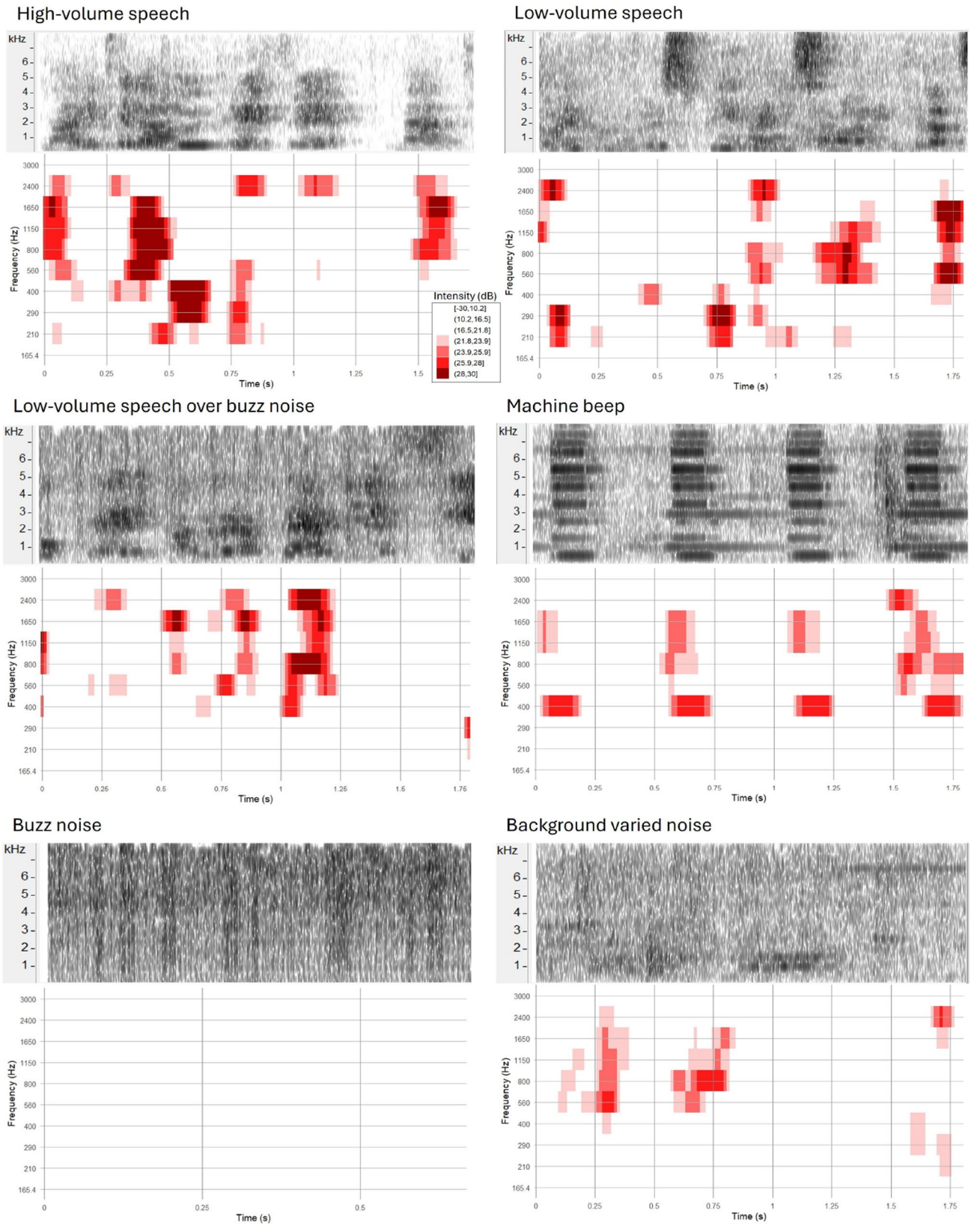

Fig. 4 Comparison of the traditional spectrograms, based on 32 ms-length windows (upper grey-coloured images), and the Modulation Spectrograms, based on 250 ms-length windows (lower red-coloured images), across six audio samples presenting different conditions of speech volumes and noise levels and types. The Modulation Spectrogram legend is the same for all charts

3.3.2 Energy and pitch

The standard short-window spectrograms in our case studies showed high background noise levels, making it difficult to reliably detect ineffective verbal communication. For this reason, our workflow focused on extracting syllabic-scale features that captured the rhythmic organisation of the acoustic signal, specifically the energy and pitch contours (Fig. 2-step 3), which reflected the intonation and emphasis patterns of the utterances.

Syllabic-scale energy was computed as the discrete-time signal power, expressed as the root mean square (RMS) of the sample amplitudes within a 100-ms analysis window:

$$E[n] = \sqrt{\frac{1}{N} \sum_{k=-\lfloor N/2 \rfloor}^{\lfloor N/2 \rfloor - 1} s^2[n+k]},$$

where n denotes the central sample index of the analysis window, $s[k]$ denotes the discrete-time audio signal, and N corresponds to the number of samples within the analysis window. These values were calculated over consecutive frames to produce a temporal short-syllabic-scale energy contour of the signal.

Syllabic-scale pitch corresponds to the fundamental frequency F_0 of a tone associated with a syllabic segment, and its temporal evolution reflects the musicality and intonation of the speech signal. In our workflow, pitch was estimated using an autocorrelation-based sound-to-pitch algorithm [105]. This algorithm computes the short-time autocorrelation function

$$R_s[\tau; n] = \sum_{k=0}^{N-1} s[n-k]\, s[n-k-\tau],$$

where τ is the lag in samples, and N and n are defined as they are for energy. The fundamental period $\tau_0[n]$ is then estimated as the lag that maximises the autocorrelation within a predefined range:

$$\tau_0[n] = \arg \max_{\tau \in [\tau_{\min}, \tau_{\max}]} R_s[\tau; n].$$

The pitch of the window n is finally obtained as:

$$F_0[n] = \frac{f_s}{\tau_0[n]},$$

where f_s is the sampling frequency.

The algorithm was applied with a frequency search range of 50–500 Hz and an analysis window of 100 ms. Frames for which pitch could not be reliably estimated were assigned a value of zero.

As suggested by previous studies [28], computing energy and pitch at a 100-ms temporal resolution through the methods described allows the capture of syllabic-level prosodic patterns. At this scale, the resulting trajectories typically exhibit autocorrelation structures that reflect rhythmic regularities and help highlight prominent syllables, whereas analyses performed at larger temporal scales may smooth out these syllabic dynamics. However, to ensure consistency with the MS features and enable analysis of larger syllable-scale modulations of energy and pitch, these features were time-aligned with the MS features (Fig. 2-step 4). The alignment was performed using a standard frame-averaging procedure. Specifically, for each MS frame - computed every 12.5 ms over a 250 ms analysis window - the energy and pitch frames falling within the temporal span of the MS frame were selected, and the average energy and average pitch were calculated. This process produced two synchronised sequences of syllabic-scale, super-segmental acoustic features, one with 8 MS features and the other with 2 energy-pitch features.

3.4 Noise removal

Our workflow includes a pipeline that processes Modulation Spectrogram features from the input audio recording to distinguish anomalous from non-anomalous TUs. The underlying assumption is that anomalous TUs may correspond to audio segments dominated by non-speech events, as supported by the analysis presented in Sect. 3.3.1. As explained in the following sections, the pipeline consists of two components: one that detects anomalous MS feature vectors (anomaly detection) and another that detects anomalous TUs based on these vectors (noise detection).

3.4.1 Anomaly detection

The workflow incorporates a deep-learning model that captures non-linear relationships among the elements of each feature vector and infers the latent variables that may have generated the observed MS features (Fig. 2-step 5). Specifically, this model is a Variational Autoencoder (VAE) [106], which estimates the probability distribution of latent variables $\bar{z}$ using an unsupervised probabilistic Artificial Neural Network (ANN) and then reconstructs the input vectors through another ANN.

Generally, a VAE models the distribution $p(\bar{x})$ of the input vector $\bar{x}$ by projecting it into a lower-dimensional latent space defined by $\bar{z}$. In our case, $\bar{x}$ corresponded to the 8 MS features associated with a 250 ms audio segment, with one vector of 8 features provided every 12.5 ms. From the latent representation $\bar{z}$, the VAE attempts to reconstruct the original input vector. If the latent variables successfully capture the generative factors underlying the feature vectors, the reconstructed vectors will closely approximate the original inputs. The feature vectors that cannot be accurately reconstructed likely correspond to atypical or anomalous patterns in the data.

A VAE internally consists of two ANNs: an *Encoder* and a *Decoder*. The Encoder maps each input vector $\bar{x}$ to the parameters (means and standard deviations) of a multivariate Gaussian distribution over the latent variables. This distribution approximates the posterior distribution of the latent variables given the input ($q(\bar{z} \mid \bar{x})$). The second network, the *Decoder*, reconstructs the input vector by generating a reconstructed vector $\tilde{x}$ from a latent sample $\bar{z}$ drawn from the encoded distribution (i.e., $\bar{z} \sim q(\bar{z} \mid \bar{x})$). The Decoder also estimates the likelihood $p(\bar{x} \mid \bar{z})$, i.e., the probability of observing the input vector given the latent variables. This probabilistic reconstruction mechanism enables the identification of anomalous vectors as inputs that the model cannot accurately reconstruct [107–113]. Specifically, the reconstruction probability of $\bar{x}$ is approximated by marginalising over the latent space:

$$p(\bar{x}) = \int p(\bar{x} \mid \bar{z})\, q(\bar{z} \mid \bar{x})\, d\bar{z}.$$

This quantity is estimated via Monte Carlo sampling by drawing N_s samples (16 in our application) $\{\bar{z}^{(i)}\}_{i=1}^{N_s} \sim q(\bar{z} \mid \bar{x})$, yielding:

$$\hat{p}(\bar{x}) = \frac{1}{N_s} \sum_{i=1}^{N_s} p(\bar{x} \mid \bar{z}^{(i)}).$$

After computing $\hat{p}(\bar{x})$ for all feature vectors, a distribution of reconstruction probabilities is obtained. Anomaly detection is then performed by selecting the vectors whose reconstruction probabilities fall below a given percentile of this distribution. Feature vectors with low reconstruction probability are interpreted as poorly explained by the learned latent structure and can therefore be considered atypical with respect to the dominant acoustic patterns of the recording. This decision rule provides a recording-adaptive definition of anomaly, where deviations are identified relative to the statistical properties of the analysed signal.

During training, the weights of both the Encoder and Decoder are optimised to maximise the reconstruction likelihood over the input vectors. This optimisation is achieved by minimizing a loss function composed of two terms: (i) the *Reconstruction Loss*, which measures the difference between the input vector $\bar{x}$ and its reconstruction $\tilde{x}$ (commonly using mean squared error or cross-entropy), and (ii) the *Kullback–Leibler (KL) divergence*, which regularizes the latent representation by measuring the divergence between the learned posterior distribution $q(\bar{z} \mid \bar{x})$ and a standard multivariate Gaussian prior with mean vector $\bar{\mu} = (0, \ldots, 0)$ and standard deviation vector $\bar{\sigma} = (1, \ldots, 1)$, where $\dim(\bar{\mu}) = \dim(\bar{\sigma}) = \dim(\bar{z})$. Formally, the VAE is trained by maximising the evidence lower bound (ELBO), which corresponds to minimising the following loss function:

$$\mathcal{L}(\bar{x}) = -\mathbb{E}_{q(\bar{z}|\bar{x})} \left[\log p(\bar{x} \mid \bar{z})\right] + D_{\mathrm{KL}}\left(q(\bar{z} \mid \bar{x}) \,\|\, p(\bar{z})\right)$$

where the first term represents the *Reconstruction Loss* and the second term is the *KL divergence* between the approximate posterior and the prior $\bar{z}$ distribution. This objective encourages the model to accurately reconstruct typical input patterns while regularising the latent space.

In our workflow, the optimal VAE architecture was predetermined through an empirical optimisation procedure performed on the development set. The optimal configuration in terms of the number of neurons and layers was selected using a *growing* strategy, whereby hidden nodes (up to 500 per layer) and hidden layers (up to 10 per ANN, allowing the exploration of both shallow and deep architectures) were progressively added until the configuration achieving the highest overall reconstruction probability across the input vectors was obtained. The optimal dimensionality of the latent variable space was determined by varying its size from 1 to $\dim(\bar{x}) = 8$ and selecting the value yielding the highest reconstruction probability. The network parameters were optimised using the Adam optimiser with Leaky ReLU activation functions in the hidden layers, after Xavier weight initialisation, over 100 training epochs.

The final architecture consisted of an Encoder with two hidden layers containing 100 and 50 neurons, respectively, followed by a latent space with the same dimensionality as the input feature vectors ($\dim(\bar{z}) = 8$). The Decoder mirrored the Encoder structure, with two hidden layers of 50 and 100 neurons, and produced the reconstructed feature vectors at the output layer.

This architecture was fixed for all applications, but the weights were randomly initialised before processing a new audio recording (i.e., no pre-training was used). During the execution of the workflow on an audio recording, the VAE adapts its weights based on MS features extracted from the analysed acoustic signal and detects anomalies in each recording independently. This design choice guarantees the model's unsupervised adaptation across very different audio-quality scenarios.

Consistent with related studies that have used VAEs for anomaly detection [112–116], our workflow performs anomaly detection by defining a threshold at the 25th percentile of the reconstruction probability distribution across the input vectors. Formally, it selects vectors - among all M vectors in the dataset - for which $\hat{p}(\bar{x}) < \theta$, with $\theta = \mathrm{Quantile}_{0.25}\left(\{\hat{p}(\bar{x}_i)\}_{i=1}^{M}\right)$. Feature vectors with reconstruction probabilities below this threshold are labelled as *anomalous*. The choice of the 25th percentile was empirically determined on the development set and is consistent with other practices in anomaly detection [112, 113], which use a fixed proportion of low-likelihood observations to identify atypical patterns in complex systems.

This approach thus defines "normality" relative to the dominant acoustic patterns within each analysed recording. Consequently, anomaly detection identifies deviations from within-session behaviour rather than departures from an external or absolute reference of effective communication. This design enables adaptation to highly heterogeneous and noisy conditions but implies that the system captures relative variations within each session.

As an output of this process, each aligned feature vector from the input recording is classified as either *anomalous* or *non-anomalous*.

3.4.2 Noise detection

Based on the anomalous feature vectors detected, our workflow identifies noise-related TUs through a two-step procedure (Figure 2-step 6). In the first step, for each TU, the number of anomalous feature vectors falling in the TU is computed. If this number is relatively high - indicating that a substantial portion of the TU consists of anomalous vectors - the tone unit is considered to mostly correspond to a non-speech audio segment because it would provide limited information for dialogue analysis.

In the second step, the proportion of anomalous audio duration relative to total TU duration is calculated. A threshold on this proportion is applied to identify anomalous TUs. Those exceeding the threshold are marked for *exclusion* from subsequent processing stages (i.e., as *noise*). The optimal threshold, estimated on the development set, was 50%. This threshold was identified as the one that excluded 100% of the development-set tone units dominated by non-speech frames while preserving segments containing relevant speech information.

As output from this phase, the workflow annotates the original audio file by generating a label file in LAB format, which is natively supported by the free WaveSurfer educational speech analysis software [117]. The label file is associated with the audio recording and contains lines specifying the start and end markers (in seconds) of the noise segments along with an annotation (e.g., "1.5 3.9 'noise' "). This simple, easy-to-parse format facilitates integration with other speech analysis tools and complex analytical pipelines. Furthermore, visualising the audio signal alongside the generated annotations enables trainers to rapidly navigate long recordings and quickly identify audio content during the most critical moments of the simulation, thereby supporting the preparation for the subsequent debriefing phase.

Overall, the noise removal process described so far can be summarised as follows:

1. Retrieve the MS feature vectors ($\dim(\bar{x}) = 8$, every 12.5 ms, each corresponding to a 250 ms frame)
2. Initialise a Variational Autoencoder with the following characteristics: Encoder with 100×50 nodes; Decoder with 50×100 layers; $\dim(\bar{z}) = 8$; and Xavier weight initialisation
3. Train the VAE on the input vectors
4. Calculate the reconstruction probability $p(\bar{x} \mid \bar{z})$ for each vector $\bar{x}$
5. Calculate the distribution of the reconstruction probability over all $\bar{x}$ vectors
6. Calculate the 25th percentile of the distribution and select the vectors $\bar{x}_{low}$ falling in this percentile
7. For each TU from the tone unit segmentation algorithm
   Calculate the number of $\bar{x}_{low}$ vectors falling in the TU
   Calculate the total duration ($D(TU_{low})$), in seconds, associated with these vectors (*non-speech* duration)
   Calculate the proportion between non-speech and total TU duration ($D_{ns} = \frac{D(TU_{low})}{D(TU)}$)
   If $D_{ns} > 50\% \rightarrow$ mark the TU as *noise*
9. Return the *noise*-TU time markers in a LAB file with corresponding 'noise' annotations
10. Return the *non-noise*-TU time markers to the next phase.

Algorithm 2 Noise removal algorithm

3.5 Ineffective verbal communication detection

To detect ineffective verbal communication within the *non-noise* tone units, the workflow relies on the evidence discussed in Sect. 2, which indicates that ineffective communication is reflected, at least partially, in measurable alterations of acoustic-prosodic features, such as increases in vocal intensity, loudness, and pitch. Accordingly,

the workflow does not directly identify communication errors at the semantic or interactional level, but instead includes a dedicated pipeline for detecting ineffective communication that searches for anomalies in syllabic-scale energy and pitch, as well as their temporal variations (Fig. 3). This approach provides an indirect yet robust proxy when higher-level linguistic analysis is unreliable.

This pipeline processes the energy and pitch (EP) features from phase 1, along with their respective first-order derivatives (deltas), within the *non-noise* TUs, yielding 4 features per 250 ms frame (with 12.5 ms frame overlap) (Fig. 3-step 1). These feature vectors are analysed to *identify* and *select* anomalies associated with alterations in syllabic-scale speech behaviour. The detection scheme follows a similar approach to the noise-detection stage. Specifically, the feature vectors are processed by a VAE, which attempts to reconstruct them (Fig. 3-step 2). Alterations in energy and pitch patterns are selected as anomalous vectors, characterised by low reconstruction probabilities that fall below the 25th percentile of the reconstruction probability distribution across the input vectors.

The optimal VAE architecture was identified through the same process used for noise detection, resulting in an Encoder with two hidden layers containing 60 and 30 neurons, a Decoder with a mirrored architecture (30–60), and a latent space with dimension 4. This architecture remained fixed, but the model was not pretrained because the workflow was designed to fully adapt this VAE to the analysed input audio.

The workflow subsequently maps the detected anomalous vectors to their corresponding TUs (Fig. 3-step 3). A threshold of 20% on the proportion of anomalous vectors within a TU is applied to identify those that likely contain speech alterations associated with ineffective verbal communication. These segments are then annotated and returned to the trainers for inspection. The 20% threshold was calibrated on the development set to maximise precision (i.e., to minimise false positives). It allowed the detection of consistent patterns of altered speech while reducing the impact of isolated acoustic fluctuations. Classifying ineffective communication at the TU level rather than the frame level relaxes the assumption that anomalies directly correspond to ineffective communication and provides a controllable parameter to adjust the sensitivity of the detection process. If anomalies are uncommon, the workflow will likely and correctly detect no TU associated with ineffective communication. However, if energy-pitch anomalies occur frequently, they may be partially normalised. In this case, the 20% threshold serves as a reasonable tolerance to include TUs with infrequent anomalies and consequently account for this normalisation.

As in the previous phase, the workflow produces an output label file containing annotations aligned with the original audio recording. These annotations (labelled 'ineffective communication') indicate audio intervals corresponding to ineffective verbal communication between medical team members.

In addition, the workflow computes an overall *Training Quality Score* (TQS) - expressed as a percentage - defined as

$$\mathrm{TQS} = \frac{D(\text{speech}) - D(\text{ineffective speech})}{D(\text{speech})} \cdot 100$$

where $D(\text{speech})$ denotes the total duration (in seconds) of TUs likely containing speech (i.e., excluding noise-related TUs), and $D(\text{ineffective speech})$ represents the total duration (in seconds) of TUs identified as containing ineffective verbal communication.

A linear classification over this score - set to give more tolerance to average effectiveness - transforms it into an interpretable label for the trainer (Fig. 3-step 4):

- $\mathrm{TQS} \leq 25\% \rightarrow$ *poor training effectiveness*
- $25\% < \mathrm{TQS} \leq 75\% \rightarrow$ *average training effectiveness*
- $\mathrm{TQS} > 75\% \rightarrow$ *good training effectiveness*

The TQS interval boundaries were defined to provide a balanced categorisation of training effectiveness, with wider tolerance for average performance, and were validated against expert assessments on the case studies.

Notably, this phase does not use the MS features, as they proved less effective than energy and pitch in detecting altered speech patterns (Sect. 4). This is not surprising, as the audio signal processed at this stage primarily consists of spoken dialogue, and the slow speech modulations captured by the MS features only partially reflect rapid rhythmic alterations. Moreover, since the energy-pitch frames had large spans (250 ms) and a high rate (12.5 ms), energy-pitch change velocity (deltas) could capture rapid shifts in syllabic energy and pitch, making acceleration (second-order derivatives, double deltas) less informative (as demonstrated in Sect. 4).

Overall, the ineffective verbal communication process can be summarised as follows:

1. Retrieve the EP+Δ feature vectors ($\dim(\bar{x}) = 4$) from the TUs, excluding the *noise*-TUs
2. Train a Variational Autoencoder (Encoder with 60×30 nodes; Decoder with 30×60 layers; $\dim(\bar{z}) = 4$; and Xavier weight initialisation) and select the vectors $\bar{x}_{low}$ falling in the 25th percentile of the reconstruction probability distribution
3. For each TU from the tone unit segmentation algorithm
   Calculate the total duration (in seconds) associated with the $\bar{x}_{low}$ vectors falling in the TU ($D(TU_{low})$) (*altered speech* duration)
   Calculate the proportion between *altered speech* and the total TU duration ($D_{as} = \frac{D(TU_{low})}{D(TU)}$)
   If $D_{as} > 0.2$ ($i.e., 20\%$) $\rightarrow$ mark the TU as *ineffective verbal communication*
5. Return the *ineffective verbal communication*-TU time markers in a LAB file with a corresponding 'ineffective communication' annotation
6. Calculate the *Training Quality Score* as $\text{TQS} = \frac{D(\text{speech}) - D(\text{ineffective speech})}{D(\text{speech})} \cdot 100$, and linearly interpret it as overall *poor/average/good* training effectiveness.
7. Return the annotation file, the TQS and its interpretation (in a text file) to the trainer.

Algorithm 3 Ineffective verbal communication detection algorithm

3.6 Evaluation methodology

As the first evaluation step, the performance of ineffective verbal communication (IVC) detection was assessed against manual TU-level annotations. Among the 4,385 TUs included in the test, those labelled as containing IVC by both the manual annotators and the automatic system were considered true positives (TP), whereas those labelled by both as noise or effective communication (i.e., non-IVC) were considered true negatives (TN). False positives (FP) corresponded to TUs incorrectly classified by the automatic system as containing IVC, while false negatives (FN) were TUs manually annotated as IVC but classified as non-IVC by the system.

Based on these indicators, accuracy was computed as

$$A = \frac{TP + TN}{TP + TN + FP + FN}.$$

Agreement between manual and automatic annotations was further assessed using Cohen's kappa coefficient [118], which measures accuracy relative to chance agreement (P_e):

$$\text{kappa} = \frac{A - P_e}{1 - P_e}$$

with $P_e = \frac{(TP+FP)(TP+FN)+(TN+FN)(TN+FP)}{(TP+TN+FP+FN)^2}$.

These metrics were compared with those reported in [28], which used cluster analysis to distinguish IVC from non-IVC segments on the same corpus.

As the second evaluation step, the automatic training-effectiveness interpretation was compared with the independent assessments of two expert evaluators across the case studies (Sect. 3.1). An agreement matrix involving the three assessors was constructed using Cohen's kappa values, with the resulting agreement levels interpreted according to the guidelines proposed by Fleiss [119].

The third evaluation step assessed the contribution of the Modulation Spectrogram features to the noise-removal procedure. Noise detection accuracy was first computed using this feature set. Then, the variation in noise-detection accuracy was measured when using the Modulation Spectrogram features in combination with deltas (MS change velocity and acceleration) and energy–pitch features, or using energy–pitch features alone, to quantify the relative contributions of the features.

Finally, the fourth evaluation step consisted of a sensitivity analysis of different feature combinations in the IVC-detection VAE model to verify that energy-pitch-delta features were optimal for this model when used after the noise-removal stage.

4 Results

The following sections report the results of the four evaluation steps described in Sect. 3.6.

4.1 Performance of the ineffective verbal communication detection

The performance of the proposed workflow across the case studies was generally high, particularly compared with the traditional cluster analysis approach used in [28] (Table 2). Accuracy ranged from a maximum of 91.17% with a Cohen's kappa of 0.61 (case study T1) to a minimum of 83.87% with a kappa of 0.53 (case study T10).

The relative accuracy gain over the cluster-analysis baseline, normalised by the performance of the proposed method ($(A_{new} - A_{baseline})/A_{new}) \times 100$) measured the proportion of the proposed method's achieved accuracy that would be lost if the baseline system was used instead. This quantity varied across the datasets, ranging from ∼15% (T2) to ∼59% (T9), thereby indicating that the advantage of the proposed workflow over the baseline was not uniform but became particularly substantial in some challenging scenarios, especially under highly adverse acoustic conditions.

Notably, in the case study with the highest noise level (T8), the workflow achieved an accuracy of 88.41% with *moderate* agreement with manual annotations (kappa $= 0.54$).

Table 2 Performance (accuracy and Cohen's kappa) of our workflow in detecting tone units that contain ineffective verbal communication across the case studies, along with indications of (i) the signal-to-noise ratio (SNR), (ii) the accuracy and agreement (with Fleiss' interpretation) achieved by our workflow, and (iii) by a baseline cluster analysis, and (iv) the accuracy loss if the baseline system was used instead (accuracy improvement relative to the proposed method's accuracy)

Test	SNR (dB)	Accuracy (%)	Agreement (kappa)	Agreement interpretation	Accuracy of cluster analysis (%)	Agreement of cluster analysis (kappa)	Agreement interpretation of cluster analysis	Relative accuracy gain relative to proposed method performance (%)
T1	14.2	91.17	0.61	substantial	64.62	0.22	fair	29.13
T2	6.8	85.79	0.41	moderate	73.08	0.46	moderate	14.82
T3	11.0	86.89	0.37	fair	71.15	0.21	fair	18.11
T4	11.1	88.09	0.60	moderate	70.91	0.41	moderate	19.50
T5	6.8	90.90	0.58	moderate	70.00	0.07	slight	22.99
T6	11.5	86.51	0.53	moderate	64.58	0.28	fair	25.35
T7	12.3	85.27	0.48	moderate	57.69	0.14	slight	32.34
T8	4.6	88.41	0.54	moderate	50.00	0.14	slight	43.45
T9	4.8	90.19	0.60	moderate	36.84	0.15	slight	59.15
T10	10.1	83.87	0.47	moderate	63.41	0.24	fair	24.39

Table 3 Comparison matrix of the agreements (as Cohen's kappa, alongside Fleiss' interpretation) on training effectiveness assessments as *poor*, *average*, or *good* across the case studies by two experts and our automatic workflow

	Automatic	Expert 1	Expert 2
Automatic		0.85 (almost perfect)	0.52 (moderate)
Expert 1	0.85 (almost perfect)		0.38 (fair)
Expert 2	0.52 (moderate)	0.38 (fair)	

Table 4 Average noise detection accuracy across the case studies using the Modulation Spectrogram (MS) features, along with relative accuracy loss when using other types of feature combinations, including energy-pitch (EP) and the first and second order derivatives of the features (Δ and $\Delta\Delta$)

Average accuracy (%)	Relative avg. performance loss when using one feature type (%)			
MS	MS+Δ+$\Delta\Delta$	MS+EP	EP	EP+Δ+$\Delta\Delta$
91.72	~ -1	~ -1	~ -19	~ -24

Considering the case studies as mutually independent evaluations, the average accuracy reached 87.71% with a standard deviation of 2.35%. In addition, the average agreement improved from the generally *slight* level observed with the cluster-analysis approach to a generally *moderate* level.

Overall, these results demonstrate the added value of combining super-segmental acoustic features with deep learning modelling. In particular, the workflow demonstrated strong robustness to noise, with no clear relationship observed between increasing noise levels and decreasing accuracy.

4.2 Performance of the training effectiveness assessment

The agreement matrix (Table 3) comparing the workflow training effectiveness assessment with the assessments of two experts, and the agreement between the experts themselves, revealed several noteworthy observations. The two experts exhibited complementary perspectives: Expert 1 focused more on acoustic characteristics, whereas Expert 2, with domain-specific expertise, paid greater attention to the semantic content and practical aspects of the interaction. This difference in evaluation criteria was reflected in a *fair* level of agreement between the two experts (Cohen's kappa $= 0.38$).

In contrast, the automatic system showed greater agreement with the individual experts. In particular, it achieved *almost perfect* agreement with Expert 1 (kappa $= 0.85$) and *moderate* agreement with Expert 2 (kappa $= 0.38$). These results suggest that the system formed an intermediate perspective between the two experts in assessing training effectiveness. Notably, because the workflow did not rely on speech transcription and primarily analysed acoustic and prosodic cues, its assessments tended to align more closely with the acoustically oriented perspective of Expert 1. At the same time, the system captured several domain-related situations of uncertainty and stress that were not detected by Expert 1 but were identified by Expert 2, highlighting a complementary rationale for the automated analysis relative to Expert 1.

4.3 Contribution of the Modulation Spectrogram to noise removal

The VAE model using the MS features correctly detected 91.72% of the noisy TUs (Table 4), which included various types of noise and non-speech events.

When the MS+Δ+$\Delta\Delta$ features were used, the model incurred a modest $\sim$1% relative loss in accuracy, indicating that temporal variations relative to previous frames provided only marginal information for noise removal. The MS generally was the most informative feature set. In fact, when combined with EP, the relative accuracy loss was $\sim$1% because most of the information was contained in the MS features. In contrast, when only EP or EP+Δ+$\Delta\Delta$ features were used, accuracy losses were $\sim$19% and $\sim$24% respectively.

These results highlight the crucial role of the Modulation Spectrogram features in distinguishing speech from non-speech segments. They also confirm the preliminary analysis reported in Sect. 3.3.1, which showed statistically significant differences in MS intensity across frequency bands and provided the basis for the design of the noise-removal pipeline. Overall, these findings support the use of MS features for noise removal in our workflow.

4.4 Sensitivity analysis of feature selection on ineffective verbal communication detection

The performance of the ineffective verbal communication detection model varied depending on the feature set used in the VAE, after noise removal (Table 5).

Overall, all features contained useful information for detecting ineffective communication. However, the optimal feature combination could be identified by ranking the combinations according to the number of times they achieved top-5 accuracy across the experiments (Table 6). This analysis revealed that EP+Δ features were the most consistently represented in the top-performing configurations, appearing in 90% of the cases, followed by EP+Δ+$\Delta\Delta$ features (70%). In general, combinations including energy and pitch features frequently appeared in the top 5. In contrast, energy features alone were insufficient to achieve top performance. Furthermore, using the Modulation Spectrogram in isolation from energy–pitch features introduced additional variability, thereby reducing the effectiveness of verbal communication detection.

This sensitivity analysis indicates that the EP+Δ feature combination represents the most effective configuration for the second VAE model in our workflow, when processing the TUs left by the noise-removal pipeline. At the same time, the results confirm that the Modulation Spectrogram plays a crucial role in the first stage of the workflow, where it contributes to separating speech from noise, but appears to encode information less suited to distinguishing between altered and non-altered speech patterns. Once speech segments have been isolated, syllabic-scale rhythmic and intensity cues become highly more informative for detecting ineffective verbal communication, in line with the psychoacoustic studies reported in Sect. 2.

5 Discussion and conclusions

This study has described a psychoacoustically inspired, deep-learning-based workflow for automatically detecting ineffective verbal communication during neonatal simulation training sessions. The proposed approach addresses a challenging real-world problem: analysing communication behaviour in recordings characterised by very low signal-to-noise ratios, heterogeneous background noise, overlapping speech, and highly spontaneous interactions. Instead of relying on phonetic-scale representations or speech transcription - approaches that become unreliable under such conditions - the workflow leverages syllabic-scale super-segmental acoustic features and unsupervised deep learning to analyse communication dynamics. The system processes simulation recordings by segmenting the signal into tone units, extracting features related to slow modulations of speech and super-segmental features, filtering noise segments through anomaly detection, and identifying speech segments containing acoustic-prosodic alterations associated with potentially ineffective communication. In addition, the workflow produces an interpretable Training Quality Score (TQS) that summarises the relative prevalence of acoustic-prosodic patterns associated with potentially ineffective communication during a simulation session.

The experimental evaluation has demonstrated that the proposed workflow achieved strong performance across diverse simulation scenarios. The detection of acoustic-prosodic patterns associated with potentially ineffective communication achieved an average accuracy of approximately 87.7%, with substantial-to-moderate agreement with manual annotations and substantial gain over a previously proposed cluster analysis baseline. Importantly, the system maintained stable performance across case studies with different acoustic conditions, including recordings with very low signal-to-noise ratios. This robustness highlights the advantage of combining super-segmental acoustic features with deep generative modelling for analysing speech behaviour in adverse acoustic environments. Moreover, the training effectiveness indicator produced by the workflow showed moderate-to-strong

agreement with expert assessments, suggesting that the workflow can capture acoustic-prosodic patterns considered relevant by both acoustic- and domain-oriented evaluators.

A central contribution of the proposed workflow is the use of features extracted from the Modulation Spectrogram (MS) in the noise-removal stage. The MS captures slow temporal modulations of the speech signal, which are known to reflect syllabic-scale speech modulations, and simulates temporal-sensitivity aspects of the human auditory cortex. The MS features proved highly effective for distinguishing speech from non-speech segments under severe noise conditions. The results showed that removing the MS representation significantly reduced noise-detection accuracy, confirming that modulation-based features encode crucial information for separating speech activity from environmental and machine noise. This finding aligns with psychoacoustic and computational speech-processing studies indicating that low-frequency amplitude modulations provide robust cues for speech perception and speech detection in noisy environments (Sect. 2).

When interpreting these results, it is important to differentiate between the concept of ineffective communication and the observable signals that suggest it. The proposed workflow identifies acoustic-prosodic anomalies that are statistically linked to communication challenges, rather than directly analysing the meaning or context of the speech itself. Therefore, certain aspects of ineffective communication - especially those related to clinical reasoning, specific terminology, or strategic discussions - are not captured by this approach. The system should be seen as a tool for highlighting segments that may require further review by human experts, rather than as a comprehensive evaluation of communication quality. On the other hand, the findings emphasise the significance of super-segmental acoustic features in analysing communication behaviour in complex conversational environments. Notably, variations in syllabic-scale energy and pitch were highly effective for identifying speech changes associated with ineffective communication, such as sudden spikes in vocal intensity or pitch during moments of stress, uncertainty, or conflict. Unlike features on a phonetic scale, super-segmental cues reflected behavioural aspects of speech that tended to remain consistent, even under challenging acoustic conditions. The sensitivity analysis revealed that energy and pitch features, along with their temporal changes, provided the most informative representation for detecting communication anomalies once speech segments were isolated. This observation supports the hypothesis that high-stress communication difficulties are often accompanied by measurable alterations in prosody and vocal intensity.

From a practical perspective, the proposed workflow provides a valuable tool for simulation trainers and healthcare educators, as well as for re-evaluating historical audio recordings from hospital archives. Automatically identifying tone units that likely contain ineffective communication enables trainers to focus directly on the most critical segments of the simulation. This capability would substantially reduce the time required for reviewing recordings and facilitate more targeted debriefing discussions. Furthermore, the TQS provides an interpretable summary indicator of the relative prevalence of acoustic-prosodic alterations associated with potentially ineffective communication within a session, enabling trainers to track the evolution of trainees' communication skills across multiple simulation sessions. However, it is influenced by the accuracy of the underlying detection process, the threshold used to identify ineffective communication, and the recording-adaptive nature of the workflow. As a result, TQS reflects the relative prevalence of altered speech patterns within a session rather than their contextual severity, and should be interpreted as a comparative indicator rather than an absolute measure of communication quality.

The proposed workflow should also be considered in light of recent advances in self-supervised and embedding-based speech modelling (Sect. 2). While these methods have demonstrated strong performance across a wide range of speech analysis tasks, their effectiveness typically depends on large-scale pretraining on relatively clean, homogeneous datasets. In contrast, the present work addresses a low-resource, high-noise scenario, where the direct application of these techniques may pose additional challenges. The experimental comparison in this study was limited to a traditional clustering-based baseline, motivated by the objective of evaluating the proposed workflow against the previous methodology applied to the same corpus. However, direct benchmarking against recent self-supervised and representation-learning approaches is an important direction for future investigation, particularly under low-resource and highly noisy operational conditions.

https://doi.org/10.1007/s00521-026-12319-1

Table 5 Performance measurements - accuracy, agreement as Cohen's kappa, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) - of the ineffective verbal communication detection model across the case studies at the variation of the feature set used: Modulation Spectrogram, energy, and pitch, and the respective first-order (Δ) and second-order (ΔΔ) derivatives. Bold-highlighted rows indicate the optimal feature set for each case study; red-highlighted rows indicate the finally selected set (energy + pitch +Δ)

	Accuracy (%)	Agreement	TP	TN	FP	FN	Accuracy (%)	Agreement	TP	TN	FP	FN
	T1						**T2**					
Energy	83.06	0.37	39	422	80	14	82.24	0.36	25	253	45	15
Pitch	91.53	0.55	35	473	29	18	84.31	0.37	22	263	35	18
Energy + Pitch	91.53	0.56	36	472	30	17	84.61	0.36	21	265	33	19
Energy + Pitch +Δ	*91.17*	*0.57*	*39*	*467*	*35*	*14*	*85.79*	*0.41*	*23*	*267*	*31*	*17*
Energy + Pitch +Δ+ΔΔ	**91.53**	**0.59**	**41**	**467**	**35**	**12**	87.86	0.46	23	274	24	17
Mod. Spec.	83.60	0.32	31	433	69	22	**88.46**	**0.51**	**26**	**273**	**25**	**14**
Mod. Spec. +Δ	83.78	0.33	32	433	69	21	88.46	0.48	23	276	22	17
Mod. Spec. +Δ+ΔΔ	83.78	0.32	31	434	68	22	86.39	0.41	22	270	28	18
Mod. Spec. + Energy + Pitch	84.50	0.35	32	437	65	21	88.46	0.48	23	276	22	17
Mod. Spec. + Energy + Pitch +Δ	83.96	0.32	30	436	66	23	86.68	0.43	23	270	28	17
Mod. Spec. + Energy + Pitch +Δ+ΔΔ	84.32	0.34	32	436	66	21	86.39	0.42	23	269	29	17
	T3						**T4**					
Energy	79.03	0.19	19	343	81	15	79.62	0.34	33	268	49	28
Pitch	85.58	0.33	22	370	54	12	88.62	0.60	43	292	25	18
Energy + Pitch	85.58	0.32	21	371	53	13	88.35	0.59	43	291	26	18
Energy + Pitch +Δ	*86.89*	*0.37*	*23*	*375*	*49*	*11*	*88.09*	*0.60*	*46*	*287*	*30*	*15*
Energy + Pitch +Δ+ΔΔ	86.68	0.37	23	374	50	11	**88.62**	**0.62**	**48**	**287**	**30**	**13**
Mod. Spec.	88.42	0.37	20	385	39	14	86.24	0.49	35	291	26	26
Mod. Spec. +Δ	**88.64**	**0.39**	**21**	**385**	**39**	**13**	84.65	0.43	32	288	29	29
Mod. Spec. +Δ+ΔΔ	87.77	0.37	21	381	43	13	84.12	0.41	30	288	29	31
Mod. Spec. + Energy + Pitch	86.24	0.33	21	374	50	13	87.30	0.53	37	293	24	24
Mod. Spec. + Energy + Pitch +Δ	86.89	0.37	23	375	49	11	86.24	0.50	36	315	35	31
Mod. Spec. + Energy + Pitch +Δ+ΔΔ	86.46	0.35	22	374	50	12	85.71	0.37	31	315	35	31
	T5						**T6**					
Energy	74.08	0.12	22	361	102	32	82.58	0.47	42	252	42	20
Pitch	88.97	0.51	38	422	41	16	88.48	0.59	40	275	19	22
Energy + Pitch	89.94	0.54	39	426	37	15	88.76	0.61	42	274	20	20
Energy + Pitch +Δ	*90.90*	*0.58*	*40*	*430*	*33*	*14*	*86.51*	*0.53*	*38*	*270*	*24*	*24*
Energy + Pitch +Δ+ΔΔ	**91.10**	**0.58**	**39**	**432**	**31**	**15**	**87.92**	**0.57**	**38**	**275**	**19**	**24**
Mod. Spec.	89.16	0.50	35	426	37	19	80.05	0.34	31	254	40	31
Mod. Spec. +Δ	88.58	0.48	35	423	40	19	79.21	0.31	29	253	41	33
Mod. Spec. +Δ+ΔΔ	88.97	0.50	36	424	39	18	78.65	0.30	29	251	43	33
Mod. Spec. + Energy + Pitch	89.94	0.54	39	426	37	15	81.17	0.37	32	257	37	30
Mod. Spec. + Energy + Pitch +Δ	89.16	0.52	38	423	40	16	81.17	0.37	31	258	36	31
Mod. Spec. + Energy + Pitch +Δ+ΔΔ	88.97	0.50	36	424	39	18	81.46	0.37	31	259	35	31
	T7						**T8**					

https://doi.org/10.1007/s00521-026-12319-1

Table 5 (continued)

	Accuracy (%)	Agreement	TP	TN	FP	FN	Accuracy (%)	Agreement	TP	TN	FP	FN
Energy	83.46	0.45	39	284	37	27	79.22	0.22	29	402	63	50
Pitch	84.75	0.46	36	292	29	30	87.86	0.54	51	427	38	28
Energy + Pitch	84.49	0.44	34	293	28	32	87.31	0.51	48	427	38	31
Energy + Pitch +Δ	*85.27*	*0.48*	*37*	*293*	*28*	*29*	*88.41*	*0.54*	*48*	*433*	*32*	*31*
Energy + Pitch +Δ+ΔΔ	83.72	0.41	33	291	30	33	88.41	0.53	47	434	31	32
Mod. Spec.	84.75	0.46	36	292	29	30	86.39	0.46	43	427	38	36
Mod. Spec. +Δ	84.49	0.44	34	293	28	32	85.84	0.43	41	426	39	38
Mod. Spec. +Δ+ΔΔ	84.49	0.42	32	295	26	34	86.39	0.46	43	427	38	36
Mod. Spec. + Energy + Pitch	85.01	0.46	36	293	28	30	86.39	0.46	44	426	39	35
Mod. Spec. + Energy + Pitch +Δ	83.72	0.39	30	294	27	36	87.31	0.49	45	430	35	34
Mod. Spec. + Energy + Pitch +Δ+ΔΔ	83.72	0.40	31	293	28	35	87.13	0.49	46	428	37	33
	T9						**T10**					
Energy	81.10	0.30	28	311	42	37	74.42	0.27	40	283	83	28
Pitch	86.36	0.48	37	324	29	28	**84.79**	**0.47**	**43**	**325**	**41**	**25**
Energy + Pitch	88.27	0.56	41	328	25	24	85.94	0.51	45	328	38	23
Energy + Pitch +Δ	*90.19*	*0.60*	*40*	*337*	*16*	*25*	*83.87*	*0.47*	*45*	*319*	*47*	*23*
Energy + Pitch +Δ+ΔΔ	89.95	0.59	38	338	15	27	81.79	0.37	36	319	47	32
Mod. Spec.	89.71	0.60	42	333	20	23	79.49	0.29	31	314	52	37
Mod. Spec. +Δ	**90.43**	**0.63**	**43**	**335**	**18**	**22**	79.26	0.28	30	314	52	38
Mod. Spec. +Δ+ΔΔ	89.95	0.61	43	333	20	22	79.03	0.27	30	313	53	38
Mod. Spec. + Energy + Pitch	88.99	0.57	40	332	21	25	80.18	0.32	33	315	51	35
Mod. Spec. + Energy + Pitch +Δ	89.47	0.59	41	333	20	24	79.72	0.29	31	315	51	37
Mod. Spec. + Energy + Pitch +Δ+ΔΔ	89.23	0.57	38	335	18	27	79.95	0.31	33	314	52	35

Table 6 Fractions of case studies (as percentages) for which each feature set resulted in accuracy included in the top-5 ranking per case study, when used in the ineffective verbal communication detection model. The compared feature sets include Modulation Spectrogram, energy, and pitch, along with their respective first-order (Δ) and second-order (ΔΔ) derivatives. Colours indicate the suitability of each feature set for detecting ineffective verbal communication, from green (high) to orange (low)

	Percentage of times in the top five performance ranking
Energy + Pitch + Δ	90%
Energy + Pitch + Δ + ΔΔ	70%
Pitch	60%
Energy + Pitch	60%
Mod. Spec. + Energy + Pitch	60%
Mod. Spec.	50%
Mod. Spec. + Δ	30%
Mod. Spec. + Energy + Pitch + Δ	30%
Energy	10%
Mod. Spec. + Δ + ΔΔ	10%
Mod. Spec. + Energy + Pitch + Δ + ΔΔ	10%

Despite these promising results, several limitations of the current approach should be acknowledged. First, the evaluation was conducted on a relatively small corpus comprising 10 simulation sessions from a single training centre. Although these recordings covered diverse acoustic conditions and team compositions, and a significant number of hand-annotated tone units (4,385), further validation on larger datasets from different simulation environments will be necessary to assess the generalisability of the proposed approach. Additionally, the workflow focuses exclusively on acoustic and prosodic cues and does not incorporate semantic or linguistic information from the spoken content. Furthermore, the system currently operates as an offline analysis pipeline rather than as an integrated component of the simulation infrastructure, a crucial step toward operationalisation. Finally, due to its recording-adaptive, unsupervised design, the workflow defines normality relative to the dominant acoustic patterns within each session. Therefore, the detected anomalies and the resulting TQS should be interpreted comparatively within the analysed session rather than as absolute measures of communication quality across sessions or teams.

Although the use of tone-unit-level aggregation and a threshold on the proportion of anomalous frames mitigates sensitivity to isolated fluctuations, frequent anomalous patterns may be partially normalised, potentially reducing the system's ability to consistently detect altered communication behaviours. Finally, the workflow relies on a set of empirically defined thresholds, including the percentile-based anomaly threshold, the noise proportion criterion, the tone-unit-level ineffective communication threshold, and the TQS classification intervals. These parameters were calibrated on a development set and, where available, informed by prior studies. While the sensitivity analysis presented focuses on feature selection, additional analyses on threshold variations were not included to limit the scope and length of the study. The adopted thresholds reflect a trade-off between sensitivity and robustness. In particular, the percentile-based anomaly detection provides a relative, recording-adaptive criterion, while the tone-unit aggregation thresholds mitigate the effect of isolated anomalies. Variations in these parameters primarily affect the stringency of detection: higher thresholds reduce sensitivity, while lower thresholds increase false positives. A more extensive sensitivity analysis across all parameters represents a relevant direction for future work, especially in the context of larger datasets. Future work will address all reported limitations and explore several promising research directions. The workflow will be integrated into the simulation infrastructure used by the Centro di Simulazione Neonatale NINA at the Santa Chiara Hospital. This integration will enable near-real-time analysis during training sessions, as the workflow requires only a few minutes to process a 10-minute audio on a common laptop - e.g., with an Intel i7 Processor and at least 3 GB of free Random Access Memory - and does not require a Graphics Processing Unit. Testing the system in operational conditions will provide valuable insights into its usability and effectiveness in supporting simulation-based education. Moreover, additional case studies will be collected to refine the workflow parameterisation and further

evaluate its robustness across a wider range of training scenarios. Future experiments will also analyse recordings acquired under cleaner acoustic conditions to determine whether some stages of the workflow - particularly the noise removal pipeline - can be simplified when higher-quality audio signals are available. Future versions of the system will investigate integrating the proposed acoustic analysis pipeline with speech recognition models fine-tuned to the specific grammar and vocabulary used in neonatal simulation scenarios. Such integration would enable the combination of acoustic, prosodic, and textual information, potentially improving the detection of communication ineffectiveness and providing richer feedback to trainers. Finally, although the workflow was designed to process recordings from neonatology training, its independence from the speech content makes it theoretically suitable for applications in other contexts, e.g., debriefing sessions in neonatology or training sessions in other medical fields, and for other training domains beyond medical. Future work will also explore these directions.

In conclusion, this study has demonstrated that psychoacoustics-inspired speech representations, combined with unsupervised deep learning, provide a powerful framework for analysing communication behaviour in noisy, heterogeneous simulation environments. By focusing on syllabic-scale super-segmental features rather than phonetic transcription, the proposed workflow achieved robust detection of ineffective verbal communication even under extremely challenging acoustic conditions. The resulting system offers a practical tool for supporting simulation-based training by automatically highlighting acoustic-prosodic anomalies potentially associated with communication difficulties and by providing relative indicators of altered communication dynamics within a session. More broadly, this work contributes to the growing integration of Artificial Intelligence and learning analytics into healthcare education, illustrating how advanced speech-processing techniques can enhance the evaluation and improvement of team communication in high-risk clinical contexts.

Acknowledgements The authors wish to thank all involved AOUP medical staff for their voluntary participation in these experiments.

Author contributions CRediT author statement: Gianpaolo Coro: Conceptualization, Methodology, Validation, Visualization, Writing, Software; Armando Cuttano: Resources, Validation; Serena Bardelli: Investigation, Validation, Writing.

Funding Open access funding provided by Consiglio Nazionale Delle Ricerche (CNR) within the CRUI-CARE Agreement. The paper is the result of a self-funded research initiative (ALTER EGO project, Prot. Nr. 8638 of 20260112, code 2026-CNR0A00-0008638) within the neonatology unit's activities of the Santa Chiara Hospital of Pisa, Italy.

Data availability The workflow code and related modules and models are entirely Java-based and available as open-source software on GitHub at the following links: Workflow: https://github.com/cybprojects65/IneffectiveVerbalCommunicationAnalysis, Variational Autoencoder: https://github.com/cybprojects65/VariationalAutoencoder, Tone Unit Marker: https://github.com/cybprojects65/ToneUnitMarker, Modulation Spectrogram: https://github.com/cybprojects65/ModulationSpectrogram. Data are the property of the Azienda Ospedaliero-Universitaria Pisana and are available only upon request to the authors. Audio samples corresponding to the examples in Figure 4 are available at https://github.com/cybprojects65/IneffectiveVerbalCommunicationAnalysis/tree/main/public_audio_examples.

Declarations

Ethical approval Not applicable.

Competing interests The authors declare no conflict of interest.

References

1. Lateef F (2010) Simulation-based learning: just like the real thing. J Emerg Trauma Shock 3(4):348
2. Cardoso SA, Suyambu J, Iqbal J, Jaimes DCC, Amin A, Sikto JT, Valderrama M, Aulakh SS, Ramana V, Shaukat B (2023) Exploring the role of simulation training in improving surgical skills among residents: a narrative review. Cureus. https://doi.org/10.7759/cureus.44654
3. Zhang C (2023) A literature study of medical simulations for non-technical skills training in emergency medicine: twenty years of progress, an integrated research framework, and future research avenues. Int J Environ Res Public Health 20(5):4487
4. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, Elendu ID (2024) The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 103(27):38813
5. Al-Elq AH (2010) Simulation-based medical teaching and learning. J Fam Community Med 17(1):35–40
6. Buléon C, Mattatia L, Minehart RD, Rudolph JW, Lois FJ, Guillouet E, Philippon A-L, Brissaud O, Lefevre-Scelles A, Benhamou D et al (2022) Simulation-based summative assessment in healthcare: an overview of key principles for practice. Adv Simul 7(1):42
7. Kingston L, Markey K, Doody O, Murphy L, Meskell P, Hennessy T, Moloney M (2025) Key informants' perspectives on simulation-based learning experiences: a qualitative study. BMC Nurs 24(1):1076
8. Cuttano A, Scaramuzzo RT, Gentile M, Moscuzza F, Ciantelli M, Sigali E, Boldrini A (2012) High-fidelity simulation in neonatology and the italian experience of nina. J Pediatr Neonatal Individ Med (JPNIM) 1(1):67–72
9. Thomas E, Sexton J, Lasky R, Helmreich R, Crandell D, Tyson J (2006) Teamwork and quality during neonatal care in the delivery room. J Perinatol 26(3):163–169
10. Ediger K, Rashid M, Law BHY (2022) What is teamwork? a mixed methods study on the perception of teamwork in a specialized neonatal resuscitation team. Front Pediatr 10:845671
11. Harris J, Beck S, Ayers N, Bick D, Lamb BW, Aref-Adib M, Kelly T, Green JS, Taylor C (2022) Improving teamwork in maternity services: a rapid review of interventions. Midwifery 108:103285
12. Jepkosgei J, English M, Adam MB, Nzinga J (2022) Understanding intra-and interprofessional team and teamwork processes by exploring facility-based neonatal care in kenyan hospitals. BMC Health Serv Res 22(1):636
13. Hayden EM, Wong AH, Ackerman J, Sande MK, Lei C, Kobayashi L, Cassara M, Cooper DD, Perry K, Lewandowski WE et al (2018) Human factors and simulation in emergency medicine. Acad Emerg Med 25(2):221–229
14. Moroney WF, Lilienthal MG (2008) Human Factors in Simulation and Training. CRC Press, pp 3–38
15. Stone KP, Huang L, Reid JR, Deutsch ES (2016) Systems integration, human factors, and simulation. Comprehensive Healthcare Simulation: Pediatrics. Springer, Cham, pp 67–75
16. Zajac S, Woods A, Tannenbaum S, Salas E, Holladay CL (2021) Overcoming challenges to teamwork in healthcare: a team effectiveness framework and evidence-based guidance. Front Commun 6:606445
17. Atinga RA, Gmaligan MN, Ayawine A, Yambah JK (2024) It's the patient that suffers from poor communication: analyzing communication gaps and associated consequences in handover events from nurses' experiences. SSM-Qualit Res Health 6:100482
18. Neethling J, Lubbe W, Scholtz S, Botha J (2026) Improving team performance in neonatal resuscitation: a systematic review of training programs. Mater Health Neonatol Perinatol 12(1):6
19. Brutschi R, Wang R, Kolbe M, Weiss K, Lohmeyer Q, Meboldt M (2024) Speech recognition technology for assessing team debriefing communication and interaction patterns: an algorithmic toolkit for healthcare simulation educators. Adv Simul 9(1):42
20. Li KY, Huang K, Harmer B, Ducharme C, Heasman B, Falahee E, Cooke J, Cole M, Sample A, Popov V (2026) Autoclc: towards automated assessment and feedback on closed-loop communication in team-based healthcare simulation training. ACM Trans Comput Healthc 7(1):1–31
21. Tscholl DW, Ebensperger M, Rahrisch A, Wang H, Heckel H, Thomasius M, Kaserer A, Grande B, Seelandt JC, Kolbe M (2026) Generative ai in simulation debriefings: an exploratory study using the team-first framework and qualitative feedback from simulation experts and learners. Adv Simul 11, https://doi.org/10.1186/s41077-026-00407-
22. Coro G, Bardelli S, Cuttano A, Fossati N (2022) Automatic detection of potentially ineffective verbal communication for training through simulation in neonatology. Educ Inf Technol 27(7):9181–9203
23. Echeverria V, Zhao L, Alfredo R, Milesi ME, Jin Y, Abel S, Fan JX, Yan L, Dix S, Wotherspoon R, Li X, Jaggard HA, Osborne A, Buckingham Shum S, Gasevic D, Martinez-Maldonado R (2025) Teamvision: an ai-powered learning analytics system for supporting reflection in team-based healthcare simulation. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. CHI '25. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3706598.3713395
24. Escribano S, Sánchez-Marco M, Juliá-Sanchís R, Pérez-Manzano A, Krauss A, Macur M, Kalender-Smajlović S, García-Sanjuán S, Fernández-Alcántara M, Cabañero-Martínez MJ (2025) Healthcomm simulator: virtual simulation tool for training communication skills. Comput Biol Med 197:110893

25. Herschbach L, Festl-Wietek T, Stegemann-Philipps C, Sonanini A, Herrmann B, Erschens R, Herrmann-Werner A, Holderried F (2025) Evaluation of an AI-based chatbot providing real-time feedback in communication training for mental health care professionals: proof-of-concept observational study. J Med Internet Res 27:82818
26. Wang X-J, Song L-J, Jiao X-P, Chen S-Q (2025) Integration of artificial intelligence in nursing clinical education: enhancing the mini-cex model. J Multidiscip Healthc. https://doi.org/10.2147/JMDH.S550145
27. Miner AS, Haque A, Fries JA, Fleming SL, Wilfley DE, Terence Wilson G, Milstein A, Jurafsky D, Arnow BA, Stewart Agras W et al (2020) Assessing the accuracy of automatic speech recognition for psychotherapy. NPJ Digit Med 3(1):82
28. Coro G, Bardelli S, Cuttano A, Fossati N (2022) Automatic detection of potentially ineffective verbal communication for training through simulation in neonatology. Educ Inf Technol 27(7):9181–9203. https://doi.org/10.1007/s10639-022-11000-z
29. Russell SO, Gessinger I, Krason A, Vigliocco G, Harte N (2024) What automatic speech recognition can and cannot do for conversational speech transcription. Res Methods Appl Linguist 3(3):100163
30. Parente R, Kock N, Sonsini J (2004) An analysis of the implementation and impact of speech-recognition technology in the healthcare sector. Perspect Health Inf Manag 1(1)
31. Siegert I, Bock R, Wendemuth A, Vlasenko B, Ohnemus K (2015) Overlapping speech, utterance duration and affective content in hhi and hci-an comparison. 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). IEEE, pp 83–88
32. Coro G, Massoli FV, Origlia A, Cutugno F (2021) Psycho-acoustics inspired automatic speech recognition. Comput Electr Eng 93:107238
33. Vitale VN, Cutugno F, Origlia A, Coro G (2024) Exploring emergent syllables in end-to-end automatic speech recognizers through model explainability technique. Neural Comput Appl 36(12):6875–6901
34. Srivastava M, Ferro M, Pirrelli V, Coro G (2025) Enhancing token boundary detection in disfluent speech. Intelligent Systems with Applications. https://doi.org/10.1016/j.iswa.2025.200614
35. Greenberg S, Kingsbury BE (1997) The modulation spectrogram: In pursuit of an invariant representation of speech. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol 3. IEEE, pp 1647–1650
36. Wu S-L, Kingsbury E, Morgan N, Greenberg S (1998) Incorporating information from syllable-length time scales into automatic speech recognition. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), vol 2. IEEE, pp 721–724
37. Abo El-Enen M, Saad S, Nazmy T (2025) A survey on retrieval-augmentation generation (rag) models for healthcare applications. Neural Comput Appl 37(33):28191–28267
38. Cole HS, Lindley M, Senetza A (2025) Artificial intelligence meets best practice: a scoping review of ai integration in simulation-based education. Clin Simul Nurs 109:101858
39. Gonzalez L, Nagendran A (2025) Artificial intelligence (ai)-facilitated debriefing: a pilot study. Clin Simul Nurs 105:101782. https://doi.org/10.1016/j.ecns.2025.101782
40. Rashid Z, Ahmed H, Nadeem N, Zafar SB, Yousaf MZ (2025) The paradigm of digital health: Ai applications and transformative trends. Neural Comput Appl 37(17):11039–11070
41. Mousikou P, Strycharczuk P, Rastle K (2024) Acoustic correlates of stress in speech perception. J Mem Lang 136:104509. https://doi.org/10.1016/j.jml.2024.104509
42. Schewski L, Doss MM, Beldi G, Keller S (2025) Measuring negative emotions and stress through acoustic correlates in speech: a systematic review. PLoS ONE 20(7):1–28. https://doi.org/10.1371/journal.pone.0328833
43. Protopapas A, Lieberman P (1997) Fundamental frequency of phonation and perceived emotional stress. J Acoust Soc Am 101(4):2267–2277
44. Paulmann S, Furnes D, Bøkenes AM, Cozzolino PJ (2016) How psychological stress affects emotional prosody. PLoS One 11(11):0165022
45. Park Y, Stepp CE (2019) The effects of stress type, vowel identity, baseline f0, and loudness on the relative fundamental frequency of individuals with healthy voices. J Voice 33(5):603–610
46. Kappen M, Van Der Donckt J, Vanhollebeke G, Allaert J, Degraeve V, Madhu N, Van Hoecke S, Vanderhasselt M-A (2022) Acoustic speech features in social comparison: how stress impacts the way you sound. Sci Rep 12(1):22022
47. Hajek P, Munk M (2023) Speech emotion recognition and text sentiment analysis for financial distress prediction. Neural Comput Appl 35(29):21463–21477
48. Origlia A, Galatà V, Ludusan B (2010) Automatic classification of emotions via global and local prosodic features on a multilingual emotional database. Proc of Speech Prosody, vol 21437. Citeseer, pp 2010–122
49. Rao KS, Koolagudi SG, Vempada RR (2013) Emotion recognition from speech using global and local prosodic features. Int J Speech Technol 16(2):143–160
50. Origlia A, Cutugno F, Galatà V (2014) Continuous emotion recognition with phonetic syllables. Speech Commun 57:155–169

51. Tepperman J, Narayanan S (2005) Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners. In: Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., vol. 1, p. 937 IEEE
52. Segbroeck MV, Travadi R, Vaz C, Kim J, Black MP, Potamianos A, Narayanan SS (2014) Classification of cognitive load from speech using an i-vector framework. Proc. Interspeech. pp 751–755
53. Yache VP, Moradbakhti L, Neuner I, Veselinovic T (2025) Predicting affective engagement and mental strain from prosodic speech features. Front Psych 16:1656292
54. Alamri F, Mirdad A, Saba T, Raza M, Siddique U (2026) Investigating the impact of combined spectral and prosodic features on speech emotion recognition. IEEE Access PP, 1–1 https://doi.org/10.1109/ACCESS.2025.3650618
55. Kobayashi M, Katayama M, Hayashi T, Hashiyama T, Iyanagi T, Une S, Honda M (2023) Effect of multimodal comprehensive communication skills training with video analysis by artificial intelligence for physicians on acute geriatric care: a mixed-methods study. BMJ Open 13(3):065477
56. Coro G, Bardelli S, Cuttano A, Scaramuzzo RT, Ciantelli M (2023) A self-training automatic infant-cry detector. Neural Comput Appl 35(11):8543–8559
57. Luo X, Zhou L, Adelgais K, Zhang Z (2025) Assessing the effectiveness of automatic speech recognition technology in emergency medicine settings: a comparative study of four ai-powered engines. J Healthc Inform Res 9(3):494–512
58. Moser D, Stanic N, Sariyar M (2025) Benchmarking speech-to-text robustness in noisy emergency medical dialogues: an evaluation of models under realistic acoustic conditions. JAMIA open 8(6):147
59. Shi H, Kawahara T (2024) Investigation of adapter for automatic speech recognition in noisy environment. arXiv:2402.18275 arXiv preprint
60. Shah MA, Raj B (2025) Revisiting acoustic features for robust asr. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 IEEE
61. Basire C, Lorenzini I, Cabrera L (2025) Neural and psychoacoustic insights into amplitude modulation processing and its link with speech perception in noise. J Neurophysiol 133(6):1828–1835
62. Shahamiri SR, Mandal K, Sarkar S (2025) Dysarthric speech recognition: an investigation on using depthwise separable convolutions and residual connections. Neural Comput Appl 37(12):7991–8005
63. Alsedais N, Mansour MA, Aly AM, Abdelsalam SI (2024) Artificial intelligence-driven FVM-ANN model for entropy analysis of MHD natural bioconvection in nanofluid-filled porous cavities. Front Heat Mass Transf 22(5):1277–1307. https://doi.org/10.32604/fhmt.2024.056087
64. Alsedais N, Mansour MA, Aly AM, Abdelsalam SI (2025) Artificial neural network validation of mhd natural bioconvection in a square enclosure: entropic analysis and optimization. Acta Mech Sin 41(9):724507
65. Hyder A-A, Budak H, Aly AM, Abdelsalam SI (2025) Exploring fractional bullen-type inequalities via second-derivative bounds and artificial neural networks. Eng Appl Artif Intell 162:112619
66. Schneider S, Baevski A, Collobert R, Auli M (2019) wav2vec: Unsupervised pre-training for speech recognition. arXiv:1904.05862 arXiv preprint
67. Baevski A, Zhou Y, Mohamed A, Auli M (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv Neural Inf Process Syst 33:12449–12460
68. Chung Y-A, Zhang Y, Han W, Chiu C-C, Qin J, Pang R, Wu Y (2021) W2v-bert: combining contrastive learning and masked language modeling for self-supervised speech pre-training. 2021 IEEE Automat Speech Recogni Understand Workshop (ASRU). IEEE, pp 244–250
69. Hsu W-N, Bolte B, Tsai Y-HH, Lakhotia K, Salakhutdinov R, Mohamed A (2021) Hubert: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans Audio Speech Language Processing 29:3451–3460
70. Chen S, Wang C, Chen Z, Wu Y, Liu S, Chen Z, Li J, Kanda N, Yoshioka T, Xiao X et al (2022) Wavlm: large-scale self-supervised pre-training for full stack speech processing. IEEE J Select Top Sign Process 16(6):1505–1518
71. Baevski A, Hsu W-N, Xu Q, Babu A, Gu J, Auli M (2022) Data2vec: a general framework for self-supervised learning in speech, vision and language. International Conference on Machine Learning. pp 1298–1312 (**PMLR**)
72. Purohit H, Tanabe R, Endo T, Suefusa K, Nikaido Y, Kawaguchi Y (2020) Deep autoencoding GMM-based unsupervised anomaly detection in acoustic signals and its hyper-parameter optimization https://arxiv.org/abs/2009.12042
73. Hojjati H, Armanfard N (2022) Self-supervised acoustic anomaly detection via contrastive learning. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3253–3257 IEEE
74. Liu S, Mallol-Ragolta A, Parada-Cabaleiro E, Qian K, Jing X, Kathan A, Hu B, Schuller BW (2022) Audio self-supervised learning: a survey. Patterns (N Y). https://doi.org/10.1016/j.patter.2022.100616
75. Zorin I, Makarov I (2022) Anomaly detection with self-supervised audio embeddings. Detect Classif Acoustic Scenes Events. 1–4
76. Feng C, Chen Z, Owens A (2023) Self-supervised video forensics by audio-visual anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 10491–10503
77. Zhang Z, Geiger J, Pohjalainen J, Mousa AE-D, Jin W, Schuller B (2018) Deep learning for environmentally robust speech recognition: an overview of recent developments. ACM Trans Intell Syst Technol (TIST) 9(5):1–28

78. Zhu Q-S, Zhang J, Zhang Z-Q, Wu M-H, Fang X, Dai L-R (2022) A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3174–3178. IEEE, https://doi.org/10.1109/icassp43922.2022.9747379
79. Coro G, Cutugno F, Schettino L, Tanda E, Vietti A, Vitale VN (2025) Phoné: an initiative to develop a dataset for the automatic recognition of spoken Italian. Oral Arch J 1:89–107
80. Laerdal s.r.l. (2026) Laerdal medical simulation software. https://laerdal.com/
81. Halliday MAK (2015) Intonation and Grammar in British English, vol 48. Walter de Gruyter GmbH & Co KG, Berlin
82. Cutugno F, D'Anna L (2003) Segmenting the speech chain into tone units: human behaviour vs automatic process. In: Proceedings of The XVth International Congress of Phonetic Sciences (ICPhS), pp. 1233–1236
83. Selkirk E (1986) Phonology and syntax: the relation between sound and structure mit press. Studies in Language. Int J sponsored Found "Foundations of Language" 10(1): 235–241 . https://doi.org/10.1075/sl.10.1.21coo
84. Shattuck-Hufnagel S (2019) Toward an (even) more comprehensive model of speech production planning. Lang Cogni Neurosci 34(9):1202–1213
85. Ludusan B, Ziegler S, Gravier G (2014) Is syllable stress information robust for asr in adverse conditions? In: International Conference on Speech Prosody. pp 939–943
86. Mermelstein P (1975) Automatic segmentation of speech into syllabic units. J Acoust Soc Am 58(4):880–883
87. Cutugno F, D'Anna L, Petrillo M, Zovato E (2002) Apa: Towards an automatic tool for prosodic analysis. In: Proceedings of Speech Prosody 2002, Aix-en-Provence, France, pp. 1–4 ISCA
88. Cutugno F, Coro G, Petrillo M (2005) Multigranular scale speech recognizers: technological and cognitive view. Congress of the Italian Association for Artificial Intelligence. Springer, pp 327–330
89. D'Anna L, Petrillo M (2003) Sistemi automatici per la segmentazione in unità tonali. In: ETS (ed.) Atti delle XIII Giornate di Studio del Gruppo di Fonetica Sperimentale (GFS), pp. 285–290. Associazione Italiana di Acustica, Roma
90. Coro G (2008) A step forward in multi-granular automatic speech recognition. Phd thesis, Università degli Studi di Napoli Federico II, Napoli, Italy https://doi.org/10.6092/UNINA/FEDOA/1995. http://www.fedoa.unina.it/id/eprint/1995
91. MacNeilage PF, Davis BL (2000) On the origin of internal structure of word forms. Science 288(5465):527–531
92. Fujimura O (1994) Syllable timing computation in the c/d model. Third International Conference on Spoken Language Processing. Yokohama, Japan, pp 519–522 (**ICLPS 1994**)
93. Arnal LH, Poeppel D, Giraud A-L (2016) Chapter 38 - a neurophysiological perspective on speech processing in "the neurobiology of language." In: Hickok G, Small SL (eds) Neurobiology of Language. Academic Press, San Diego, pp 463–478. https://doi.org/10.1016/B978-0-12-407794-2.00038-9
94. Cutugno F, Leone E, Ludusan B, Origlia A (2012) Investigating syllabic prominence with conditional random fields and latent-dynamic conditional random fields. Thirteenth Annual Conference of the International Speech Communication Association. pp 2402–2405
95. Coro G, Cutugno F, Caropreso F (2007) Speech recognition with factorial-hmm syllabic acoustic models. INTERSPEECH, pp 870–873 (**International Speech Communication Association**)
96. Kahn D (2015) Syllable-Based Generalizations in English Phonology. Routledge Library Editions. Routledge, London and New York
97. Kingsbury BE, Morgan N, Greenberg S (1998) Robust speech recognition using the modulation spectrogram. Speech Commun 25(1–3):117–132
98. Baby D et al (2015) Investigating modulation spectrogram features for deep neural network-based automatic speech recognition. Proceedings Interspeech 2015(2015):2479–2483
99. Gallardo-Antolín A, Montero JM (2021) On combining acoustic and modulation spectrograms in an attention lstm-based system for speech intelligibility level classification. Neurocomputing 456:49–60. https://doi.org/10.1016/j.neucom.2021.05.065
100. Mirzaei S, Jazani IK (2023) Acoustic scene classification with multi-temporal complex modulation spectrogram features and a convolutional lstm network. Mult Tools Appl 82(11):16395–16408
101. Magazine R, Agarwal A, Hedge A, Prasanna SM (2022) Fake speech detection using modulation spectrogram. International Conference on Speech and Computer. Springer, pp 451–463
102. Massaro DW (ed) (1975) Understanding Language: An Information-Processing Analysis of Speech Perception, Reading, and Psycholinguistics. Academic Press, New York
103. Kingsbury BE, Morgan N (1997) Recognizing reverberant speech with rasta-plp. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol 2. IEEE, pp 1259–1262
104. Greenwood DD (1961) Critical bandwidth and the frequency coordinates of the basilar membrane. J Acoust Soc Am 33(10):1344–1356
105. Boersma P (1993) Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proceedings of the Institute of Phonetic Sciences, vol 17. Citeseer, pp 97–110
106. Kingma DP, Welling M (2013) Auto-encoding variational bayes.arXiv:1312.6114 arXiv preprint

107. Zhang C, Peng Y (2018) Stacking vae and gan for context-aware text-to-image generation. In: 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), pp. 1–5. IEEE, Xi'an, China. IEEE
108. Lin S, Clark R, Birke R, Schönborn S, Trigoni N, Roberts S (2020) Anomaly detection for time series using vae-lstm hybrid model. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4322–4326 Ieee
109. Liu K, Shuai R, Ma L et al (2021) Cells image generation method based on vae-sgan. Proced Comput Sci 183:589–595
110. Coelho G, Matos LM, Pereira PJ, Ferreira A, Pilastri A, Cortez P (2022) Deep autoencoders for acoustic anomaly detection: experiments with working machine and in-vehicle audio. Neural Comput Appl 34(22):19485–19499
111. Powers H, Edoh K (2025) Outlier detection variational autoencoder. Neural Comput Appl 37(21):16871–16882
112. Pavirani L, Bove P, Coro G (2025) Assessing marine ecosystem risks through unsupervised methods. Eco Inform 103334
113. Bove P, Bertini A, Coro G (2026) Estimating species commonness and prevalence through unsupervised methods. Sci Rep. https://doi.org/10.1038/s41598-026-38900-1
114. Pavirani L, Bove P, Coro G (2025) Ecosystem risk assessment through stressor concurrency identification: a comparative analysis. OCEANS 2025 Brest. IEEE, pp 1–6
115. Coro G, Pavirani L, Ellenbroek A (2025) Computing ecosystem risk hotspots: a mediterranean case study. Eco Inform 85:102918
116. Coro G, Bove P, Ellenbroek A (2026) Detecting fishing areas from navigation radar detector data. Int J Digit Earth 19(1):2617004
117. WaveSurfer: Software Guide. http://www.fb10.uni-bremen.de/anglistik/ling/gk-lingintro08/software/WaveSurfer (2026)
118. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Measur 20(1):37–46
119. Fleiss JL, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Measur 33(3):613–619

Authors and Affiliations

Gianpaolo Coro[1] · Armando Cuttano[2,3] · Serena Bardelli[2]

✉ Gianpaolo Coro
gianpaolo.coro@cnr.it

[1] Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Consiglio Nazionale delle Ricerche, Via Moruzzi 1, Pisa 56124, Italy

[2] Centro di Formazione e Simulazione Neonatale NINA, Dipartimento Materno-Infantile, AOUP, Via Roma 67, Pisa 56126, Italy

[3] Unitá Operativa Neonatologia, Dipartimento Materno-Infantile, AOUP, Via Roma 67, Pisa 56126, Italy