Automatic detection of potentially ineffective verbal communication for training through simulation in neonatology

Gianpaolo Coro[1*], Serena Bardelli[2], Armando Cuttano[2] and Nicoletta Fossati[3]

[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, Azienda Ospedaliero Universitaria Pisana, via Roma 67, Pisa, 56126, Italy.
[3]St George's University Hospitals, Blackshaw Road, London, SW17 0QT, United Kingdom.

*Corresponding author(s). E-mail(s): gianpaolo.coro@cnr.it;
Contributing authors: s.bardelli@centronina.it;
a.cuttano@centronina.it; nfossati@sgul.ac.uk;

Abstract

Training through simulation in neonatology relies on sophisticated simulation devices that give realistic feedback to trainees during simulated scenarios. It aims at training highly specialised medical teams in established operational skills, timely clinical manoeuvres, and successful synergy with other professionals. For effective teaching, it is essential to tailor simulation to trainees' emotional status and communication abilities (human factors), which in turn affect their interaction with the equipment, the environment, and the rest of the team. These factors are crucial to achieving optimal timing and cooperation during a clinical intervention, to the point that they can determine the success of a complex operation such as neonatal resuscitation. Ineffective teams perform in a slow and/or poorly coordinated way and therefore jeopardise positive outcomes. Expert trainers consider human factors as crucial as technical skills. In this context, new technology can help measure learning improvement by quantitatively analysing verbal communication within a medical team. For example, Artificial Intelligence models can

work on audio recordings, and draw from extensive historical archives, to extract useful human-factor related information for the trainers. In this study, we present an automatic workflow that supports training through simulation in neonatology by automatically detecting dialogue segments of a simulation session with potentially ineffective communication between team members due to anger, stress, fear, or misunderstandings. Rather than working on audio transcriptions, the workflow analyses syllabic-scale (100-200 ms) spoken dialogue energy and intonation. It uses cluster analysis to identify potentially ineffective communication and extracts the most important related words after audio transcription. Performance is measured against a gold standard containing annotations of 79 minutes of audio recordings from neonatal simulations, in Italian, under different noise conditions (from 4.63 to 14.17 SNR). Our workflow achieves a detection accuracy of 64% and a fair agreement with the gold standard in a challenging context for a speech-processing system, where a commercial automatic speech recogniser reaches just a 9.37% sentence accuracy. The workflow also identifies viable words for trainers to conduct the debriefing session, and can be easily extended to other languages and applications in healthcare. We consider it a promising first step towards introducing new technology to support training through simulation centred on human factors.

Keywords: training through simulation, neonatal simulation, human factors, cluster analysis, speech processing, automatic speech recognition, text processing, named entity recognition

MSC Classification: 68T10 , 68T50 , 62H30 , 97U99 , 92C50

1 Introduction

Training through simulation is a methodology that uses learning environments mirroring real-life scenarios. It allows specialty trainees to put knowledge and skills into practice through hands-on activities, and has been proven effective in assessing trainees' learning. In medicine, this methodology is widely used to help trainees acquire sufficient skills to face real-life situations while keeping patients and medical staff safe [1].

Neonatology is a medical specialty where training through simulation is practiced extensively and effectively. Neonatology focusses on the care of the newborns, their development, and their illnesses. Neonatologists are doctors specialised in the care of newborns, especially those presenting peculiar health challenges, such as premature newborns and infants with underdeveloped organs. These newborns often require hospitalization in a neonatal intensive care unit, because of illness, injury, or birth defects. Neonatologists also provide general newborn evaluation and care in the hospital where they are based. They are trained to handle complex and high-risk situations during and after delivery, and are consulted during pregnancy if a problem is identified.

Neonatology training revolves around developing a team's operative skills. A newborn's well-being is achieved through the correct and timely execution of essential actions by the clinical team. As a consequence, limited or ineffective team communication would put a newborn at higher risk of a negative outcome through slow or ineffective interventions. Therefore, specialty training aims to foster effective team practices to develop established operational skills, rapidity in clinical manoeuvres, and effective synergy. In fact, reaching high team effectiveness is more crucial to good outcomes than developing individual expertise [2]. In neonatology, training through simulation can achieve positive outcomes by simulating emergency scenarios (during or after delivery) via high-fidelity simulators (e.g., manikins that simulate clinical feedback) [3, 4]. The main aim is to develop well-honed technical skills that allow trainees to work effectively within a team [5]. In this context, effective teaching and learning in simulation should take into account *human factors*, i.e. non-technical skills that can affect operators' interactions with the equipment, the environment, and the team [6]. Attention to non-technical skills is therefore crucial in all contexts where operational efficiency and effectiveness as a team are essential; this includes neonatology. Neonatal resuscitation requires excellent co-ordination between several operators, to achieve optimal timing and effectiveness in mutual interactions and, therefore, ensure successful resuscitation outcomes. In this context, simulation devices and environments are also designed with a strong focus on human factors [7, 8]. However, these training strategies require expert trainers with multi-disciplinary skills, e.g., from psychology to neonatology, who can properly recognise and manage human factors. Many factors can positively impact learning quality, e.g., (i) making the simulation environment comfortable, (ii) allowing familiarisation with the simulation room spaces, (iii) expertly managing briefing and debriefing sessions taking psychological and social factors into account, and (iv) describing all evaluation aspects in a report. Experts commonly adopt observation and memorisation for this type of training. In addition, new technology can effectively support trainers in several ways [9], e.g., through wearable microphones, high-quality audio-video recording devices, and augmented reality glasses. Finally, Artificial Intelligence (AI) can complement experts' knowledge and observation in measuring human factors. AI can further improve the long-standing practice of training through simulation by exploring the details of ineffective learning through powerful analysis models. This technological support extends trainers' abilities and helps design radically new approaches [10, 11].

The evaluation of a simulation session in neonatology is usually based on the study of audio and video recordings by expert trainers. The trainees are usually doctors who must be able to manage a simulated emergency or critical scenario with the help of assistants. Operations commonly monitored through recordings are equipment check, neonatal care, assisted ventilation, chest drain insertion, chest compressions, endotracheal intubation and drug administration. Some of these operations can also be simultaneous, requiring excellent team organisation and task execution. For example, neonatal resuscitation is

an uncommon but critical emergency where an effective team organisation is crucial to achieving success. The newborn's outcome entirely depends on the actions of different neonatologists and anaesthetists, each of them having a specific role. Cardiopulmonary resuscitation requires at least two operators constantly communicating with each other. Other interventions require three or more operators, who may not be present in the delivery room and can be summoned via emergency operational protocols ('crash calls'). By studying the audio and video recordings of a training session, trainers analyse the verbal and non-verbal components of team interactions, any silences, individual choices of team members, their communication, and involvement of all team members. Video recordings, in particular, allow inferences about the symmetric/asymmetric relationships between operators, their gestural interactions, intervention times, and the level of collaboration. Based on these aspects, expert trainers can elaborate on why an intervention (e.g., neonatal resuscitation) was successful or not. These experts can consequently study the human factors at play, and infer possible interaction, equipment, and simulation shortcomings. Observed human factors include the trainees' emotional and psychological status during the simulation in relation to the patient's clinical state, and the communication between the team members. During an emergency, the medical team follows a strict timeline that requires precise action times. A correctly-executed timeline requires successful and timely practical manoeuvre execution, equipment use, and mutual understanding. This condition should also persist in emergency scenarios. Newborn assistance can be a quiet routine moment if the newborn's clinical state is good. However, it is highly stressful if the newborn's clinical state is critical. An emergency requires quick actions for the newborn's survival; the intervention rate increases (i.e., the action timeline shortens), but the operational precision and communication quality must remain high. Emotional and psychological stress can indeed introduce mistakes and misunderstandings that would (i) slow down the timeline, (ii) overload some operators and exclude others, and (iii) potentially end in the newborn's death. For this reason, training through simulation uses emergency scenarios that test the trainees' emotional status in stressful situations to verify that the action timeline maintains its effectiveness. When the timeline is ineffective, the trainer searches for the reasons in the lack of individual technical skills (i.e., manoeuvre execution and equipment use) and in poor mutual understanding. While individual technical skills can be directly improved through practice, mutual understanding mainly depends on communication effectiveness and can consequently affect collaboration in technical equipment use. Specifically, the trainees' dialogue intonation and communication modalities directly affect communication effectiveness during simulation. Thus, optimising these aspects, through good communication practices in quiet-to-stressful situations, can improve the operational timeline effectiveness. To this aim, training through simulation in neonatology addresses the achievement of effective and efficient team dynamics with excellent synchronicity and little room

for misunderstandings. Every communication episode addresses the appropriate team member, who in turn asks back for confirmation of their correct understanding. The communication style is assertive but not aggressive, and emotional content and verbal intonation are controlled.

In this paper, an automatic workflow is proposed to support training through simulation in neonatology. The workflow automatically highlights segments of *potentially ineffective* communication in the recorded dialogues of a simulation session, through audio signal processing and unsupervised machine-learning. It then transcribes the detected dialogue segments and extracts the conversational keywords with their related semantic weights. The proposed workflow can rapidly provide trainers with a summary of the dialogues that should be analysed first, i.e., largely ineffective verbal communication between the team members due to anger, stress, or misunderstandings. The workflow significantly reduces the time for trainers to get relevant audio segments after long recording sessions. It also identifies verbal communication elements (verbs, tenses, nouns, adjectives) that should be discussed and improved upon. This way, the workflow complements the work of the trainers in the analysis of a simulation session and can in turn generate advice for the trainers, after a debriefing session, about their verbal communication with the trainees.

Unlike other approaches [12–14], our workflow detects potentially ineffective verbal communication by processing audio at a syllabic temporal scale (100-200 ms), without transcribing the speech, and is robust to a very high and varied noise level. It overcomes a common limitation of other approaches [15, 16] that use high-quality recording devices as it can also process low-quality audio recordings, e.g., historical recording archives. The workflow was tested on ten real case studies containing spoken dialogues in Italian. Common approaches based on transcribed audio processing would not capture enough detail due to the high and varied noise present. Environmental noise was principally generated by clinical tools operating during the simulation. Performance was measured as the accuracy, precision, and recall at detecting potentially ineffective communication against a gold standard prepared by two researchers with complementary expertise on training through simulation and spoken dialogue analysis respectively. The practical value of the extracted keywords in pointing out ineffective attitude and positive aspects from the simulations was also demonstrated.

2 Methods

2.1 Workflow for potentially ineffective communication detection

Our workflow processes a one-dimensional audio signal, i.e., a sequence of sound-amplitude samples (measured in dB) recorded by a microphone at a 44,100 Hz sampling frequency. The input audio relates to a recorded neonatology emergency simulation session. As output, the workflow labels segments of the input audio containing potentially ineffective communication due to

alterations of speech energy and prosody (intonation). The detected segments are eventually transcribed using a state-of-the-art automatic speech recogniser (ASR). The transcribed text is further processed to extract the words containing enough semantic information to understand the dialogue contents. In the following sub-sections, the workflow is described through the steps reported in Figure 1. The workflow was designed and developed in JAVA and is open source (Supplemental information).

2.1.1 Digital signal processing

The input of our workflow is an audio signal recorded by either a wearable or room microphone. The workflow was designed to also process historical recordings with a low signal-to-noise ratio (SNR), i.e., a high noise level. In our study cases, barely intelligible recordings were present (with a ∼4.6 SNR) along with good-quality recordings (with a ∼14.2 SNR). Modern ASRs cannot transcribe speech with such a high noise level (Section 3.1), mainly because the phonetic structure is compromised [17]. In these cases, using acoustic syllables as a speech analysis unit is convenient because the syllabic structure is robust to a high noise level, and syllables are key to human speech recognition robustness [18]. An acoustic syllable is a 100–250 ms segment of speech signal built around a high energy peak (nucleus), often preceded by an increasing energy slope (onset) and followed by a decreasing energy tail (coda). A few acoustic features can characterise a syllable directly without referring to its phonetic structure. Among these, energy and pitch are commonly used [19].

Syllabic energy is here intended as the squared sum of the audio samples of a syllable divided by the number of samples, i.e., formally it is the *power* of a syllabic signal segment. The time series of syllabic energy correlates with the intensity and rate of the speech. Our workflow normalises the energy time series with respect to the maximum energy to produce commensurable data from different recordings. Syllabic pitch is an acoustic correlate of tone and intonation, and is the frequency of a tone associable to a syllable. The time series of syllabic pitch represents the musicality and intonation of the speech. Our workflow uses syllabic-scale sliding windows on the signal to extract energy and pitch, with 50% superposition between the windows. It calculates pitch for each signal window using the Boersma's sound-to-pitch algorithm [20] - which is based on the windowed-signal autocorrelation - with a lower cut-off frequency at 60 Hz. The extracted time series of energy and pitch contain one value for each window. They constitute the basis of all further analyses of the workflow. The window length used for energy is 100 ms, which allows to catch short pauses and is functional to signal segmentation. The pitch window length is 200 ms, which estimates long tones only but also increases the reliability of the detected pitch. In fact, this window size likely selects the tones of prominent syllables, which preserve their structure even when a high and varied noise level is present [21].

2.1.2 Dialogue segmentation

Ineffective verbal communication is commonly studied from dialogue units containing finite and meaningful interactions. An acoustic proxy of these dialogue units is the *tone unit*. A tone unit is a portion of speech uttered within a coherent intonation contour. Tone units are also valuable for increasing ASR performance as they mostly contain only speech and complete sentences. They also serve the final text analysis phase of our workflow. Our workflow detects potentially ineffective verbal communication among the tone units extracted from the audio signal by first segmenting the audio into tone units, then passing these units to the subsequent classification phase. Our workflow embeds a fast algorithm based on the signal energy for tone unit detection [19]:

Algorithm 1 Tone unit detection algorithm

Step 1: calculate the derivative of the energy time series;
Step 2: for each energy time series sample:
Detection condition: if the derivative to the sample is negative and energy is below a tone unit threshold -> mark the energy window beginning as a tone unit end.

The tone unit threshold is an adaptive and iterative energy threshold that starts from a minimal value (0.001 dB) and doubles until at least three tone units are found. This algorithm identifies a tone unit as the end of a sequence of high-energy speech and has demonstrated robustness in speech segmentation under varied noise conditions [22].

The initial audio recording is divided into shorter recordings corresponding to tone units. Units with lengths under 3 s are not processed further as they contain either noise only or too short sentences (i.e., one or two words before a long pause) that are not useful to our workflow. The overall output of this workflow step is a sequence of dialogue segments likely corresponding to tone units, with two vectors of energy and pitch time series associated with each segment.

2.1.3 Cluster analysis

As a further processing step, our workflow labels each dialogue segment extracted through the previous step as either *potentially ineffective* or *viable* verbal communication. *Potentially ineffective* communication refers to segments that an expert trainer or a social analyst can use to study critical situations of misunderstanding, stress, fear, and agitation. *Viable* communication indicates either informative dialogues or effective communication. Our automatic communication labelling is fully based on energy and pitch. Annotated collections of ineffective dialogue examples (corpora) were not available for model training. Moreover, there was a large variability of noise and speech

across the analysed data that would have required collecting a huge amount of data to build a supervised model (e.g., an Artificial Neural Network, a Support Vector Machine, etc.). This scenario imposed using an unsupervised modelling approach for automatic communication labelling. Unsupervised modelling usually achieves lower performance than supervised modelling; however, as it does not require annotated corpora, is cost- and development-time saving. We therefore adopted a cautious approach to dialogue segmentation: possible mislabelling of communication segments had to be tolerated and yet kept to a minimum, to make automatic labelling useful for the experts. We also chose to label segments as *potentially ineffective*, rather than *ineffective*, leaving to the experts the final decision to use or discard them.

Our workflow uses cluster analysis based on the K-means algorithm [23] to distinguish between *potentially ineffective* and *viable* communication. K-means is a widely used unsupervised model that divides a vector space into a fixed number of clusters without prior knowledge of the vector distribution. The K-means iterative processing that optimally assigns the vectors to k clusters. A vector is assigned to the nearest cluster using Euclidean distance as a measure of proximity. As for our workflow, the vector space is composed of multidimensional vectors of energy and pitch values from the extracted tone units. Since K-means requires sample vectors to have the same lengths, our process uses the vector of the central 3 seconds of each extracted dialogue segment. As a result, the extracted vectors consist of 90 values (60 energy and 30 pitch values), i.e., the vector space dimension is 90. Each vector is assigned to one among two clusters ($k = 2$) after 1000 optimisation iterations. Experiments using more than two clusters were also conducted, but yielded more false negatives (i.e., mislabelled ineffective communication) and were less valuable for the expert analysis.

After the K-means assignment, our process calculates the average energy and pitch of each cluster. Vectors associated with potentially ineffective communication are identified as those falling in the cluster with the highest average energy and pitch, i.e., those with the highest $c = avg(energy) \cdot avg(pitch)$ score. The cluster with the highest c score indicates those dialogue segments with the highest energy (potentially associable with anger, stress, fear, and misunderstandings) and pitch (potentially associable with a high conversational tone and agitation).

The output of this workflow step is therefore the labelling of the dialogue segments as containing either *potentially ineffective* or *viable* verbal communication. This information undergoes signal annotation and audio transcription.

2.1.4 Speech annotation

After audio segment labelling, our workflow annotates the original audio file by producing a plain text file associated with the audio file. The LAB file format [24] is used for the annotation, which contains lines with start and

end seconds and an associated comment, e.g., "12.5 15.9 'Potentially ineffective communication' ". It is an easy to parse format, suitable for integrating our workflow with commonly used speech analysis tools (e.g., WaveSurfer and Praat). Visualising speech with annotations allows experts to rapidly browse through a long audio recording, quickly identifying potentially ineffective communication during the most critical moments of the simulation before starting the debriefing phase.

2.1.5 Speech recognition and text analysis

As a further optional step, our workflow uses a large-vocabulary automatic speech recogniser for Italian to transcribe dialogues containing potentially ineffective communication. This step is optional due to the high sensitivity of modern state-of-the-art ASRs to noise [17, 25] and therefore is more suited to higher SNR cases. It can be enabled through a Boolean flag at the start of the workflow. As the default ASR, our workflow uses the Google Speech-to-Text cloud service [26]. This ASR internally uses deep-learning phoneme models trained with thousands of hours of annotated speech. It has top-level performance and a high response efficiency, and is used by most of the Google technology. Moreover, Google constantly improves its performance with new data. The Google ASR can recognise ten times the number of words of an entire language dictionary and supports 120 languages. However, the Google ASR allows transcribing a maximum of 60 minutes of audio per month with a free account and would require high costs for more extensive analyses. As an alternative, our workflow embeds a free-to-use large-vocabulary speech recogniser for Italian [17], based on state-of-the-art technology (i.e., the KALDI ASR toolkit [27]), trained with 20 hours of open-access annotated corpora (VoxForge [28]). The transcription performance of this ASR is about 22% lower than the Google ASR one on moderately noisy audio [17], but the gap decreases on cleaner audio.

In our workflow, the ASR transcription is functional to a *keyword* extraction process. This process extracts the words that refer to the main topics in the text, i.e., the words with the highest amount of semantic information about the transcribed dialogue. In particular, our workflow uses the free-to-use *Keywords* Named Entity Recogniser Web service of the NLPHub platform [29]. The input of this service is a text file and the specification of the text language (Italian in our case, although 23 languages are supported); the output is a "cloud" of keywords (*word cloud*) comprising verbs and nouns with associated weights. These weights are proportional to the word occurrence-frequency in the text. Among the frequent words, the algorithm retains only those with a frequency close to the geometric mean (within 1.5 standard deviations). The underlying hypothesis is that the distribution of the occurrence frequencies of meaningful words across a document is log-normal, i.e., the geometric mean is the distributional mean. This fast process was used in awarded digital assistants and has demonstrated effectiveness in multiple domains [29].

It also attenuates the bias of ASR transcription errors as it reports repeated words and eliminates nonfunctional words (stop words).

After running the ASR and *Keywords* on the identified potentially ineffective dialogue segments, a *word cloud* is reported to the trainers to highlight the most important words uttered and repeated in these dialogues. This information is essential to identify patterns of errors and sub-optimal practices (Section 3.3). Repeated verb tenses and persons, questions, exclamations, references to time and missing equipment constitute information that experts can use to rapidly infer the main issues in the training session.

2.1.6 Workflow summary

Referring to Figure 1 and the explanations given in the previous sections, the complete workflow can summarised through the following steps:

1. *Digital signal processing*: Energy and pitch are calculated for syllabic-scale sliding windows (100-200 ms) over the signal (Section 2.1.1);
2. *Dialogue segmentation*: Tone units are detected and their audio segments are extracted. Only segments with length over 3 seconds are retained (Section 2.1.2);
3. *Cluster analysis*: Tone unit segments are labelled as either *potentially ineffective* or *viable* verbal communication, through cluster analysis of energy and pitch (Section 2.1.3);
4. *Speech annotation*: A LAB file is produced as the annotation of the original audio file. It contains the specification of the intervals where communication is *potentially ineffective* (Section 2.1.4);
5. *Speech recognition and text analysis*: The audio of the tone units containing *potentially ineffective* communication is transcribed through an automatic speech recogniser. A keyword extraction process is then used to produce a *word cloud* that highlights the most important nouns and verbs in the dialogues (Section 2.1.5).

The workflow produces audio annotations that can be inspected through widely used speech analysis tools (e.g., WaveSurfer and Praat); these allow to rapidly browse through *potentially ineffective* communication. On the other hand, the output *word cloud* works as a summary for trainers to identify the main reasons for major communication issues. For example, in Figure 1, the greater weight of the first person plural "facciamo" (let us do) than the one of the first person singular "faccio" (I do) suggests that the team leader is uncertain about which actions to take. The much greater weight of these two verbs compared with all the other words indicates that they are very frequent in the dialogues; thus, it can be inferred that the team leader is discussing a strategy rather than giving precise instructions. This situation is risky, because timing and organisation are crucial for newborn survival.

2.2 Study cases

Ten study cases were selected from historical recordings from the "Centro di Simulazione e Formazione Neonatale" (Centro NINA) to evaluate the performance of our workflow on the detection of ineffective verbal communication. These recordings belonged to the simulation activities practised within the "Dipartimento Materno Infantile" of the Azienda Ospedaliero-Universitaria Pisana for paediatrics speciality training. As a preliminary session, trainees supported an expert trainer who managed a simulated emergency. All involved experts had Paediatric Basic Life Support Defibrillation (PBLSD) certification. After one month, trainees were asked to manage the simulated scenario without the trainer's support. At the end of the session, the trainer conducted a debriefing session to analyse and evaluate the simulation. After two-three months (without a pre-scheduled date), trainees were asked to repeat the simulation with a trainer being among the assistants. Audio recordings from these last sessions were used as study cases for our workflow. A total number of 10 sessions were selected (Table 1), for an overall duration of 79 min 5s, under different noise conditions (from 4.63 to 14.17 SNR) and team and gender composition (two or three team members, male and female trainees).

The simulations were conducted in a dedicated room of Centro NINA inside the Santa Chiara Hospital of Pisa. The room was equipped with a real-life neonatal island, a delivery room and clinical tools to improve the attendees' emotional involvement and give the procedures realistic feel. The simulated scenarios revolved around neonatal resuscitation. A neonatal simulation manikin (SimNewB [30]) was used to reproduce the same critical clinical scenarios for all trainees. SimNewB was managed through the LLEAP simulation software. This software simulated clinical states and was sensorised to record the effectiveness of the clinical interventions and give feedback to the medical team. The feedback and interaction were not supervised by an expert behind the scenes, i.e., they entirely depended on the trainees' interventions and the simulator automatic feedback. The SimView software [31] was used to record the sessions from four wireless cameras and one environmental microphone (Behringer B2 PRO). The microphone was placed 1.5 m above the simulation manikin, in the middle of the room, without noise filtering enabled.

Our workflow processed the study cases to tag potentially ineffective audio segments and extract keywords related to these segments. A gold standard of annotated audio segment as *potentially ineffective* or *viable* was prepared by two researchers with different expertise in the field. The first researcher had long experience on dialogues and training through simulation in neonatology and could identify, annotate, label, and comment potentially ineffective communication. The second researcher had long experience on spoken dialogue communication, analysis, and computational linguistics, and revised the annotations of the former expert to further comment the labelled audio segments. The researchers did not know the automatic tagging result. Their comments identified and explained the critical issues in the communication. Based on this gold standard, the audio segments that our workflow correctly labelled

as *potentially ineffective* communication were considered *true positives* (TP), and *false positives* otherwise (FP). Segments correctly labelled as *viable* communication were considered *true negatives* (TN), and *false negatives* (FN) otherwise. Performance on audio tagging was measured through the following standard metrics:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$Precision = \frac{(TP)}{(TP + FP)}$$

$$Recall = \frac{(TP)}{(TP + FN)}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{(Precision + Recall)}$$

Overall, accuracy is the total portion of correctly labelled audio segments containing either *potentially ineffective* or *viable* communication; precision is the ratio of correctly detected *potentially ineffective* communication segments to all ineffective-communication segments detected by the workflow; recall measures the sensitivity of the model, i.e., the ratio of correctly labelled *potentially ineffective* communication segments to all ineffective-communication segments reported in the gold standard. F1 is the harmonic mean of precision and recall and indicates how balanced the workflow is between these two indicators. Finally, Cohen's kappa [32] was used to measure agreement between the classifications produced by our workflow and those reported in the gold standard, with respect to chance. Fleiss' labels [33] were used to interpret the values.

As a further evaluation, the researchers' comments associated with potentially ineffective communication were checked to report words included in the extracted *word cloud*. These comments often cited sentences uttered during the dialogue to highlight ineffective communication and suggest corrections. Examples of comments including citations are: "the statement '*io non sono capace*' (I am not able to) should be discussed with the trainees", "the learner does not remember the procedure and says '*non mi ricordo*' (I do not remember)", "valuable information is contained in this segment for the debriefing session, with the interesting statement '*stai facendo poca pressione*' (you are applying little pressure)", "there was confusion between the terms 'glucosata' (glucose solution) and 'fisiologica' (saline solution)", "too much emotional stress in '*e io come vado?*' (how am I doing?)". The total percentage of cited words (both verbs and nouns) of the gold standard included in the *word cloud* extracted from all study cases was used as an effectiveness measure of the keyword extraction process. Although this was not the optimal approach to evaluate the *word cloud* effectiveness, it was the best way to produce an approximate numerical and objective assessment with the data at hand.

3 Results

3.1 Limits of automatic speech recognition on our study cases

The averagely low SNR of the study case recordings prevents achieving a high transcription performance through state-of-the-art ASRs. The Google Speech-to-Text service correctly recognised 46.87% of the words transcribed from all the study case recordings (Table 2). The performance had high variability (48%) across the study cases and depended on the noise level. The minimum 23.31% accuracy was gained on the 4.63 SNR audio, whereas the maximum 71.33% accuracy was gained on the 14.17 SNR audio. However, the ASR completely missed 88.4% of the uttered words, as they were not present in the transcription. Thus, the recognition accuracy on the entire sentences (9.37%) was much lower than the one on words and ranged from 4.66% to 14.27% over the study cases.

This low performance made it difficult to use the Google ASR to directly produce the text and keywords of the simulation dialogues and infer communication effectiveness. The chance to catch a critical keyword from our cases can be approximated as the chance that the keyword was not omitted $(1 - P_{omission})$ multiplied by the chance that the keyword was correctly transcribed ($P_{recognized}$), i.e., $(1 - P_{omission}) \cdot P_{recognized} = (1 - 0.884) \cdot 0.4687 = 0.054 = 5.4\%$. There was therefore only a 5.4% chance to catch a keyword from our simulation dialogues using a state-of-the-art commercial ASR. This scenario justifies our choice to use a completely unsupervised approach based on syllabic audio features to detect potentially ineffective communication, rather than process the transcribed text. Our workflow uses the ASR only on the already detected potentially ineffective communication segments. This operation eliminates segments containing noise only and focuses the ASR on shorter segments, which improves the transcription performance [17]. Furthermore, the final keyword extraction process lowers the dependency on transcription errors by reporting frequent words only.

3.2 Performance on potentially ineffective communication detection

The performance of potentially ineffective communication detection was evaluated on each study case, and also on all recordings considered as one study case. The use of syllabic-scale energy and pitch instead of phonetic features also improved processing speed. The workflow required ∼10 seconds to process a 10-minute recording on an Intel i7-7700HQ CPU machine with 16GB RAM. Referring to the measurements reported in Table 3 and Figure 2, the following considerations can be extracted: Accuracy ranged from 37% (on T9) to 73% (on T2) and reached an overall 64% on all recordings. The agreement between the unsupervised model and the gold standard ranged from 0.066 (on T5) to

0.455 (on T2). It was *fair* over all recordings, according to Fleiss' interpretation. Of note, our workflow reached optimal performance on T2, although the SNR was low (6.76). In other cases (e.g., T5), accuracy (70%) was higher than the average, but the agreement was low because communication was infrequent (few cases to evaluate), i.e., agreement by chance was possible. F1 ranged from 0.25 (on T5) to 0.7 (on T2) and was overall 0.47. These measurements should be considered within the limitations of working on very noisy audio and low dialogue intelligibility. In this context, a 64% [37; 73%] accuracy, a 0.47 [0.25; 0.7] F1, and a *fair* agreement with the gold standard can be considered a satisfactory achievement. In fact, using an alternative approach based on audio transcription, with a 9.37% sentence accuracy, would have yielded much lower quality results.

The percentage of unlabelled potentially ineffective communication, i.e., the false negatives, was 11.7% overall (50 segments over 426) (Table 4) and indicates that the workflow misidentified only few potentially ineffective communication segments. The actual percentage of audio segments containing potentially ineffective communication was 27.5% (117 segments over 426), whereas 40% (170 over 426) was automatically labelled as potentially ineffective. Although our workflow was cautious at indicating ineffective communication, it helped save experts' time by reducing listening time by 60% compared with exhaustive searching among the dialogues contained in the complete audio recording.

The main difference from transcription-based approaches is that a higher SNR does not necessarily correspond to lower performance. As reported by other studies [17, 21], this characteristic is due to the robustness of energy and pitch features to noise. In particular, sample cross-correlation between performance measurements and SNR was maximal for accuracy (0.52) and minimal for precision (0.28) (Table 5). Referring to Figure 3 and using F1 as a measure of the overall performance quality (since it balances precision and recall), higher performance ($F1 \geq 0.4$) was achieved both on clean ($SNR \geq 11$) and noisy audio ($SNR = 6.76$) study cases. However, low performance ($F1 < 0.4$) was also observed both on clean ($SNR = 12.3$) and noisy audio ($SNR = 12.3$) study cases. The same considerations are valid for Cohen's kappa, whose modulations replicate those of F1. On the other hand, accuracy showed a higher dependency on noise when $SNR < 6.76$, due to an increased number of missed labels (false negatives) when noise completely masked dialogues. With $SNR \geq 6.76$, accuracy always remained above 58%. Overall, these considerations indicate that our workflow is poorly sensitive to SNR. Rather, it is more sensitive to other factors like the frequency of the interactions, intonation, and audio volume.

3.3 Text analysis

Keyword extraction from all audio transcriptions (i.e., across all study cases) that contained potentially ineffective communication revealed the presence of many valuable verbs for expert analysis. Repeated words with a greater

weight in the *word cloud* in the first person plural - e.g., "*abbiamo*" (we have), "*vediamo*" (let us see), "*andiamo*" (let us go), "*dobbiamo*" (we have to) - and expressions of uncertainty - e.g., "*vabbè*" (oh well), "*dimmi*" (tell me), "*bisogna*" (we ought to) - were generally present (Figure 4 and Table 6). These repetitions suggest the absence of a team leader in most study cases, with uncertainty expressions underlining low self-confidence and unclear instructions. The doctor leading operations during the simulation should give instructions to his/her assistants rather than discuss strategies during the emergency. This situation may indicate low expertise in the specific task or the lack of a structured approach. Of note, words referring to time - e.g., "*presto*" (hurry up) – were not in the cloud, suggesting that limited time was never perceived as an issue, which is a general indicator of good procedure knowledge. Likewise, words referring to equipment, and markers of issues in the execution of a procedure - e.g., "*peccato*" (what a pity) - were not in the cloud, indicating that the simulation environment and the equipment were sufficient for the experiments. This information was useful for the trainers to understand the critical issues and trainees' attitudes preventing effective communication. It can also help create strategies to avoid common misunderstandings and improve learning and team organisation.

The viability of the detected keywords, measured as the percentage of gold-standard words contained in the *word cloud*, was 59%. Most of the gold-standard words were therefore included in our workflow output (Table 4). The excluded words were mostly auxiliary verbs that could not be eliminated automatically before the evaluation. Identifying and deleting these words would indeed have required a higher word-repetition rate in the gold standard comments.

4 Discussion

In this paper, a new technique to detect potentially ineffective verbal communication between team members during a neonatal simulation session has been described. Our methodology is simple, in that it requires only the audio recorded during the simulation. Furthermore, it overcomes difficulties due to high and varied audio noise level through syllabic-scale speech processing. Our method is less dependent on noise level than speech recognition-based solutions. It uses text analysis based on automatic speech recognition only on potentially ineffective communication, to extract keywords that social analysts can use. Our system achieved satisfactory performance to support expert analysis on 10 study cases. Moreover, the extracted keywords included most of the words contained in the gold standard. The produced *word cloud* allowed inferring general verbal communication biases (e.g., absence of a team leader, inexperience) and positive aspects (e.g., sufficient equipment and time for the simulation).

Our approach represents a first further step towards the proposal of new technology to support experts in studying the communication between medical team members. It can be extended to other languages than Italian as it uses speech transcription and text processing models covering many languages. The syllabic-scale speech processing itself is independent of the language. Ideally, our workflow should be conceived as part of a more general workflow that digests and summarises information from a simulation session for expert trainers to use. Based on the information produced by our system, the experts could rapidly identify issues by looking at the keyword cloud and listening to potentially ineffective communication. Our tool is also suited to analysing long debriefing sessions and producing valuable information about the verbal communication between trainers and trainees. Our workflow robustness to noise allows processing of archived sessions that used lower-quality technology, which can help revaluate historical corpora.

The workflow performance was evaluated on extreme cases, which are less frequent in modern scenarios where high-quality wearable microphones allow labelling and identifying speakers. Modern devices can indeed reduce noise and improve the workflow performance by allowing the development of new specific models. For example, the degree of participation of a team member in the dialogue could be assessed through speech activity measurement. This measure would also allow experts to evaluate the consistency of the team structure and the member's role. Detecting and analysing effective communication through specific models would help contextualise the simulation emotional context and the empathy between the team members. These features are not included in our current workflow, as they would have required more reliable textual transcription and a low audio noise level, for example to run *sentiment* and *speech-emotion* analyses. Future applications and developments will embed these techniques to produce an overall quality score for simulation and debriefing sessions. This score will consider emotional and communication aspects and will also analyse each trainee separately.

Overall, our workflow and its future versions support two essential dimensions of training through simulation: a *horizontal* dimension, where communication practices are abstracted from multiple simulations and experts design new communication protocols for all medical teams; and a *vertical* dimension, where one team or person develops communication skills thanks to personalised practices. We believe these to be important lines of future research into new technology to support training through simulation centred on human factors.

Supplementary information. The source code of our workflow is available on the D4Science e-Infrastructure at
https://svn.research-infrastructures.eu/public/d4science/gcube/trunk/data-analysis/NINASpeechCommunicationAnalyser/

Definitions

Definitions of key concepts used in the paper are reported in the following:

- *Neonatology*: the branch of medicine that focusses on the treatment and care of newborns;
- *Training through simulation*: a training methodology that relies on artificial learning environments (manikins, sensors, virtual and augmented reality, etc.) that mirror real-life scenarios;
- *Human factors*: factors related to the environmental, organisational, working context, and individual characteristics that influence behaviour, performance, health and safety at work;
- *Paediatric Basic Life Support Defibrillation certification*: a certification programme in how to manage cardiac arrest of a paediatric subject using an automatic external defibrillator.
- *Speech signal*: a sequence of air pressure amplitudes over time captured by a microphone, which transforms air fluctuations into an electrical signal;
- *Sampling frequency*: the average number of samples per seconds stored by a sound recording device, measured as the number of samples recorded per second;
- *Speech signal energy*: the squared sum of the audio samples of an audio segment;
- *Speech signal power*: the squared sum of the audio samples of an audio segment divided by the number of samples;
- *Speech prosody*: the rhythm, intonation and stress of speech;
- *Speech pitch*: the fundamental frequency of the speech signal, corresponding to the vibration frequency of the vocal cords during the production of sonorant consonants or vowels;
- *Signal-to-noise ratio*: a measure comparing the intensity of a signal to the intensity of background noise, i.e., the ratio of signal power (energy per unit of time) to background noise power;
- *Acoustic syllable*: a 100–250 ms segment of speech signal built around a high energy peak (nucleus), often preceded by an increasing energy slope (onset) and followed by a decreasing energy tail (coda);
- *Tone unit*: a portion of speech uttered within a coherent intonation contour;
- *Potentially ineffective communication in training sessions*: dialogues of interest for an expert trainer or a social analyst to study critical situations of misunderstanding, stress, fear, and agitation;
- *Viable communication in training sessions*: informative or complex dialogues with effective communication;
- *Cluster analysis*: a set of computational methods to group a set of data so that data in the same group (cluster) are close with respect to a similarity measurement (e.g., a distance, the angle between two vectors, etc.);
- *Automatic speech recogniser*: an automatic system that can transform an audio signal into a sequence of words from a language dictionary according to grammar rules;
- *Syllabic temporal scale*: time scale of speech phenomena with an average duration similar to that of an acoustic syllable (usually 100-200 ms);

- *Named entity recogniser*: an information extraction system that locates and classifies the words mentioned in an unstructured text associated to predefined categories (e.g., persons, organisations, locations, etc.).
- *Word cloud*: a set of verbs and nouns with associated weights proportional to their occurrence frequencies in the text;
- *NLPHub*: a free-to-use cloud computing service that orchestrates several named entity recognisers to extract entities from an input text and produces a *word cloud*.

Declarations

- Funding: Not applicable.
- Conflict of interest/Competing interests: The authors declare no conflict of interest.
- Ethics approval: Not applicable.
- Consent to participate: Not applicable.
- Consent for publication: Not applicable.
- Availability of data and materials: Not applicable.
- Code availability: The source of the described workflow is linked in the Supplementary information.
- Authors' contributions: blinded for review.

References

[1] Lateef, F.: Simulation-based learning: Just like the real thing. Journal of Emergencies, Trauma and Shock **3**(4), 348 (2010)

[2] Cuttano, A., Scaramuzzo, R.T., Gentile, M., Moscuzza, F., Ciantelli, M., Sigali, E., Boldrini, A.: High-fidelity simulation in neonatology and the italian experience of nina. Journal of Pediatric and Neonatal Individualized Medicine (JPNIM) **1**(1), 67–72 (2012)

[3] Kurup, V., Matei, V., Ray, J.: Role of in-situ simulation for training in healthcare: opportunities and challenges. Current opinion in anaesthesiology **30**(6), 755–760 (2017)

[4] Satin, A.J.: Simulation in obstetrics. Obstetrics & Gynecology **132**(1), 199–209 (2018)

[5] Halamek, L.P.: The simulated delivery-room environment as the future modality for acquiring and maintaining skills in fetal and neonatal resuscitation. In: Seminars in Fetal and Neonatal Medicine, vol. 13, pp. 448–453 (2008). Elsevier

[6] Hayden, E.M., Wong, A.H., Ackerman, J., Sande, M.K., Lei, C., Kobayashi, L., Cassara, M., Cooper, D.D., Perry, K., Lewandowski, W.E.,

et al.: Human factors and simulation in emergency medicine. Academic Emergency Medicine **25**(2), 221–229 (2018)

[7] Moroney, W.F., Lilienthal, M.G.: Human factors in simulation and training. Human Factors in Simulation and Training. CRC Press, 3–38 (2008)

[8] Stone, K.P., Huang, L., Reid, J.R., Deutsch, E.S.: Systems integration, human factors, and simulation. In: Comprehensive Healthcare Simulation: Pediatrics, pp. 67–75. Springer, Cham (2016)

[9] McGrath, J.L., Taekman, J.M., Dev, P., Danforth, D.R., Mohan, D., Kman, N., Crichlow, A., Bond, W.F., Riker, S., Lemheney, A., *et al.*: Using virtual reality simulation environments to assess competence for emergency medicine learners. Academic Emergency Medicine **25**(2), 186–195 (2018)

[10] Roussin, C.J., Weinstock, P.: Simzones: an organizational innovation for simulation programs and centers. Academic Medicine **92**(8), 1114–1120 (2017)

[11] Birt, J., Stromberga, Z., Cowling, M., Moro, C.: Mobile mixed reality for experiential learning and simulation in medical and health sciences education. Information **9**(2), 31 (2018)

[12] Anderson, C.: Presenting and evaluating qualitative research. American journal of pharmaceutical education **74**(8) (2010)

[13] Georgiou, P.G., Black, M.P., Lammert, A.C., Baucom, B.R., Narayanan, S.S.: "that's aggravating, very aggravating": is it possible to classify behaviors in couple interactions using automatically derived lexical features? In: International Conference on Affective Computing and Intelligent Interaction, pp. 87–96 (2011). Springer

[14] Aufegger, L., Bicknell, C., Soane, E., Ashrafian, H., Darzi, A.: Understanding health management and safety decisions using signal processing and machine learning. BMC medical research methodology **19**(1), 1–12 (2019)

[15] Dadaleares, T.S., Crawford, S.B.: The healthcare simulation technology specialist and audio/video technology. In: Comprehensive Healthcare Simulation: Operations, Technology, and Innovative Practice, pp. 159–187. Springer, Cham (2019)

[16] Ito-Masui, A., Kawamoto, E., Esumi, R., Imai, H., Shimaoka, M.: Sociometric wearable devices for studying human behavior in corporate and healthcare workplaces. BioTechniques **71**(1), 392–399 (2021)

[17] Coro, G., Massoli, F.V., Origlia, A., Cutugno, F.: Psycho-acoustics inspired automatic speech recognition. Computers & Electrical Engineering **93**, 107238 (2021)

[18] Cutugno, F., Leone, E., Ludusan, B., Origlia, A.: Investigating syllabic prominence with conditional random fields and latent-dynamic conditional random fields. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)

[19] D'Anna, L., Petrillo, M.: Sistemi automatici per la segmentazione in unità tonali. In: Atti delle XIII Giornate di Studio del Gruppo di Fonetica Sperimentale (GFS), pp. 285–290 (2003)

[20] Boersma, P., *et al.*: 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, pp. 97–110 (1993). Citeseer

[21] Cutugno, F., Coro, G., Petrillo, M.: Multigranular scale speech recognizers: Technological and cognitive view. In: Congress of the Italian Association for Artificial Intelligence, pp. 327–330 (2005). Springer

[22] Cutugno, F., D'Anna, L., Petrillo, M., Zovato, E.: Apa: Towards an automatic tool for prosodic analysis. In: Speech Prosody 2002, International Conference (2002)

[23] MacQueen, J., *et al.*: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967). Oakland, CA, USA

[24] WaveSurfer: Software Guide for L541. https://phonlab.sitehost.iu.edu/wsman157/wsman10.htm (2021)

[25] Coro, G., Cutugno, F., Caropreso, F.: Speech recognition with factorial-hmm syllabic acoustic models. In: INTERSPEECH, pp. 870–873 (2007)

[26] Google LLC: Speech-to-Text Cloud Service. https://cloud.google.com/speech-to-text/docs/basics (2021)

[27] Povey, D.: The KALDI ASR toolkit. https://kaldi-asr.org/ (2021)

[28] Voxforge: VoxForge speech corpora. http://www.voxforge.org (2021)

[29] Coro, G., Panichi, G., Pagano, P., Perrone, E.: Nlphub: An e-infrastructure-based text mining hub. Concurrency and Computation: Practice and Experience **33**(5), 5986 (2021)

[30] Laerdal s.r.l.: SimNewB simulator. https://laerdal.com/it/products/simulation-training/obstetrics--paediatrics/simnewb/ (2021)

[31] Laerdal s.r.l.: SimView software. https://laerdal.com/it/ProductDownloads.aspx?productId=382 (2021)

[32] Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)

[33] Fleiss, J.L., Levin, B., Paik, M.C.: Statistical Methods for Rates and Proportions. John Wiley & Sons, 111 River Street Hoboken NJ, 07030. United States (2013)

Table 1 Descriptions, signal-to-noise ratio (SNR), and duration of the study cases used for the performance evaluation of the proposed workflow.

Study case	Description	SNR	Duration
T1	Simulation of newborn with respiratory failure. Two team members, doctor (female) and assistant (female). Good quality audio and low noise level.	14.17	10 min 57 s
T2	Simulation of hypotonic newborn. Two team members, doctor (female) and assistant (female). Medium-low audio quality and substantial stationary white noise.	6.76	6 min 36 s
T3	Simulation of newborn with respiratory failure. Two team members, doctor (female) and assistant (female). Medium quality audio and stationary white and violet noise.	11	8 min 8 s
T4	Simulation of newborn with respiratory failure. Two team members, doctor (female) and assistant (female). Good quality audio and a low level of stationary white and blue noise.	11.14	6 min 48 s
T5	Simulation of hypotonic newborn. Two team members, doctor (female) and assistant (female). Low quality audio, low volume, and substantial stationary white and violet noise. Very long silences (infrequent speech).	6.8	9 min 21 s
T6	Simulation of hypotonic newborn. Three team members, doctor (male) and 2 assistants (female and male). Good quality audio and high volume. Low violet noise.	11.5	6 min 26 s
T7	Simulation of hypotonic newborn. Three team members, doctor (female) and 2 assistants (female and male). Good quality audio and high volume. Low amount of violet noise.	12.3	6 min 36 s
T8	Simulation of newborn with respiratory failure. Two team members, doctor (female) and assistant (female). Low audio quality and low volume. High level of white noise. Long silences (infrequent speech).	4.63	9 min 19 s
T9	Simulation of newborn with respiratory failure. Two team members, doctor (male) and assistant (female). Low audio quality and low volume. High level of white noise.	4.77	7 min 35 s
T10	Simulation of hypotonic newborn. Three team members, doctor (female) and 2 assistants (females). Good quality audio and high volume. Medium amount of white noise.	10.05	7 min 23 s

Table 2 Word and sentence recognition accuracy of the Google Speech-to-Text service on our 10 study cases. Word accuracy refers to the words transcribed by the Google service, whereas sentence accuracy also accounts for missing words in the transcription.

	Average	**Minimum**	**Maximum**
Word accuracy	46.87%	23.31%	71.33%
Sentence accuracy	9.37%	4.66%	14.27%

Table 3 Performance evaluation of our workflow over our 10 study cases, and on all recordings considered as one overall case (*Overall*). Fleiss' agreement interpretation is reported for kappa values.

Study case	**Accuracy**	**Precision**	**Recall**	**F1**	**Kappa**	**Kappa Interpretation**	**SNR**
T1	0.65	0.38	0.59	0.47	0.218	Fair	14.17
T2	0.73	0.67	0.73	0.7	0.455	Moderate	6.76
T3	0.71	0.45	0.36	0.4	0.214	Fair	11
T4	0.71	0.44	0.92	0.6	0.413	Moderate	11.14
T5	0.7	0.22	0.29	0.25	0.066	Slight	6.8
T6	0.65	0.39	0.75	0.51	0.277	Fair	11.5
T7	0.58	0.28	0.64	0.39	0.135	Slight	12.3
T8	0.5	0.25	0.2	0.22	0.14	Slight	4.63
T9	0.37	0.25	0.5	0.33	0.152	Slight	4.77
T10	0.63	0.53	0.56	0.55	0.24	Fair	10.05
Overall	0.64	0.39	0.57	0.47	0.21	Fair	9.3

Table 4 Summary of the proposed workflow performance on all study cases.

Number of overall extracted dialogue segments	426
Number of potentially ineffective communication segments	170
Correctly labelled potentially ineffective communication segments	57.3%
Correctly labelled viable communication segments	66.7%
Correctly labelled communication segments overall	64.1%
Actual percentage of potentially ineffective communication segments over total segments	27.5%
Percentage of detected potentially ineffective communication segments over total segments	40%
Gold-standard words from comments on potentially-ineffective dialogues that were contained in the automatic *word cloud*	59%

Table 5 Cross-correlation between performance metrics (on 10 study cases) and signal-to-noise ratio (SNR).

	Cross-correlation with SNR
Accuracy	0.52
Precision	0.28
Recall	0.5
F1	0.39
Kappa	0.3

Table 6 Percent weight (with respect to the sum) of the most important nouns and verbs of the *word cloud* produced by our workflow out of all study cases. All reported numbers are percentages referring to the importance and repetition of the words in the audio segments containing *potentially ineffective* communication.

abbiamo	3.16	bisogna	1.05	sanno	0.76	aiuto	0.48	lavo	0.48
vediamo	2.65	nata	1.05	vuoi	0.76	pochino	0.48	frequenza	0.48
sono	2.55	leggo	1.05	hai	0.69	sistema	0.48	ricordo	0.48
faccio	2.48	tirarlo	1.05	fatto	0.69	basta	0.48	mantenere	0.48
lascio	2.11	giorno	1.05	gruppo	0.63	quant	0.48	corsa	0.48
messo	1.81	favore	1.05	fai	0.63	saturimetro	0.48	passivo	0.48
ho	1.73	serve	1.05	caso	0.63	dire	0.48	passati	0.42
piange	1.68	dimmi	1.05	ritirare	0.63	voglia	0.48	dimenticare	0.42
fare	1.52	dicono	1.05	mantenerli	0.63	dare	0.48	funzione	0.42
inizio	1.47	soccorso	1.05	chiama	0.63	tenerlo	0.48	credo	0.42
minuti	1.47	andiamo	1.05	smettere	0.63	doccia	0.48	senti	0.42
vabbè	1.39	tubo	1.05	vedo	0.63	raga	0.48	ritirato	0.42
va	1.24	studiamo	1.05	andava	0.63	bisogno	0.48	riscaldato	0.42
devo	1.12	confessioni	1.05	so	0.63	metti	0.48	prende	0.42
terapia	1.05	astrazione	1.05	accesa	0.63	materiale	0.48	dovuto	0.42
bracciale	1.05	adrenalina	1.05	caldo	0.63	casino	0.48	mettila	0.42
facendo	1.05	sai	1.05	continuare	0.63	chiamiamo	0.48	guardarlo	0.42
facciamo	1.05	respira	1.05	saturazione	0.63	scrivi	0.48	ventilare	0.42
carmine	1.05	pressione	1.05	voglio	0.63	finisci	0.48	dice	0.42
metterlo	1.05	aveva	1.05	aspettiamo	0.63	illusione	0.48	situazione	0.42
pesa	1.05	carlo	1.05	macchina	0.63	attivo	0.48	lasci	0.42
riesce	1.05	essere	0.95	segnale	0.63	prendere	0.48	teniamo	0.42
vincere	1.05	guarda	0.91	fa	0.55	arrivi	0.48	dobbiamo	0.42
stanno	1.05	siamo	0.84	smettila	0.53	ascolta	0.48	scende	0.42

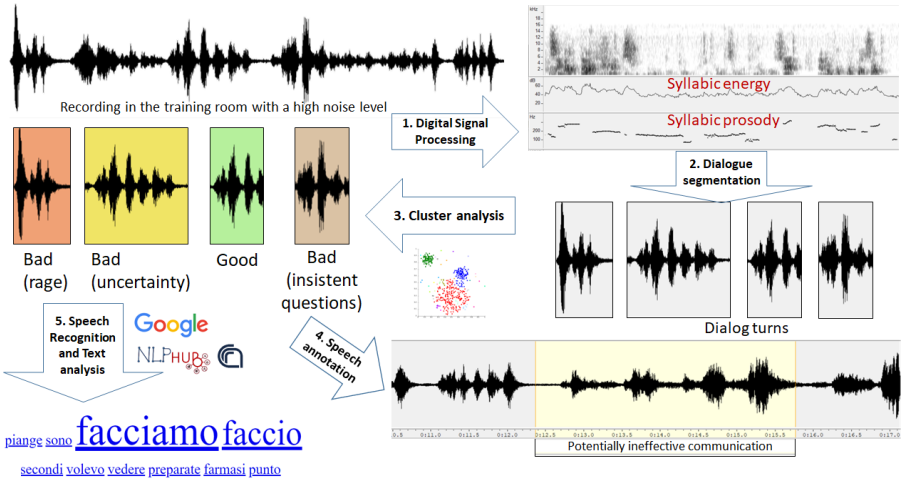

Fig. 1 Overall scheme of the proposed workflow: Step 1 (digital signal processing) calculates energy and pitch at a syllabic scale; step 2 (dialogue segmentation) divides the audio into portions with coherent intonation contours (tone units); step 3 (cluster analysis) detects the tone units containing *potentially ineffective* verbal communication; step 4 (speech annotation) produces an annotation file specifying the intervals of *potentially ineffective* communication; step 5 (speech recognition and text analysis) transcribes the audio of the tone units through an automatic speech recogniser and extracts a *word cloud* of ineffective communication keywords. The cloud indicates major communication issues, e.g., "facciamo" (let us do) has a higher weight than "faccio" (I do), suggesting that the team leader is uncertain about which actions to take and would require support to improve leadership and organisation skills.

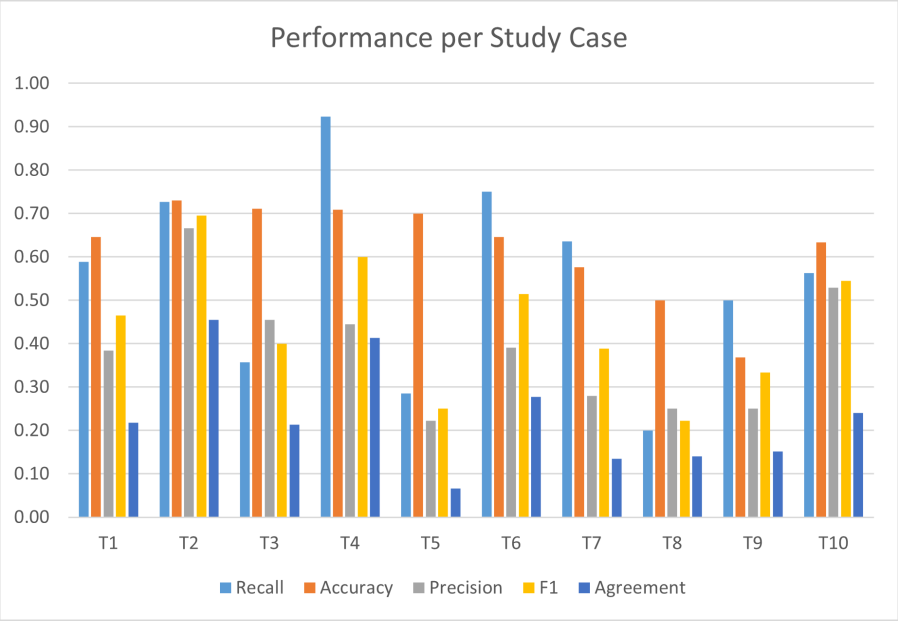

Fig. 2 Distribution of performance measurements over our 10 study cases: accuracy is the total portion of correctly labelled audio segments containing either *potentially ineffective* (PI) or *viable* communication (V); precision measures the ratio of correctly detected PI segments to the total detected PI segments; recall measures the ratio of correctly labelled PI segments to all PI segments of the gold standard; F1 is the harmonic mean of precision and recall and indicates how much the workflow is balanced between them; agreement, measured through Cohen's kappa, measures the agreement between the classifications produced by our workflow and those contained in the gold standard, weighted with respect to chance agreement. All performance measurements are maximum for T2. The worst study case in terms of agreement and F1 is T5, whereas T9 has the lowest accuracy.

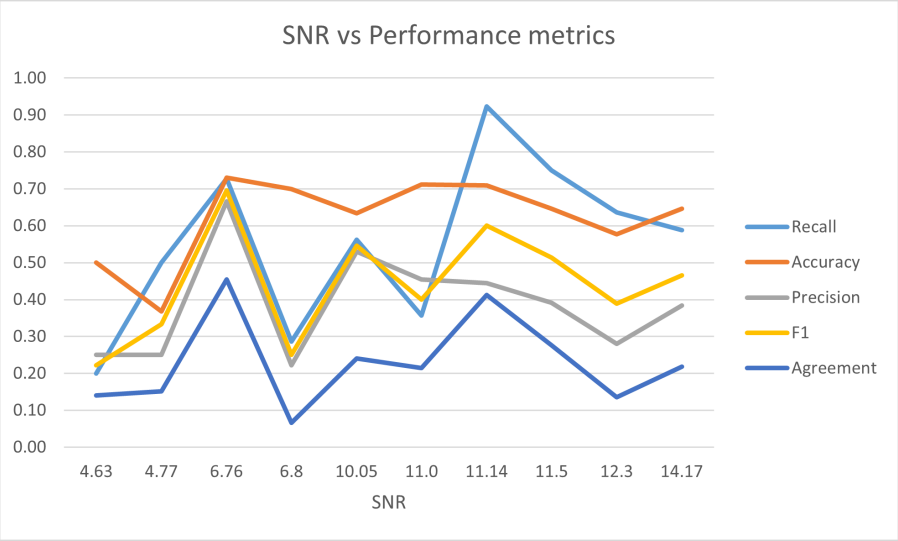

Fig. 3 Trends of the performance metrics at the lowering of audio noise level, i.e., at the increase of the signal-to-noise ratio (SNR). The chart highlights a poor correlation between SNR and performance. Higher and lower F1 and agreement are evenly achieved on cleaner audio and noisy audio. Accuracy decreases when noise is very high ($SNR < 6.76$) due to complete dialogue unintelligibility, preventing correct detection of *potentially ineffective* communication.

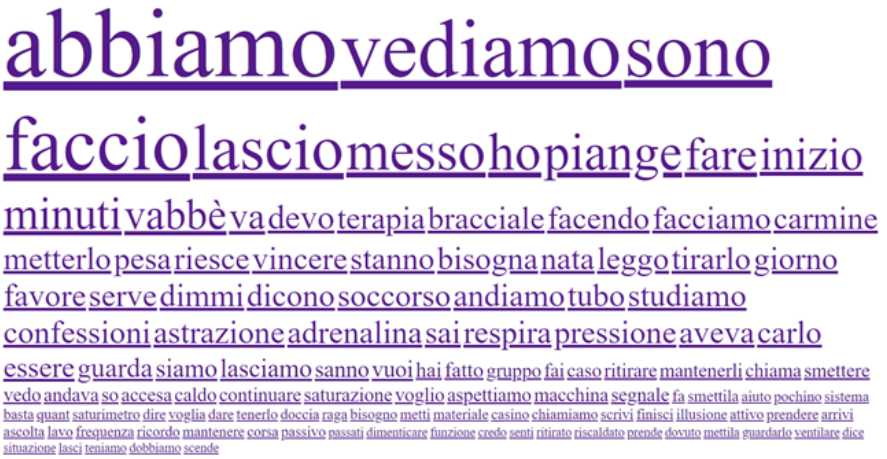

Fig. 4 *Word cloud* produced by the proposed workflow out of the transcriptions of all study cases. The font sizes (weights) are proportional to the frequencies of the words during *potentially ineffective* communication. This representation summarises information that allows trainers to identify major communication issues. For example, greater-weight words in the first person plural - e.g., "*abbiamo*" (we have), "*vabbè*" (oh well) - indicate general operational uncertainty and low self-confidence of the team leader.