Multimedia Tools and Applications (2026) 85:555
https://doi.org/10.1007/s11042-026-21723-7

Real-time speech-to-speech translation with urgency preservation in time-sensitive scenarios

Simone Pinna[1] · **Francesca Maridina Malloci**[1] · **Mirko Marras**[1] · **Diego Reforgiato Recupero**[1,2] · **Daniele Riboni**[1] · **Giuseppe Scarpi**[1,2]

Received: 15 January 2026 / Revised: 12 May 2026 / Accepted: 1 June 2026
© The Author(s) 2026

Abstract
In a global society where collaboration increasingly involves speakers of different languages, communication barriers remain a significant challenge. The problem becomes particularly relevant in urgent scenarios, where critical information must be delivered rapidly and understood correctly across speakers of several languages. To address this challenging issue, we introduce a multilingual framework for near real-time speech-to-speech translation that explicitly detects and preserves urgency. The system extends the cascade paradigm by integrating audio- and text-based urgency recognition with urgency-aware speech synthesis, and has been implemented as an open-source prototype accessible through WebRTC. As an additional contribution, we created the first publicly available multilingual dataset explicitly annotated for urgency, comprising 200 recordings in English, French, German, Dutch, and Italian. The framework has been evaluated through both quantitative and qualitative analyses. The experimental results indicate that our system achieves real-time performance while maintaining translation accuracy and effectively preserving perceived urgency. While most current solutions are offered by commercial providers and deployed on the cloud, raising significant privacy concerns, our system is released open-source and can be executed on premises using a mid-range server. The dataset is publicly available at https://zenodo.org/records/17502378.

Keywords Speech-to-speech translation · Urgency detection · Multilingual speech dataset

1 Highlights

- Near real-time speech-to-speech translation preserving urgency tone.
- Multilingual dataset of 200 urgent and non-urgent speech recordings.
- Dual urgency detection using audio prosody and text-based emotion cues.
- Modular, privacy-preserving, on-premise REST-API and WebRTC architecture.

Extended author information available on the last page of the article

Published online: 10 June 2026

- End-to-end latency < 3s with strong BLEU, WER, and human-rated MOS scores.

2 Introduction

In multicultural environments, interactions among speakers of different languages are common. Although English is often used as a lingua franca, it frequently obscures the nuances and prosody of native speech. While language barriers may be inconvenient in everyday situations, they can become dangerous in emergencies. For example, language misunderstandings on construction sites may cause safety incidents, while in healthcare they can delay treatment and worsen patient outcomes [1]. In the healthcare sector, poor communication often leads to rapid deterioration in the patient's condition, both during check-ups and in immediate emergencies where timeliness is crucial [2]. These problems can be mitigated by using speech-to-speech translation (S2ST) technologies. The requirements of these systems depend on their usage context and on their ability to accurately recognize, translate, and synthesize speech. In emergency contexts, low latency is crucial, and systems lacking near real-time capabilities are ineffective. The most widely used solutions are Google Translate[1] (Live mode) and Microsoft Translator[2]. Their widespread use is motivated both by the popularity of the underlying platform in which they are integrated and by their extensive coverage of language pairs. However, in emergency situations, they have limitations, both in terms of the reliability of translations (often due to errors in speech-to-text) and in terms of interaction with the speaker, due in part to excessive latency [3]. Proprietary systems such as those mentioned above also raise privacy concerns. These systems rely on corporate cloud infrastructures to process incoming speech. Consequently, their use requires trust in vendors' data-handling policies. The use of Google Translate in healthcare raises concerns among users who perceive privacy as a frequent risk [4]. Furthermore, they are highly dependent on the availability of a fast internet connection, which may be unavailable in potential emergency situations. Neither system accounts for the communicative context in which translation occurs, and the audio generated does not reflect the user's mood, determining possible misinterpretation of the content. This problem is addressed by Meta's SeamlessExpressive[3], an end-to-end S2ST model capable of replicating the speaker's prosodic characteristics. However, it is designed to preserve emotional tone in normal situations and not in situations where prosodic values are extreme, such as in emergencies.

In light of these open issues, in [5] we have developed a preliminary demonstrator of a cascade pipeline for S2ST in near real time, preserving any urgency tone used. That system combines Automatic Speech Recognition (ASR), Machine Translation (MT), and Text to Speech (TTS) models applied to audio chunks. Urgency is recognized by detecting prosodic features in the speaker's voice, and it is reflected in the output audio through the use of urgent reference samples. While that prototype validated the feasibility of near real-time urgency-aware translation, it was limited to a proof-of-concept demonstration. In this work, we present a new and fully developed framework that builds upon our earlier preliminary demonstrator, extending it with additional modules, a multilingual urgency dataset, and a

[1] https://translate.google.com/

[2] https://www.microsoft.com/it-it/translator/

[3] https://seamless.metademolab.com/expressive/

comprehensive experimental validation. We extend and improve the initial architecture in several dimensions.

Specifically, to enhance the accuracy of urgency classification, we introduced text-based emotion recognition based on speech transcription. Indeed, emotions such as fear and anger are common in emergency situations and are a significant indicator for detecting urgency, to be used alongside audio-based recognition [6]. To train the urgency recognition modules and for evaluating our system, it is essential to use a dataset of urgent and non-urgent speech samples. Although several public speech datasets are available for evaluating generic S2ST systems, to the best of our knowledge no urgency-aware dataset is currently available. Existing emotional datasets such as EmoDB [7], TESS [8], RAVDESS [9], and IEMOCAP [10] do not provide a specific label marked as "urgency". In response to this need, we created the first multilingual audio-transcription dataset explicitly annotated for urgency. The dataset consists of 200 recordings, equally divided into urgent and non-urgent ones, covering the following languages: English, French, German, Dutch, and Italian. To support robustness evaluation, each recording was additionally augmented with five acoustic degradation conditions, yielding a total of 1,000 additional audio samples.

Moreover, we implemented a context-aware chunking mechanism with overlapping windows and adaptive silence detection to improve segmentation and transcription robustness. The architecture was redesigned using a modular REST-API approach, allowing flexible integration into remote conferencing and collaborative applications. The new version supports deployment through WebRTC, enabling distributed real-time communication, and introduces an on-premise open-source implementation that mitigates privacy concerns associated with cloud-based systems.

Furthermore, we performed an experimental evaluation with the urgency-aware dataset to assess our solution in terms of efficiency and effectiveness, considering both quantitative and qualitative indicators. The quantitative analysis considered end-to-end latency, which consistently remained below three seconds. We also evaluated transcription accuracy using Word Error Rate (WER), translation quality using Bilingual Evaluation Understudy (BLEU), and system determinism through consistency across repeated trials. Our system achieves an average end-to-end latency below three seconds, a Word Error Rate (WER) of 8.28% for non-urgent audios and 10.30% for urgent audios, BLEU scores of 0.40 (for non-urgent audios) and 0.33 (for urgent audios), and full consistency across repeated trials. The multimodal urgency classifier, combining audio-based and text-based detection through an OR fusion strategy, achieves an F1-score of 0.929. To complement these automatic measures, we conducted a qualitative evaluation based on human judgments. Native speakers rated the outputs using the Mean Opinion Score (MOS) on a five-point Likert scale, assessing both translation quality and urgency preservation. Integrating these measures is essential, since latency alone does not capture the perceptual and affective dimensions of urgency-aware speech-to-speech translation. Qualitatively, native speakers rated translation quality with MOS values of 3.69 for non-urgent audios and 3.38 for urgent ones, and urgency preservation with a MOS of 3.34. These results indicate that the system supports near real-time communication while maintaining translation accuracy and effectively conveying urgency.

Compared to existing solutions, our system mitigates privacy concerns avoiding the need to communicate speech data to third-party cloud infrastructures. Indeed, our software pipeline is lightweight and can be run entirely locally on an on-premise server. Its hardware

requirements depend on the chosen model configuration. The default version, designed for near real-time requirements, can run efficiently on a mid-range server. Furthermore, the project is open source and uses standard technologies, allowing seamless integration into existing communication platforms.

Finally, we release the project as a fully open-source framework[4] that consolidates all components into a unified, production-ready system.

The final pipeline integrates modular elements for audio chunking, urgency detection at both audio and text levels, speech-to-text conversion, text translation, and speech synthesis with urgency preservation. It supports five languages (English, German, French, Dutch, and Italian) and is designed for efficient, on-premise execution as well as seamless integration into existing communication platforms.

Overall, the original contribution of this paper does not lie in the introduction of a novel S2ST system per se, but rather in the integration of urgency detection and preservation, a fundamental feature for many application domains that is not specifically addressed by existing systems.

In contrast to existing S2ST systems, which primarily focus on semantic accuracy or general prosodic expressiveness, current approaches do not explicitly model urgency as a communicative property, nor do they integrate urgency detection and preservation within a real-time multilingual pipeline.

In particular, our framework provides innovative contributions to the S2ST domain through:

(i) a multimodal algorithm for real-time urgency detection exploiting both audio prosodic features and transcribed text;
(ii) a technique for urgency-aware speech synthesis;
(iii) the public release of the first multilingual urgency-annotated dataset for this task;
(iv) an open-source near-real-time implementation of the full urgency-aware S2ST pipeline, evaluated using both quantitative metrics and human judgments.

This research has been carried out within the framework of the CORTEX^2 project (COoperative Real-Time EXperiences with EXtended Reality)[5]. CORTEX^2 is a Horizon Europe initiative aimed at enabling real-time, immersive, and interactive experiences by combining extended reality (XR), edge computing, and artificial intelligence technologies. The project focuses on developing open, interoperable frameworks to support distributed and low-latency applications in collaborative environments. Within this context, our work contributes to advancing multimodal communication systems capable of operating in near real time, with a specific focus on speech-to-speech translation in urgent scenarios.

The remainder of the paper is organized as follows. Section 2 discusses the most common S2ST tools, focusing on their usability in emergency situations. The datasets that can be used in such a context are also explored in detail. Section 3 introduces the proposed framework, describing the dataset and the architecture. Section 4 describes the experiments performed to establish the system's efficiency and effectiveness, both qualitatively and

[4] https://github.com/peter-cortex/peter_mediationgateway

[5] https://cortex2.eu/project/

quantitatively. In Section 5 the results achieved and the limitations are discussed. Finally, Section 6 summarizes the study carried out and outlines future work.

3 Related work

To contextualize our contribution within existing research efforts, this section reviews prior work related to urgency-oriented speech resources, real-time speech-to-speech translation systems, and expressive speech synthesis. Rather than providing a catalog of existing solutions, we critically analyze their limitations with respect to urgency modeling, real-time constraints, and deployment requirements.

3.1 Urgency and emotional speech resources

Emotional speech datasets have been widely used to study vocal expression and paralinguistic features. Well-known resources include the Berlin Emotional Speech Database (EmoDB) [7], the Toronto Emotional Speech Set (TESS) [8], RAVDESS [9], and IEMOCAP [10]. These datasets provide acted emotional speech labeled with categories such as anger, fear, happiness, sadness, and neutral states, and have supported advances in emotion recognition and expressive speech synthesis.

Despite their importance, these datasets focus on general affective states and do not explicitly model urgency. Urgency differs from emotion because it reflects communicative intent in time-critical situations rather than purely affective expression. Only a few studies have attempted to collect urgency-oriented speech corpora, mainly for specialized domains such as automotive alerts [11] or voicemail analysis [12]. However, these datasets are not publicly available, are domain-specific, and do not support multilingual evaluation.

This gap highlights the lack of publicly available multilingual datasets explicitly annotated for urgency, motivating the creation of the resource introduced in this work.

3.2 Real-time speech-to-speech translation systems

S2ST systems generally follow either cascade or end-to-end architectures [13].

Cascade approaches Combine automatic speech recognition, machine translation, and text-to-speech modules in sequential pipelines [14]. These systems are modular and robust, which explains their adoption in widely used commercial tools such as Google Translate and Microsoft Translator. However, cascade pipelines suffer from latency accumulation and error propagation across modules. Furthermore, they typically prioritize semantic accuracy and broad language coverage while neglecting prosodic aspects and communicative context. Their reliance on cloud infrastructures also raises privacy and connectivity concerns in emergency scenarios [4].

End-to-end approaches Aim to directly map source speech to target speech, reducing pipeline complexity and potentially improving naturalness [15]. Recent models such as SeamlessM4T [16], Meta SeamlessExpressive [17], and Google Translatotron [18] explore expressive translation, speaker identity preservation, and style transfer. While these sys-

tems achieve competitive translation quality, they often require large training datasets, have substantial computational demands, and rarely provide latency guarantees compatible with real-time applications.

More importantly, both cascade and end-to-end systems focus primarily on semantic fidelity and vocal style preservation. Urgency, as a communicative property influencing how information should be delivered in time-sensitive contexts, remains largely unexplored.
Beyond application-oriented S2ST systems, several complementary research areas are relevant to urgency-aware communication. In speech processing, prosodic and paralinguistic analysis has long investigated how acoustic cues such as pitch dynamics, temporal structure, spectral energy, and speaking rate convey stress, arousal, and communicative intent [19]. More recently, representation-learning approaches based on self-supervised speech models and multimodal embeddings have enabled richer modeling of latent acoustic and semantic information for affective and expressive speech tasks [20, 21]. In parallel, multimodal fusion strategies combining acoustic, linguistic, and contextual information have been increasingly explored for emotion recognition and speech understanding [22–25]. Although these approaches are not specifically designed for urgency-aware speech-to-speech translation, they provide important methodological foundations for the modeling of urgency as a communicative property.

3.3 Expressive and prosody-aware speech synthesis

Modern text-to-speech systems rely on neural acoustic models that generate intermediate speech representations, which are subsequently converted into waveforms through neural vocoders. Architectures based on sequence-to-sequence and transformer models enable high naturalness and fine-grained control over speech characteristics.

Recent research has focused on expressive and emotional speech synthesis, where prosodic features such as pitch, speaking rate, and intensity are manipulated to convey affective states. Style-transfer techniques further allow models to mimic speaker identity and vocal style using reference audio samples [17].

However, urgency differs from emotional expressiveness. Emotional TTS aims to reproduce affective states such as happiness or sadness, whereas urgency reflects communicative intent under time-critical conditions. Urgent speech is characterized by functional prosodic patterns, including increased speech rate, pitch variability, and vocal intensity, that serve to accelerate comprehension rather than convey emotion.

Existing acoustic modeling and waveform-generation approaches do not explicitly incorporate urgency as a distinct communicative dimension, nor do they integrate urgency detection with multilingual real-time translation pipelines.

3.4 Positioning of the proposed work

In summary, prior research provides important advances in emotional speech modeling, real-time translation, and expressive synthesis. However:

- no publicly available multilingual dataset explicitly targets urgency;
- existing S2ST systems prioritize semantic fidelity or vocal style rather than communica-

tive urgency;
- expressive TTS approaches focus on emotion and speaker identity without modeling urgency as an operational constraint;
- deployment aspects such as privacy-preserving on-premise execution are often neglected.

Our work addresses these gaps by integrating multilingual speech-to-speech translation with urgency detection and prosody-aware synthesis in a unified real-time framework. This combination enables timely, accurate, and context-aware communication in emergency and XR-assisted scenarios. Recent end-to-end S2ST models have shown promising results in terms of translation accuracy and expressive speech generation. However, these architectures are typically less modular and less easily customizable than cascade-based pipelines. In this work, we adopt a cascade approach because urgency detection is performed at both the acoustic and textual levels, which requires explicit access to intermediate representations such as prosodic features and speech transcriptions. Furthermore, current end-to-end S2ST systems do not explicitly model urgency as a communicative state, and adapting them to this task would require large-scale audio–text corpora annotated for urgency, which are currently unavailable. Moreover, the cascade design allows parallel processing and fine-grained control over individual components, which is essential to meet strict latency constraints in emergency scenarios. Finally, many of the technologies underlying state-of-the-art S2ST models are either not fully open source or require substantial computational resources. These characteristics are incompatible with the goals of universality, portability, and on-premise deployment that guide the design of the proposed framework. Table 1 provides a concise comparison of the main emotional speech datasets discussed in this section. It highlights key aspects such as language, modality, and domain, emphasizing that only a few datasets explicitly address urgency, and none in a multilingual or publicly available form. This comparison further underscores the need for a new resource tailored to urgent vs. non-urgent speech. To further clarify the practical differences between existing systems and our approach, Table 2 summarizes the main characteristics of representative tools. It can be seen that proprietary tools can only be used on the cloud and do not offer services designed

Table 1 Comparison between existing emotional speech datasets and our proposed dataset

Dataset	Language	Modality	Urgency	Domain	Public
EmoDB (Berlin)	German	Audio	x	Emotion recog.	✓
TESS (Toronto)	English	Audio	x	Emotion recog.	✓
RAVDESS	English	Audio	x	Emotion recog.	✓
IEMOCAP	English	Audio–Video–Text	x	Multi-modal interaction	✓
Urgency–AV [11]	English	Audio	✓	Automotive	x
Urgency–Voicemail [12]	English	Audio	✓	Voicemail	x
Our Dataset	Multiple	Audio–Text	✓	Urgency recog.	✓

Table 2 Comparison between major speech-to-speech translation tools and our proposed system

System	Openness	Urgency pres.	Multiling.	Deploy
Google Translate (live)	Proprietary	No	Yes	Cloud
Microsoft Translator	Proprietary	No	Yes	Cloud
Meta SeamlessExpressive	Open (research)	No	Yes	Cloud
Google Translatotron	Open (research)	No	Partial	On-device
Our System	Open (research)	Yes	Yes	Both

for emergencies. Our proposal fills the gaps left by the intersection of proprietary and open solutions, combining the advantages of both.

4 Proposed framework

In this section, we present the proposed framework for near real-time voice translation with urgency detection and preservation. We first describe the multilingual dataset that we specifically acquired to train the model for distinguishing urgent from non-urgent speech. We then introduce the architecture of the speech-to-speech translation system, which is tailored to achieve accuracy, low latency, and urgency preservation. Finally, we illustrate the end-to-end workflow of the pipeline, highlighting the interaction among its components and the temporal sequence of the processing steps.

4.1 Evaluation dataset for urgency recognition

To the best of our knowledge, currently, there is no publicly available dataset to explicitly distinguish between urgent and non-urgent speech. To address this gap, we acquired a novel multilingual dataset tailored to our goals of urgency detection. We first wrote 40 recordings in Italian (20 urgent, 20 non-urgent), each lasting approximately 10-20 seconds. The recordings emulate representative emergency scenarios (e.g., fires, structural failures, accidents, sudden illness), as well as less critical urgent situations such as airport announcements, system alerts, and time-sensitive educational communications. In addition, we considered cases related to the use of a virtual reality headset. An example of these cases is: *Hey, you! Take off that headset right now, there's a staircase right in front of you! If you take another step, you'll fall! Watch out, I'm talking to you!*

Non-urgent texts consist of general-purpose content such as narrative excerpts, informational messages, and advertisements. An example of such texts is: *Since the beginning of time, humans have built relationships with others and given meaning to the world through dialogue and storytelling.*

These 40 texts were then translated into English, French, German, and Dutch using an automated translation tool[6], obtaining a total of 200 textual items (100 urgent, 100 non-urgent) across five languages. Finally, native speakers of German, French, Dutch, and Italian validated the texts and recorded the speeches in their respective languages, carefully

[6] https://www.deepl.com/en/translator

preserving the intended urgent or neutral prosodic tone. Each language was recorded by a single native speaker. As a consequence, the dataset does not allow for cross-speaker evaluation, and robustness across different voices and accents cannot be assessed in the current study.

For the English language, a text-to-speech system (TTS)[7] was used, setting a shouting feature to simulate an urgent tone.

English recordings were generated via TTS to ensure consistent pronunciation and prosodic realization across samples. Since English appears in every translation direction, either as source or target language, using a single TTS-generated voice reduces inter-speaker variability and guarantees uniform input conditions, improving comparability and reproducibility across experiments. However, we acknowledge that TTS-generated speech tends to exhibit more regular and less variable prosody than natural human speech, which may make urgency detection on English input artificially easier compared to other languages. An overview of the dataset composition across languages, including the number of urgent and non-urgent recordings and the speaker type, is reported in Table 3.

To increase acoustic variability and support a more robust evaluation of pipeline robustness, the original 200 recordings were further augmented by applying five degradation conditions: additive white Gaussian noise at 5, 15, and 30 dB Signal-to-Noise Ratio (SNR), environmental background noise, and room impulse response reverberation, yielding 1,000 additional audio samples. This extended dataset is publicly released alongside the original recordings and is used in the robustness experiments.

Having described the dataset used to evaluate urgency recognition, we now introduce the architecture of the proposed system.

4.2 Architecture

The proposed system currently supports five languages (English, German, French, Dutch, and Italian) and has been implemented as a prototype of a remote conferencing application based on Web Real-Time Communication (WebRTC) [26], as shown in Fig. 1 (a). This technology enables the direct transmission of multimedia content, such as audio and video, between browsers with minimal latency, relying on a peer-to-peer architecture that is suitable for immersive and low-latency collaboration scenarios [27].

From an implementation perspective, the system follows a REST-API architectural approach [28, 29] designed to ensure modularity and straightforward integration into different multimedia platforms. The APIs are organized into three main functional layers:

[7] https://www.texttovoice.online

Table 3 Overview of the multilingual dataset. Each language includes 20 urgent and 20 non-urgent speeches, with recordings provided by native speakers (except for English, where urgency was simulated using TTS)

Language	# Urgent	# Non-urgent	Total	Speaker type
Italian	20	20	40	Native speaker
English	20	20	40	TTS
German	20	20	40	Native speaker
French	20	20	40	Native speaker
Dutch	20	20	40	Native speaker
Total	**100**	**100**	**200**	–

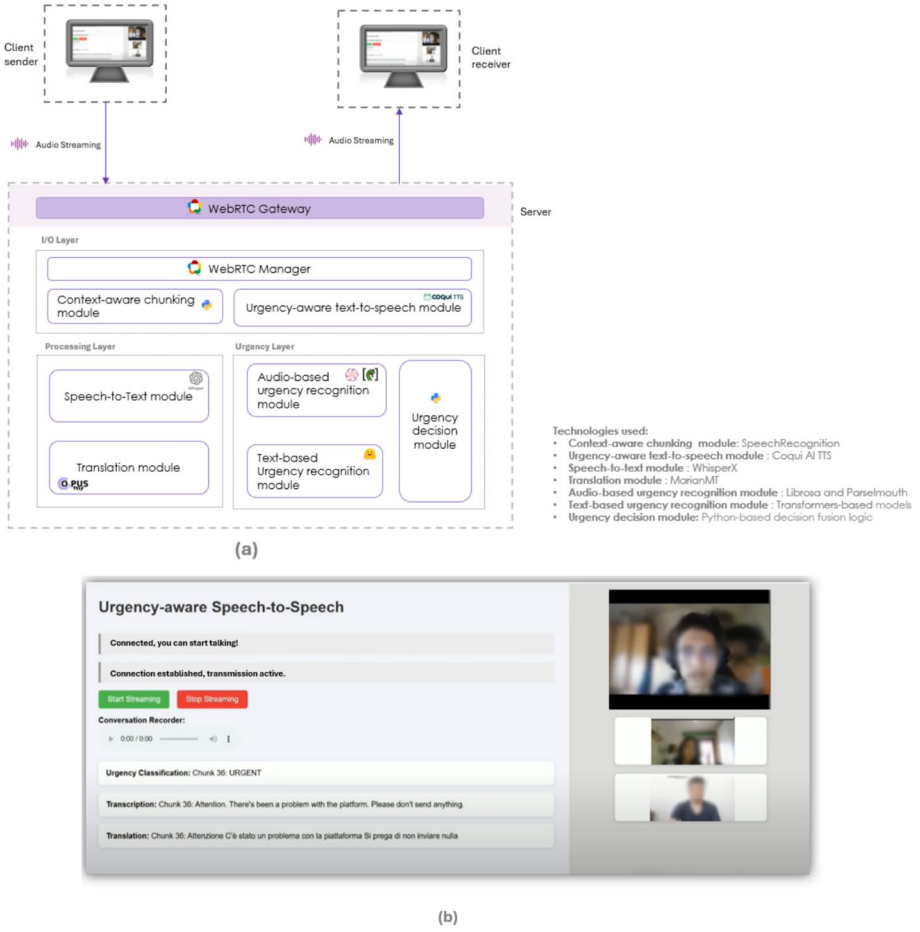


Fig. 1 System overview. (**a**) Architecture of the proposed speech-to-speech translation framework with urgency preservation. The WebRTC gateway handles client connections and routes the audio streams through three layers: I/O Layer (`WebRTC manager`, `context-aware chunking module`, `urgency-aware text-to-speech module`), Processing Layer (`speech-to-text module`, `translation module`), and Urgency Layer (`audio-based urgency recognition module`, `text-based urgency recognition module`, `urgency decision module`). (**b**) Graphical user interface developed for real-time conferencing

- *I/O Layer*: manages the input and output of the streams and adapts the signal to the pipeline. It includes the `WebRTC manager`, which handles the I/O logic with the `WebRTC gateway` (connection setup and session management), the `context-aware chunking` module for audio segmentation, and the `urgency-aware text-to-speech` module for speech synthesis with urgency preservation.
- *Processing Layer*: performs speech-to-text conversion and multilingual translation. It includes the `speech-to-text` and `translation` modules.
- *Urgency Layer*: analyzes audio and text samples to detect urgency. It includes the `audio-based urgency recognition`, `text-based urgency recognition`, and `urgency decision` modules, which combine their outputs through a

customizable strategy. Urgency detection can be performed in three modalities: audio, text, or combined. In the audio mode, urgency is inferred exclusively from prosodic features extracted by the `audio-based` module. In the text mode, it is identified through emotion recognition applied to the transcription generated by the `speech-to-text` module. In the combined mode, both audio and text analyses run in parallel, and urgency is triggered if either module provides a positive signal.

As illustrated in Fig. 1(a), communication is mediated by a `WebRTC gateway`. Each client establishes a direct connection with the gateway, which receives the streams, processes them through our pipeline, and then forwards the results to the recipient. Each client establishes two distinct WebRTC connections with the gateway, one for the input stream and one for the output stream. Figure 1(b) illustrates the user interface designed to offer an intuitive access point to the system's functionalities. The example depicts a representative educational scenario: during an online lesson, a teacher sends an urgent message (*Attention. There's been a problem with the platform. Please don't send anything.*) to students who speak different languages. The system detects the urgency, translates the message, and delivers it immediately, preserving both the semantic content and the urgent tone to ensure timely understanding and response. The figure also displays the single audio chunk in which the message was captured, along with its urgency classification, transcription and translation.

4.3 System modules

After presenting the overall architecture, we provide a detailed description of the individual modules that compose the system.

4.3.1 Context-aware chunking module

This module is responsible for segmenting the incoming audio stream into finite-length chunks that can be processed by the subsequent components of the pipeline. This segmentation step is essential, because the speech-to-text engine employed in our framework operates only on discrete chunks and cannot handle continuous audio streams. The maximum chunk duration can be dynamically configured according to application requirements. To improve the recognition of words occurring at the boundaries of consecutive chunks, the module employs an overlapping sliding window approach. In the current implementation, the maximum chunk length is set to 5 seconds with an overlap of 0.3 seconds. Additionally, a silence detection mechanism adaptively terminates a chunk after a predefined period of silence, ensuring efficient segmentation. Overall, this module manages the incoming stream and divides it into manageable units. The implementation relies on the `SpeechRecognition` library[8].

4.3.2 Speech-to-text module

This module generates the textual representation of each audio chunk. The transcriptions are placed in queues, from which the subsequent modules retrieve the textual data for further processing. The transcription is generated by an Automatic Speech Recognition (ASR)

[8] https://pypi.org/project/SpeechRecognition/

model based on `WhisperX`[9], a variant of OpenAI's Whisper that employs the Faster Whisper backend, making it more suitable for real-time applications. In our implementation, the turbo model is used as the default configuration to balance speed and accuracy. Although Whisper-based models support automatic language identification, preliminary tests showed that enabling this functionality introduces additional processing overhead for each audio chunk. Since the proposed framework targets near real-time communication in emergency scenarios, we opted for a configuration where the source language is specified at session initialization. The audio chunks are stored in a queue and processed by a thread, which in turn forwards the transcribed text to the incoming queue of the Translation module.

4.3.3 Audio-based urgency recognition module

This module recognizes urgency by analyzing audio chunks and comparing prosodic features of the speaker's voice against thresholds derived from established findings in the literature [30–32]. While the adopted thresholds are fixed heuristic values, preliminary sensitivity checks indicated that moderate variations of these parameters do not significantly alter the detection behavior, suggesting that the selected ranges capture stable prosodic trends associated with urgent speech. These ranges were determined by extracting prosodic features from audio recordings labeled as "urgent". By examining the metric values identified as relevant in the existing literature, threshold values were defined to indicate a sufficient level of urgency in the speaker's voice.

In particular, we have taken into consideration four main metrics of audio features:

- *fundamental frequency (f0) metrics*, which capture pitch level and variability (mean, standard deviation, min/max, range). Indeed, the fundamental frequency has been shown to play an important role in recognizing urgency, with higher values indicating a greater perception of urgency [31];
- *pause metrics*, which describe the timing of silences (average pause duration and pause count), where shorter and more frequent pauses reflect urgent delivery;
- *phrase metrics*, which measure the duration of speech segments between pauses, with shorter phrases indicating high-arousal or time-constrained speech; and
- *signal intensity metrics* (Root Mean Square (RMS) and peak amplitude), which quantify vocal energy and loudness, typically higher in urgent speech.

These features are extracted using `Librosa`[10] and `Parselmouth`[11] [33] libraries, allowing the system to classify speech segments as urgent or non-urgent without the use of AI models but solely through measurable acoustic properties. Table 4 details the acoustic features considered for audio-based urgency recognition.

The rule-based design was chosen to ensure low-latency processing and interpretability, which are critical requirements in real-time emergency scenarios.

The thresholds for urgency detection used in our system have been empirically chosen using a separate dataset that we collected for this purpose and that was not used in the evalu-

[9] https://github.com/m-bain/whisperX

[10] https://librosa.org/doc/latest/index.html

[11] https://pypi.org/project/praat-parselmouth/

Table 4 Acoustic features considered for audio-based urgency recognition

Metrics	Feature	Description
Fundamental frequency (f0)	Mean f0	Average pitch of the audio (Hz); higher values typically indicate urgency or emotional intensity.
	Std. deviation of f0	Variability of pitch around its mean; higher variability is often linked to urgent or expressive speech.
	Minimum f0	Lowest detected pitch (Hz).
	Maximum f0	Highest detected pitch (Hz).
	Pitch range	Difference between maximum and minimum pitch, reflecting the dynamic range of speech.
Pause	Average pause duration	Average length of pauses (s); shorter pauses suggest faster, more urgent speech.
	Number of pauses	Total number of pauses detected.
Phrase	Average phrase length	Mean duration of speech segments between pauses (s); shorter phrases often reflect urgency in high-stakes contexts.
Signal intensity	RMS amplitude	Average energy of the signal, indicating loudness; higher values are associated with louder, potentially more urgent speech.
	Peak amplitude	Maximum amplitude of the audio signal, corresponding to the loudest point in speech.

ation experiments, ensuring that the test data played no role in threshold estimation. The dataset used for threshold determination is publicly released as part of our dataset under the folder 'test_for_thresholds'.

4.3.4 Translation module

This module is responsible for translating the original text into the target language. Translation is performed using the `Marian MT` framework with `Helsinki-NLP/Opus-MT`[12] models due to their wide language coverage, open availability, and favorable trade-off between translation quality and computational efficiency. In contrast to larger neural MT systems, these models provide lower latency, making them well suited for real-time speech-to-speech applications. The translated text is finally forwarded to the Urgency-aware text-to-speech module.

4.3.5 Text-based urgency recognition module

This module is responsible for identifying urgency directly from text. It runs in parallel after the transcription stage performed by the speech-to-text module. In psychological literature, urgency is associated with states of intense negative affect, a construct often referred to as negative urgency [34, 35]. Among these, emotions such as anger and fear are particularly relevant, as they are commonly linked to heightened arousal and the perception of urgency.

[12] https://github.com/Helsinki-NLP/Opus-MT

Building on this evidence, our system performs textual urgency detection by focusing on these two emotions. The aforementioned emotions are particularly relevant in emotionally charged and high-risk contexts, as they strongly influence both decision-making processes and interpersonal communication, including verbal and non-verbal expressions [36]. Although anger and fear do not always correspond to actual emergencies in everyday life, in environments such as virtual reality simulations or high-risk workplaces their association with urgency becomes more pronounced, potentially facilitating rapid responses aimed at preventing negative outcomes. The detection is implemented using transformer-based emotion classification models available on the Hugging Face platform. These models are language-specific and allow the system to adapt to each supported language: for English, we employ a fine-tuned `DistilRoBERTa-base`[13] model for emotion classification; for French, a fine-tuned version of the `CamemBERT-base` transformer[14]; for German, a fine-tuned `RoBERTa` model[15] for emotion classification; and for Italian, the `IT-Emotion-Analyzer` model[16].

4.3.6 Urgency-aware text to speech module

This module converts translated text into speech, adapting the intonation to the detected level of urgency. The implementation is based on the `Coqui AI TTS`[17] library with the multilingual `xtts_v2` model, which supports all required languages.

XTTS-v2 is a zero-shot multi-speaker TTS system that generates speech conditioned on a short reference audio sample. From this reference, the model extracts a compact representation of the speaker's vocal characteristics, including prosodic properties such as pitch, speaking rate, and intensity, which guide the synthesis process to reproduce a similar style in the output waveform. The waveform is then produced by a neural vocoder operating on intermediate representations generated by an autoregressive language model conditioned on both the input text and the reference embeddings [37]. In our system, two sets of reference embeddings are extracted at pipeline initialization time: one from a neutral speech sample and one from an urgent speech sample exhibiting the prosodic profile typical of emergency communication. At inference time, the appropriate embedding set is selected based on the urgency label provided by the upstream decision module, without any additional computation. This design ensures that switching between urgent and non-urgent synthesis has no impact on latency. The TTS stage accounts for the largest share of the end-to-end processing time, yet the overall latency remains below three seconds across all tested configurations, confirming the suitability of the module for near real-time operation. A further consequence of this design is that urgency-aware synthesis does not alter the linguistic content of the output, only its prosodic rendering. This is consistent with the comparable BLEU scores observed across urgent and non-urgent conditions, confirming that urgency preservation does not negatively affect translation quality.

[13] https://huggingface.co/michellejieli/emotion_text_classifier

[14] https://huggingface.co/astrosbd/french_emotion_camembert

[15] https://huggingface.co/visegradmedia-emotion/Emotion_RoBERTa_german6_v7

[16] https://huggingface.co/aiknowyou/it-emotion-analyzer

[17] https://github.com/coqui-ai/TTS

4.4 Workflow

In the following, we illustrate the workflow for the combined urgency modality, which processes both text and audio to detect urgency, as illustrated in Fig. 2.

The system takes as input an audio stream (1), which is first processed by the `Context-aware Chunking` module (2). This component segments the continuous stream into finite-length chunks. Each audio chunk is then forwarded to the `Speech-to-Text` module (3), which produces the transcription of the spoken content. In parallel, the same audio chunk is sent to the `Audio-based Urgency Detection` module to detect urgency from acoustic/prosodic cues. Next, the original text produced by the `Speech-to-Text` module is passed to the `Translation` module (4) for translation into the target language and to the `Text-based Urgency Recognition` module to recognize urgency from the text. The outputs of the audio and text-based modules are combined in the `Urgency Decision` module (5), which determines whether urgency is present or not by assigning a label. This urgency label, together with the translated text, is then passed to the `Urgency-aware Text-to-Speech` module (5). Based on the detected urgency, this module converts the text into speech with a prosodic style consistent with the assigned urgency level. Finally, the generated audio, which preserves both the meaning and the urgent tone, is delivered as the system's output (6). The entire communication between peers is established through the `WebRTC` protocol, which enables real-time, peer-to-peer audio and video transmission.

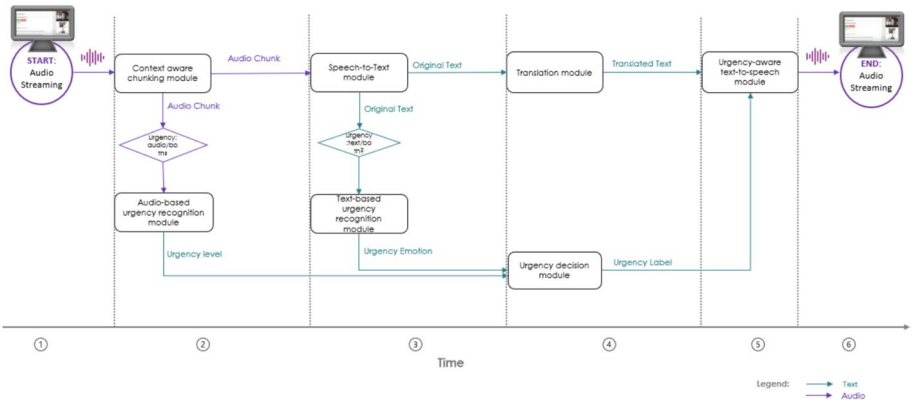


Fig. 2 Workflow of the proposed speech-to-speech translation pipeline with urgency preservation. The input audio stream is segmented into chunks by the `context-aware chunking` module and passed in parallel to the `speech-to-text` module and the `audio-based urgency recognition` module. The transcribed text is then processed by the `text-based urgency recognition` module, and both urgency signals (audio and text) are combined by the `urgency decision` module. The resulting urgency label is propagated to the `urgency-aware text-to-speech` module, which generates the translated audio while preserving urgent intonation. Green arrows indicate the text flow, while purple arrows indicate the audio flow. The horizontal axis represents the temporal sequence of the pipeline execution

5 Experimental evaluation

After detailing the system design and operational workflow, we now evaluate its effectiveness through quantitative and qualitative experiments. While quantitative metrics provide objective evidence of system performance in terms of accuracy and latency, they do not always capture perceptual and affective aspects such as the naturalness of speech or the correct rendering of urgency. Conversely, qualitative evaluations based on human judgments can highlight strengths and weaknesses that are not reflected by automatic scores alone. For this reason, we considered both perspectives, ensuring that the evaluation accounts for technical performance as well as user-perceived quality. Currently, no publicly available datasets or established benchmarks specifically target urgency detection in emergency speech. Most existing speech analysis studies focus on related but distinct tasks such as emotion recognition, stress detection, or affective computing. Consequently, direct baseline comparisons with urgency detection systems are not currently feasible.

For this reason, the proposed method should be interpreted as a lightweight heuristic baseline based on interpretable prosodic indicators rather than as a competing supervised classifier. Future work will investigate data-driven approaches, including supervised and end-to-end neural architectures, once sufficiently annotated urgency datasets become available.

By conducting these experiments, it is possible to answer Research Questions (RQs) that help summaries the results obtained. The main questions to ask are as follows:

- **RQ1**: Under different configurations, does the system achieve low latency performance ?
- **RQ2**: How accurate are the automatic transcription and translation modules ?
- **RQ3**: What is the contribution of acoustic and textual features to urgency classification, and which fusion strategy yields the best performance?
- **RQ4**: How are translation and urgency perceived by native speakers of the target language ?
- **RQ5**: Are performances consistent across different language pairs ?
- **RQ6**: How robust is the pipeline across different acoustic noise conditions ?

The answers to the above-mentioned RQs will be discussed in Section 4.3.

It should be noted that a direct comparison with commercial S2ST systems was not performed, as their proprietary implementations and lack of access to internal parameters hinder reproducibility and prevent a fair benchmarking process.

We recall that, in order to avoid data leakage, thresholds for audio-based urgency detection were chosen using audio samples not used for testing our system. For the remaining parts of our pipeline, we relied exclusively on existing pre-trained models, which were used as-is and were not fine-tuned on our data. Therefore, a conventional training/validation/test split is not applicable in this context, since no model fitting or parameter optimization was performed on the evaluation data. As a result, no model parameters were optimized on the evaluation dataset, and the reported results are not affected by overfitting to the test data.

5.1 Metrics

The evaluation was carried out using two complementary types of metrics: quantitative, capturing objective and measurable aspects of system performance, and qualitative, reflecting subjective human judgments of translation quality and urgency preservation. Quantitative metrics assess objective performance, while qualitative metrics capture human perception of translation quality and urgency preservation.

5.1.1 Quantitative metrics

- *Word Error Rate (WER)*: we use this metric to evaluate our speech-to-text module. WER measures the cost of restoring the output word sequence to the original input sequence [38]. The sequence hypothesised by the `Speech-to-Text` module is aligned with a reference transcription, and the number of errors is computed as the sum of substitutions (*S*), insertions (*I*), and deletions (*D*). If there are *N* total words in the reference transcription, the metric is defined as:

$$WER = \frac{S + I + D}{N} \times 100 \tag{1}$$

where the result is expressed as a percentage, with lower values indicating more accurate transcriptions [39].

- *Bilingual Evaluation Understudy (BLEU)*: evaluates the quality of the translation by comparing the one produced by our system with the one set as a baseline obtained through state of the art translation software. In this work, we use DeepL to obtain the baseline translation. BLEU is defined as:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \quad \text{where} \quad \text{BP} = \begin{cases} 1, & c > r, \\ \exp\left(1 - \frac{r}{c}\right), & c \leq r. \end{cases} \tag{2}$$

where p_n is the *n*-gram precision, w_n the weight of each *n*-gram (typically $w_n = 1/N$), *c* the length of the candidate translation, and *r* the length of the reference translation. Scores range from 0 to 1 [40]. Lower values indicate poorer quality, while values closer to 1 indicate higher translation accuracy.

- *Consistency*: checks the determinism of the system by verifying that identical inputs always lead to identical outputs.

$$\text{Consistency} = \frac{M}{N}, \tag{3}$$

where M is the number of times identical responses are produced for the same input, and N is the total number of repetitions of that input.

- *Latency*: measures the average processing time required to generate the system output for each audio chunk. Formally, if t_i is the processing time of chunk i and M is the total number of chunks, latency is defined as:

$$Latency = \frac{1}{M} \sum_{i=1}^{M} t_i \tag{4}$$

Lower values indicate faster system responsiveness and are essential for real-time applications.

5.1.2 Qualitative metrics

- *Mean Opinion Score (MOS)*: measures human perception of output quality on a five-point scale, following the ITU definition of opinion scores [41], which describes MOS as the *"value on a predefined scale that a subject assigns to his opinion of the performance of a system"*. Formally, if s_i is the score assigned by evaluator i and N is the number of evaluators, the MOS is defined as:

$$MOS = \frac{1}{N} \sum_{i=1}^{N} s_i \tag{5}$$

A higher MOS score indicates that the speech sounds more natural and closer to human speech, which is crucial for user satisfaction in real-world scenarios. In our evaluation, MOS was calculated for two aspects: (i) the semantic accuracy and syntactic correctness of the translations, and (ii) the degree to which the generated speech preserved the intended tone of urgency. In the following, we refer to the first metric as *MOS translation*, computed for both non-urgent and urgent audios, and to the second as *MOS urgency*, which was computed only for urgent audios. Although the MOS produces numerical ratings, it is regarded as a qualitative evaluation method because it captures human judgments rather than objective measurements. The resulting scores reflect participants' subjective perceptions of translation quality and urgency preservation, influenced by linguistic nuance, tone, and contextual understanding. Unlike the quantitative metrics indicated earlier, which are derived from automated, reproducible computations, the MOS provides insight into how users perceive the system's outputs, thus complementing quantitative analyses with a human-centered perspective on performance.

5.2 Experimental setup

The evaluation was conducted on a workstation equipped with an Intel® Core™ i3-7100 CPU, 16 GB RAM, and a 12 GB NVIDIA TITAN X GPU, with all models loaded onto the GPU.

To compute the quantitative metrics, we implemented an automated script that emulates user interaction through a WebRTC connection. The script feeds the system with audio chunks and collects the outputs for evaluation. For *WER* computation, we used the `jiwer` Python library, and for BLEU we employed the `nltk` sentence-level implementation. *Consistency* was assessed by repeatedly feeding identical inputs and verifying the invariance of the outputs. We configured all components to run deterministically (fixed decoding parameters and seeds where applicable). In our tests, identical inputs produced identical outputs, yielding consistency = 1.00. *Latency* was computed as the average processing time per chunk across multiple voice samples of varying durations and chunk sizes. This allows us to assess whether the system can operate in near real-time conditions.

To compute the qualitative metrics, namely *MOS translation* and *MOS urgency*, we involved native speakers as evaluators, one for each target language (French, German, Italian, and Dutch). Each evaluator assessed the outputs corresponding to their respective language. Due to the length of the evaluation task, participation at this stage was limited to a single evaluator per language. Consequently, the MOS results should be interpreted as an indication of linguistic accuracy and perceived urgency, rather than as statistically robust estimates. Each participant was asked to rate the system outputs on a five-point Likert scale (1 = very poor, 2 = poor, 3 = fair, 4 = good, 5 = excellent). For translation quality, participants answered the question: *"On a scale of 1 to 5, how do you rate the quality of the translation?"*. For urgency preservation, participants were asked: *"On a scale of 1 to 5, how do you rate the tone of urgency used in this situation?"*. Both questions were asked for both urgent and non-urgent audios. The experiments were conducted across eight translation directions to evaluate the system's performance in both urgency detection and preservation. Specifically, translations were performed from English to French, German, Italian, and Dutch, and reciprocally from each of these languages into English. For each language pair, a total of 40 recordings (20 urgent and 20 non-urgent) in the source language were processed through the system. Urgent audio messages contain emergency scenarios described in Section 3.1 The outputs were then translated into the target language, with the goal of assessing the system's ability to accurately render the semantic content while preserving, or intentionally disregarding, the urgency expressed in the original speech. The MOS was computed only for target languages other than English, since the English source recordings were generated using an automated text-to-speech tool, and therefore qualitative judgments on English outputs would not provide meaningful human perception data.

5.3 Results

This section reports the results obtained by responding to the RQs defined in Section 4, using the quantitative and qualitative metrics described in Section 4.1. We first report average results across all language pairs, distinguishing between urgent and non-urgent audios. This aggregated analysis offers a global perspective on translation accuracy, latency, and urgency preservation, summarizing the system's overall behavior under multilingual condi-

Table 5 Average pipeline latency measured per audio chunk, varying the maximum chunk size and the total speech duration

Chunk Length (s)	Speech Duration (s)						
	3	6	9	12	15	18	21
2	1.724	2.380	2.410	1.807	1.986	1.861	2.133
3	1.898	2.349	2.796	2.927	2.483	2.359	2.348
4	2.392	2.798	3.285	2.886	2.362	2.356	2.379
5	2.115	3.471	3.352	2.864	2.990	2.897	2.942

Table 6 Average processing time per pipeline stage, measured on natural speech chunks of 5 seconds, averaged across all 8 language pairs and both urgent and non-urgent samples

STT (s)	Translation (s)	TTS (s)	Urgency (s)	Emotion (s)	Total (s)
0.703	0.119	2.208	0.027	0.009	3.066

tions. Subsequently, we present the per-language results divided by urgent and non-urgent audios, which highlight variations across specific translation directions and allow a more detailed examination of linguistic and prosodic factors influencing performance.

5.3.1 RQ1 - latency analysis

Table 5 reports the *latency* results, averaged across all language pairs and across both urgent and non-urgent audio samples. Urgent and non-urgent recordings were aggregated because the urgency detection step does not introduce any additional computational load, and therefore does not affect the overall execution time. The table shows the average processing time per chunk for different combinations of speech duration and maximum chunk length. As expected, shorter chunks generally correspond to lower execution times: for instance, with a chunk length of 2 seconds, the average ranges from 1.724 s (speech duration 3 s) to 2.410 s (speech duration 9 s). We observe that variations in speech duration do not show a consistent trend, as values remain within a comparable range. For example, with a chunk length of 4 seconds, the execution time varies between 2.362 s (speech duration 15 s) and 3.285 s (speech duration 9 s). Across all configurations, the measured times therefore vary between a minimum of 1.724 s and a maximum of 3.471 s. These results highlight that our system provides sufficiently low latency to support several applications with near-real time requirements. Table 6 complements the previous analysis by reporting the average execution time broken down by pipeline stage, rather than varying chunk and speech duration parameters. Unlike Table 5, where fixed-length chunks were artificially imposed, these measurements were obtained by processing natural conversational audio segmented into 5-second chunks, thus reflecting a more realistic usage scenario. The results show that the Text-to-Speech stage dominates the total processing time, accounting for the majority of the end-to-end latency, while Translation contributes negligibly. The Urgency Audio and Emotion Detection stages, although present, do not impact the overall latency as they run in parallel with the main pipeline.

5.3.2 RQ2 - transcription and translation accuracy

Table 7 reports the results, averaged across all language pairs, for WER, BLEU, and consistency. The table includes two rows corresponding to non-urgent and urgent audio samples, respectively, allowing a direct comparison between the two categories.

The WER for non-urgent audios is 8.28%, indicating a limited number of transcription errors with respect to the reference. For urgent audios, the WER increases slightly to 10.30%, which may be influenced by the faster speech rate and more abrupt prosody typically associated with urgent messages. The BLEU scores for non-urgent and urgent audios are 0.40 and 0.33, respectively, suggesting that the translations maintain a reasonable correspondence with the reference outputs. Consistency for both categories is equal to 1, confirming that identical inputs consistently produce identical outputs, due to the fact that all models are deterministic. These results indicate that the inclusion of urgency-aware components does not negatively affect translation quality. To better understand the nature of the transcription and translation errors, we analyzed the most frequent edit operations produced by the ASR module. Substitutions are highly distributed and no single error pattern dominates the distribution: the most frequent substitutions (e.g., “you” → “yo” or “lastminute” → “minute”) occur only a few times each, suggesting that transcription errors are largely sporadic rather than systematic.

Deletions are more concentrated and primarily affect function words such as pronouns, articles, and prepositions (e.g., “you”, “a”, “we”, “of”), which typically have a limited impact on the intelligibility of the utterance. Insertions are less frequent and are often related to minor segmentation variations or repeated tokens. A known behavior of Whisper-based ASR systems was also observed in a small number of cases, where the model appends short closing expressions such as “thank you” even when they are not present in the audio. This phenomenon has been reported in previous analyses of Whisper models and is generally attributed to biases in the training data [42], where such expressions frequently occur at the end of spoken segments.

For the translation stage, most discrepancies involve lexical variation, paraphrasing, and minor grammatical differences rather than incorrect semantic interpretation. For example:

> REF: turn off the headset immediately it’s starting to overheat it could freeze or even worse get completely damaged if you want to keep using it you need to let it cool down for at least a few minutes

> HYP: turn off the headset right now its starting to overheat too much it might get stuck or worse damage yourself completely if you want to continue using it you have to let it cool for at least some minute

Although the semantic content of the message is largely preserved, the use of synonyms, paraphrasing, and slight structural variations leads to a reduction in BLEU score, which is known to penalize deviations from the reference formulation. Moreover, the BLEU score is indirectly influenced by transcription errors, as the translation module operates on ASR

Table 7 Averaged quantitative evaluation

Category	WER	BLEU	Consistency
non-urgent	8.28%	0.40	1.00
urgent	10.30%	0.33	1.00

Table 8 Averaged qualitative evaluation

	MOS translation	MOS urgency
non-urgent	3.69	
urgent	3.38	3.34

Table 9 p-values from the Mann-Whitney U test for WER and BLEU. Bold values indicate statistically significant differences ($p < 0.05$)

Direction	WER p-value	BLEU p-value
de→en	0.766	0.164
en→de	**0.004**	0.096
en→fr	0.051	0.074
en→it	**0.005**	0.617
en→nl	**0.022**	0.860
fr→en	0.303	0.072
it→en	0.640	0.140
nl→en	0.432	**0.0003**

outputs that may contain minor inaccuracies. Overall, these observations indicate that most errors are superficial and rarely alter the semantic content of the utterances, which is particularly important in emergency communication scenarios.

Considering the trade-off between delay and translation accuracy, previous studies showed that most people prefer a slightly longer delay (i.e., 3 or 4 seconds) to obtain more accurate translation [43]. For this reason, for the qualitative evaluation we fixed the maximum chunk length at 5 seconds, balancing latency with the need to preserve sufficient context for accurate transcription and translation. To complement the quantitative results reported in Table 7, we further assess whether the observed differences between urgent and non-urgent conditions are statistically meaningful. Specifically, we perform significance testing on sentence-level WER and BLEU scores using the Mann–Whitney U test [44], treating the two sets as independent samples. These quantitative findings should also be interpreted in light of the qualitative results reported in Table 8.

Table 9 reports the p-values from the Mann-Whitney U test across all language directions. While some statistically significant differences are observed (e.g., for WER in en→de, en→it, and en→nl, and for BLEU in nl→en), most comparisons do not show statistically significant variation. This indicates that, in the majority of cases, the differences observed may be attributed to random variation rather than systematic effects. Overall, these findings suggest that the system maintains relatively stable performance under urgent conditions. It is important to note that the urgent and non-urgent conditions are not paired at the utterance level, as they contain different input sentences. Therefore, part of the observed variation may be influenced by differences in content complexity rather than urgency alone. Future work will address this limitation by designing controlled, paired evaluations to better isolate the effect of urgency.

5.3.3 RQ3 - multimodal urgency classification performance

We evaluated the performance of the audio-based urgency recognition module and the text-based urgency recognition module, as well as their contribution to overall urgency classification. The results are summarized in Table 10. The evaluation was conducted at two different levels. The first evaluates complete audio files, where an audio clip is labeled as

Table 10 Urgency classification performance using different modalities and fusion strategies. Results are reported at both audio and chunk level

Strategy	Level	Precision	Recall	F1
Audio only	Audio	**1.000**	0.721	0.838
Text only	Audio	0.905	0.750	0.820
OR fusion	Audio	0.923	**0.936**	**0.929**
AND fusion	Audio	**1.000**	0.457	0.627
Audio only	Chunk	**1.000**	0.609	0.757
Text only	Chunk	0.918	0.388	0.546
OR fusion	Chunk	0.958	**0.785**	**0.863**
AND fusion	Chunk	**1.000**	0.213	0.351

urgent if at least one chunk within the clip is classified as urgent. The second evaluates individual chunks directly. The comparison between modalities shows that acoustic features provide a strong indicator of urgency. At the audio level, the audio-only system achieves perfect precision (1.000), indicating that whenever urgency is detected from prosodic cues the prediction is highly reliable. However, its recall remains lower (0.721), meaning that some urgent situations cannot be captured through acoustic cues alone. Text-based detection exhibits a different behavior. While it achieves a comparable recall (0.750), it introduces a small number of false positives, resulting in a slightly lower precision (0.905). These results suggest that urgency signals are not always expressed in the same modality. Some cases are conveyed primarily through prosodic characteristics, while others depend more strongly on the semantic content of the utterance.

Combining the two modalities improves performance. The OR fusion strategy exploits the complementary nature of acoustic and textual cues and achieves the best overall results, reaching a recall of 0.936 and an F1-score of 0.929 at the audio level. This represents a clear improvement over the individual modalities, whose recall values remain below this level. The same trend is observed at the chunk level, where OR fusion increases recall to 0.785 compared to 0.609 for audio-only and 0.388 for text-only detection, while maintaining a high precision of 0.958. These results indicate that combining modalities allows the system to recover urgent segments that would otherwise be missed when relying on a single source of information.

In contrast, the AND fusion strategy enforces a stricter decision rule that requires both modalities to signal urgency simultaneously. While this approach guarantees perfect precision, it significantly reduces recall, which drops to 0.457 at the audio level and 0.213 at the chunk level. This behavior reflects the fact that many urgent events are detectable in only one modality, making a strict consensus strategy unsuitable for this task. Missed urgency detections (false negatives) are reflected in recall, which is significantly improved through multimodal fusion. In many urgency detection scenarios, it may be preferable to prioritize recall over precision. Indeed, missing a genuine urgent situation may have more severe consequences than raising a small number of false alarms. From this perspective, the OR fusion strategy seems to provide the most appropriate trade-off between reliability and coverage.

5.3.4 RQ4 - perceptual evaluation of translation and urgency

Table 8 reports the results (averaged across all language pairs with a target language other than English) for the qualitative evaluation.

Table 11 Quantitative evaluation per language pair

		WER	BLEU	Consistency
en-fr	non-urgent	8.30%	0.46	1.00
en-de	non-urgent	3.10%	0.34	1.00
en-it	non-urgent	3.30%	0.33	1.00
en-nl	non-urgent	8.30%	0.41	1.00
fr-en	non-urgent	14.40%	0.35	1.00
de-en	non-urgent	7.60%	0.42	1.00
it-en	non-urgent	6.80%	0.42	1.00
nl-en	non-urgent	14.40%	0.49	1.00
en-fr	urgent	9.40%	0.34	1.00
en-de	urgent	9.90%	0.27	1.00
en-it	urgent	9.40%	0.31	1.00
en-nl	urgent	9.80%	0.44	1.00
fr-en	urgent	10.20%	0.28	1.00
de-en	urgent	11.40%	0.35	1.00
it-en	urgent	4.80%	0.35	1.00
nl-en	urgent	17.50%	0.31	1.00

Table 12 Qualitative evaluation per language pair

		MOS translation	MOS urgency
en-fr	non-urgent	4.65	
en-de	non-urgent	2.50	
en-it	non-urgent	3.70	
en-nl	non-urgent	3.90	
en-fr	urgent	4.15	4.40
en-de	urgent	2.15	3.05
en-it	urgent	3.80	2.80
en-nl	urgent	3.40	3.10

The MOS for translation quality in non-urgent audios reaches an average of 3.69 on a five-point scale, indicating that the translations were generally perceived as grammatically and semantically adequate. A comparable score of 3.38 was obtained for urgent audios, suggesting consistent translation quality across both categories. For urgent audios, the MOS for urgency preservation is 3.34, showing that participants were able to reliably perceive the intended tone of urgency in the generated speech. Tables 11 and 12 detail the quantitative and qualitative outcomes for each language pair. Qualitative MOS was collected only for target languages other than English; hence Table 12 reports only the en-X directions. Overall, the results show that system performance remains consistent across different translation directions, with some variability that can be attributed to linguistic complexity, phonetic characteristics, and language model coverage.

5.3.5 RQ5 - cross-language consistency

Looking at Table 11, from a quantitative perspective, the best WER results are observed for translations where the target languages are Italian and German, both below 4% in the non-urgent category, suggesting that these languages benefit from clearer articulation or better ASR model adaptation. Conversely, worse WER results appear in translations involving

Dutch and French, particularly for the fr–en and du–en directions (around 14%), likely due to greater pronunciation variability and less extensive model exposure to those accents. Urgent audios generally exhibit slightly higher WER across all pairs, confirming the impact of faster, more irregular speech patterns on transcription accuracy. The BLEU scores follow a similar trend: they remain moderate overall, with the highest translation quality achieved for en-fr and en-nl pairs, and lower scores for en-de, where syntactic divergence and compound structures introduce additional complexity. Consistency remains 1 across all directions, demonstrating stable model behavior and deterministic pipeline processing.

Looking at Table 12, on the qualitative side, the MOS values mirror the quantitative observations. Translation quality is perceived as adequate across all directions, with the highest scores for en-fr (4.65) and en-it (3.70) in non-urgent cases, reflecting smoother alignment between linguistic structures. Lower MOS translation values for en-de (2.50) and en-nl (3.90) suggest slightly reduced naturalness, possibly due to prosodic or word-order differences. For urgent messages, both *MOS translation* and *MOS urgency* scores remain close to those of non-urgent speech, confirming that the system preserves meaning and tone even under time-constrained conditions. The *MOS urgency* values (ranging between 2.80 and 4.40) indicate that participants could generally perceive the intended urgency, although perception tends to vary with language-specific prosody.

In summary, the per-language results confirm the robustness of the proposed framework across diverse linguistic settings. Minor variations in performance are consistent with expected differences in language morphology, speech rhythm, and translation model coverage, while overall results validate the system's capability to maintain translation accuracy and urgency preservation in a near real-time setting.

These findings provide the basis for a broader discussion of the system's strengths, limitations, and practical implications.

5.3.6 RQ6 - robustness to acoustic noise conditions

To complement the clean-condition evaluation reported in Table 11, we assessed the robustness of the transcription and translation modules under five acoustic degradation conditions: three levels of additive white Gaussian noise (AWGN at 5, 15, and 30 dB SNR), environmental noise, and room impulse response (RIR) reverberation. Results are reported separately for urgent and non-urgent categories across all eight translation directions, expressed as differences relative to the clean-condition baseline in Tables 13 (ΔWER) and 14 (Δ BLEU).

The first observation is that AWGN at higher SNR levels (15 and 30 dB) introduces minimal degradation across most language pairs and both urgency categories, with ΔWER values close to zero or even slightly negative in several non-urgent directions, reflecting the inherent variability of the evaluation rather than a genuine improvement. This confirms that mild additive noise has a limited impact on ASR accuracy. As SNR decreases to 5 dB, degradation becomes more substantial, particularly for urgent speech: the most affected directions are en$\rightarrow$fr, en$\rightarrow$de, and en$\rightarrow$nl, all showing ΔWER exceeding $+0.300$. This suggests that urgent speech, which tends to be faster and rhythmically irregular, is more easily masked by background noise at low signal-to-noise ratios.

The environmental noise condition produces degradation patterns comparable to AWGN at 15 dB across both categories, confirming that the WhisperX ASR module generalizes

Table 13 WER difference relative to clean-condition baseline. Positive values indicate higher error under noise. Bold values indicate the largest degradation per row

Direction	Category	AWGN 5 dB	AWGN 15 dB	AWGN 30 dB	Env. noise	RIR
en→fr	urgent	**+0.344**	+0.155	+0.126	+0.130	+0.136
	non-urgent	-0.042	-0.052	-0.056	-0.063	**-0.030**
en→de	urgent	**+0.339**	+0.151	+0.121	+0.123	+0.131
	non-urgent	+0.010	-0.000	-0.004	-0.011	**+0.022**
en→it	urgent	**+0.203**	+0.019	-0.001	+0.000	+0.003
	non-urgent	+0.008	-0.009	-0.006	-0.013	**+0.023**
en→nl	urgent	**+0.340**	+0.151	+0.122	+0.126	+0.132
	non-urgent	+0.002	-0.005	-0.009	-0.015	**+0.015**
fr→en	urgent	+0.122	+0.119	+0.098	+0.086	**+0.128**
	non-urgent	-0.022	-0.032	-0.027	-0.026	**+0.069**
de→en	urgent	+0.044	-0.013	-0.034	-0.018	**+0.215**
	non-urgent	+0.038	-0.001	+0.014	-0.007	**+0.244**
it→en	urgent	+0.100	+0.111	+0.103	+0.093	**+0.133**
	non-urgent	+0.011	-0.004	+0.036	+0.027	**+0.267**
nl→en	urgent	+0.097	+0.019	-0.016	-0.013	**+0.331**
	non-urgent	+0.091	-0.020	-0.022	-0.014	**+0.301**
Mean	urgent	**+0.198**	+0.089	+0.065	+0.066	+0.151
	non-urgent	+0.012	-0.015	-0.009	-0.015	**+0.114**

Table 14 BLEU variation relative to the clean-condition baseline. Negative values indicate performance degradation under noise. Bold values highlight the largest degradation per row

Direction	Category	AWGN 5 dB	AWGN 15 dB	AWGN 30 dB	Env. noise	RIR
en→fr	urgent	**-0.076**	-0.011	+0.014	-0.014	-0.011
	non-urgent	+0.024	+0.038	+0.028	+0.034	**+0.010**
en→de	urgent	**-0.025**	+0.021	+0.034	+0.018	+0.005
	non-urgent	+0.034	+0.040	+0.017	+0.018	**-0.012**
en→it	urgent	**-0.080**	+0.007	+0.040	+0.024	+0.022
	non-urgent	+0.011	+0.011	+0.021	+0.005	**+0.002**
en→nl	urgent	**-0.120**	-0.032	+0.010	+0.007	+0.025
	non-urgent	+0.019	**-0.016**	+0.028	+0.026	+0.023
fr→en	urgent	**-0.000**	+0.020	+0.026	+0.017	+0.026
	non-urgent	+0.038	+0.074	+0.093	+0.082	**+0.026**
de→en	urgent	-0.071	-0.045	-0.044	-0.054	**-0.180**
	non-urgent	+0.002	+0.021	+0.025	+0.016	**-0.123**
it→en	urgent	**-0.056**	-0.017	-0.040	-0.016	-0.029
	non-urgent	-0.002	+0.004	-0.016	+0.007	**-0.081**
nl→en	urgent	-0.093	-0.032	-0.010	-0.005	**-0.164**
	non-urgent	-0.030	+0.075	+0.023	+0.029	**-0.166**
Mean	urgent	**-0.065**	-0.011	+0.004	-0.003	-0.038
	non-urgent	+0.012	+0.031	+0.027	+0.027	**-0.040**

reasonably well to non-stationary real-world noise. This is a relevant finding in the context of emergency scenarios, where background noise from crowds, machinery, or outdoor environments is common. The impact on BLEU remains limited under both AWGN and environmental noise, suggesting that the translation module is largely robust to the level of transcription errors introduced by these conditions.

The RIR condition represents the most challenging scenario across all language pairs and both urgency categories. ΔWER increases sharply, with the largest degradations observed for nl→en (non-urgent: $+0.301$, urgent: $+0.331$) and de→en (non-urgent: $+0.244$, urgent: $+0.215$). The ΔBLEU under RIR also shows the most pronounced drops, particularly for de→en and nl→en, confirming that reverberation degrades both transcription and translation quality in a correlated manner. This degradation is consistent with well-known limitations of transformer-based ASR systems under heavy reverberation, where early reflections distort the spectral envelope of speech and reduce phoneme discriminability [45].

Overall, the results indicate that the pipeline maintains performance close to the clean-condition baseline under mild to moderate noise, while reverberation represents the primary challenge to both transcription and translation accuracy. These findings motivate future work on noise-robust front-end processing, such as spectral enhancement or de-reverberation, as a preprocessing step upstream of the ASR module.

6 Discussion and limitations

The experimental results demonstrate that the proposed system can operate under near real-time conditions, balancing linguistic accuracy with the preservation of urgency. The average latency, consistently below three seconds, highlights the feasibility of the model for time-critical XR scenarios such as remote assistance or industrial maintenance. In particular, setting the maximum chunk length to five seconds provides an effective compromise between ensuring sufficient context for translation and maintaining acceptable processing times. From a linguistic quality perspective, the system achieved an average WER of 8.28% for non-urgent audios and 10.30% for urgent audios, and average BLEU scores of 0.40 and 0.33, respectively, indicating solid transcription and translation performance across both categories. Although BLEU scores are lower than those of large-scale systems [46], this reflects the intended trade-off between low latency, prosodic preservation, and textual fidelity. This trade-off is consistent with previous research on real-time translation, where it is well established that reducing latency leads to lower BLEU scores and decreased translation quality [47–49]. For instance, Gu et al. [47] demonstrated that simultaneous neural machine translation inevitably sacrifices BLEU score when latency constraints are applied, as models must generate output with limited source context. More recently, Papi et al. [49] provided an in-depth analysis of real-time speech-to-text translation, emphasizing the inherent quality–latency trade-off, that is waiting for more input generally increases translation accuracy but at the cost of higher delay, whereas emitting output earlier reduces latency while risking lower quality. Subjective evaluation through MOS further confirmed that translations were generally perceived as adequate (3.69/5) for non-urgent audios and (3.38/5) for urgent audios and that urgency was clearly conveyed (3.34/5). These findings indicate that, from a user's perspective, the system provides sufficient quality to ensure comprehension and timeliness in critical situations. A main contribution of this study is the demonstration that

preserving urgency does not reduce perceived quality; on the contrary, it contributes to communicative naturalness in time-sensitive scenarios. Compared to general-purpose speech translation systems, which usually ignore prosody, our framework includes specific modules to detect and reproduce urgency. This represents a step forward toward translations that are both expressive and fast. The validity of this approach is supported by converging evidence from both the system design and the experimental results, providing a measurable indication of the contribution of the urgency-aware components in the absence of a direct comparison with a baseline pipeline without urgency preservation. From a technical standpoint, the XTTS-v2 model transfers the prosodic characteristics of a pre-selected urgent reference sample into the synthesized waveform through conditioning mechanisms that operate independently of the translation pipeline, as described in Section 3.3.6. This technical design is further supported by the experimental results: the MOS urgency score of 3.34 confirms that native speakers reliably perceived the intended urgency in the generated speech, while the comparable BLEU scores across urgent and non-urgent conditions confirm that this prosodic adaptation does not affect translation quality. Together, these two lines of evidence indicate the validity of the urgency preservation mechanism.

However, these results are not without limitations that should be considered. First, urgency classification is currently based on a limited set of acoustic features and pre-trained emotion recognition models with fixed thresholds. The use of emotions such as anger and fear indicates a higher probability of urgency, but on their own they may not be sufficient and may miss nuances that are unique to emergency situations. Although robust, this framework may lack flexibility in adapting to variability among speakers or more complex linguistic contexts. In principle, data-driven machine learning models could offer a more adaptive solution. However, the development of such approaches is currently hindered by the lack of sufficiently large and publicly available speech datasets explicitly annotated for urgency. Second, the developed multilingual dataset described in Section 3.1, while innovative in its explicit annotation of urgency, remains relatively small and insufficient for training more advanced models. Furthermore, the use of TTS-generated speech for English recordings may introduce a prosodic bias in urgency modeling, as synthetic voices exhibit more regular pitch and intensity patterns than natural human speech. Third, the qualitative evaluation was conducted with a limited number of participants and should be replicated on larger samples. This limits the possibility of estimating inter-rater agreement and performing statistical significance testing on the MOS scores. Finally, we mention that urgency speech often contains code-mixing under stress conditions. Hence, specific techniques for handling multilingual code-switching scenarios should be integrated in our pipeline.

These limitations suggest several directions for future work. A larger dataset, collected naturally across multiple languages and speaker profiles, would make the results more representative. Training neural models specifically for urgency recognition on large-scale datasets could improve the robustness. Finally, adding objective metrics to assess urgency preservation, alongside human ratings, would make system evaluation more reliable.

In conclusion, the results demonstrate the feasibility of a multilingual speech-to-speech translation framework that preserves urgency in near real-time. This capability distinguishes our system from existing approaches and opens avenues for the development of more effective and reliable communication technologies in time-sensitive operational scenarios.

In light of these considerations, the following section summarizes the main contributions and outlines directions for future work.

7 Conclusion and future work

Language barriers in our multicultural society represent a critical issue that can be addressed through the use of modern technologies. In this study we have demonstrated the feasibility of a multilingual framework for near real-time speech-to-speech translation with urgency detection and preservation. The proposed system extends the cascade paradigm by integrating audio and text-based urgency recognition modules together with an urgency-aware speech synthesis component. The framework has been implemented as an open-source prototype that runs locally and can be integrated into existing conferencing or XR platforms through standard technologies such as WebRTC. Beyond the system itself, this work contributes a multilingual dataset explicitly annotated for urgency. The dataset includes 200 recordings in English, French, German, Dutch, and Italian, equally divided between urgent and non-urgent speech. Although limited in size, it provides a valuable resource for evaluation and establishes a baseline for future work on urgency-aware speech technologies. The second additional contribution is a comprehensive evaluation that combines both quantitative and qualitative analyses. Quantitatively, the system achieved an average end-to-end latency consistently below three seconds, an average WER of 8.28% for non-urgent audios and 10.30% for urgent audios, BLEU scores of 0.40 and 0.33 for non-urgent and urgent audios, respectively, and full consistency across repeated trials. These results confirm the framework's feasibility for time-sensitive scenarios. Qualitatively, native speakers evaluated both translation quality and urgency preservation using the MOS. The system obtained average MOS values of 3.69 for translation in non-urgent audios, 3.38 for translation in urgent audios, and 3.34 for urgency preservation, indicating that the system reliably conveyed the intended urgent tone while maintaining semantic accuracy. Future work will extend the qualitative evaluation by increasing the number of evaluators per language, enabling a more robust and statistically reliable estimation of MOS scores. In addition, future evaluations will include quantitative comparisons with alternative S2ST pipelines without urgency preservation, in order to provide a controlled baseline and to assess the behaviour of the proposed framework against systems that rely on different architectural assumptions, including solutions that are not fully open source or require cloud-based infrastructures. Overall, the findings demonstrate that it is possible to design a speech-to-speech translation system running on a mid-range on-premise server that achieves low latency while preserving prosodic features critical for communication in time-sensitive contexts. Unlike existing commercial and research systems, which prioritize semantic accuracy or style transfer, our framework explicitly addresses urgency as a communicative property. This represents a novel contribution toward expressive and functionally reliable translation technologies.

As future work, we plan to enlarge the urgency dataset with more recordings and a broader pool of native speakers for each language. This extension will allow us to investigate the robustness of the framework across different demographic and linguistic characteristics, as well as across a wider range of emergency scenarios. A larger annotated dataset will also enable the exploration of supervised learning approaches for urgency detection and provide a benchmark for systematic comparison between different modeling strategies, thereby overcoming the limitations associated with the current approach based on static prosodic thresholds. Future work will also investigate more advanced multimodal fusion strategies, including probabilistic calibration and weighted decision mechanisms, to further improve the trade-off between precision and recall in urgency detection. Future work will

also investigate more advanced representation-learning and multimodal fusion approaches, including neural prosodic modeling and self-supervised speech representations, in order to improve the robustness and generalization of urgency detection across speakers and acoustic conditions. These directions are expected to strengthen both the robustness and the generalization of urgency-aware speech-to-speech translation.

Acknowledgements The authors would like to thank all collaborators and technical staff involved in the CORTEX^2 project for their support and constructive feedback during the development and validation of the system.

Author Contributions Simone Pinna and Francesca M. Malloci were primarily responsible for the design and implementation of the software components and conducted the experimental evaluation. Mirko Marras and Daniele Riboni contributed to the definition of the methodological framework and supported the conceptual design of the study. Giuseppe Scarpi provided and maintained the technical infrastructure required for the experimental setup and execution. Diego Reforgiato Recupero supervised the research activity, provided high-level conceptual guidance, and contributed to the overall direction and coordination of the work. All authors contributed to the discussion of the results, reviewed the manuscript, and approved the final version.

Funding Open access funding provided by Università degli Studi di Cagliari within the CRUI-CARE Agreement. This research was conducted within the CORTEX^{2} project, "COoperative Real-Time EXperiences with EXtended Reality", funded by the European Union's Horizon Europe programme under Grant Agreement No. 101070192. In particular, it includes activities carried out under the open call project PETER.

Data and materials availability The data and materials used in this study are available from the corresponding author upon reasonable request.

Declarations

Competing interests The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Trajkovski S, Loosemore M (2006) Safety implications of low-english proficiency among migrant construction site operatives. Int J Project Manage 24(5):446–452
2. Power C, O'Neill K, Ng S-K, Berry E, Grigg M, Luong A, Bloomer MJ et al (2025) Because i couldn't understand and respond: A mixed-method study examining the impact of language barriers on patient experiences of intensive care unit outreach team care. Aust Crit Care 38(3):101198
3. Hudelson P, Chappuis F (2024) Using voice-to-voice machine translation to overcome language barriers in clinical communication: an exploratory study. J Gen Intern Med 39(7):1095–1102
4. Valdez S, Guerberof-Arenas A (2025) Google translate is our best friend here A vignette-based interview study on machine translation use for health communication. Translation Spaces

5. Pinna S, Malloci FM, Marras M, Recupero DR, Riboni D, Scarpi G (2025) Translate, now! Near real-time speech-to-speech translation with urgency preservation. In: Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025). ECAI '25, Bologna, Italy. Paper ID MD41, accepted at ECAI 2025
6. Deschamps-Berger T, Lamel L, Devillers L (2021) End-to-end speech emotion recognition: challenges of real-life emergency call centers data recordings. In: 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), pp 1–8. IEEE
7. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B et al (2005) A database of german emotional speech. In: Interspeech, vol 5, pp 1517–1520
8. Pichora-Fuller MK, Dupuis K (2010) Toronto Emotional Speech Set (TESS). https://doi.org/10.5683/SP2/E8H2MF
9. Livingstone SR, Russo FA (2018) The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5):0196391
10. Busso C, Bulut M, Lee C-C, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) Iemocap: Interactive emotional dyadic motion capture database. Lang Resour Eval 42(4):335–359
11. Landesberger J, Ehrlich U, Minker W (2020) Do the urgent things first! - Detecting urgency in spoken utterances based on acoustic features. In: Adjunct Publication of the 28th ACM conference on user modeling, adaptation and personalization, pp 53–58
12. Kamiyama H, Ando A, Masumura R, Kobashikawa S, Aono Y (2019) Urgent voicemail detection focused on long-term temporal variation. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp 917–921. IEEE
13. Sarim M, Shakeel S, Javed L, Nadeem M et al (2025) Direct speech to speech translation: A review. arXiv preprint arXiv:2503.04799
14. Bentivogli, L, Cettolo M, Gaido M, Karakanta A, Martinelli A, Negri M, Turchi M (2021) Cascade versus direct speech translation: Do the differences still make a difference? arXiv preprint arXiv:2106.01045
15. Sethiya N, Maurya CK (2025) End-to-end speech-to-text translation: A survey. Computer Speech & Language 90:101751
16. Barrault L, Chung Y-A, Meglioli MC, Dale D, Dong N, Duquenne P-A, Elsahar H, Gong H, Heffernan K, Hoffman J et al (2023) SeamlessM4T: massively multilingual & multimodal machine translation. arXiv preprint arXiv:2308.11596
17. Barrault L, Chung Y-A, Meglioli MC, Dale D, Dong N, Duppenthaler M, Duquenne P-A, Ellis B, Elsahar H, Haaheim J et al (2023) Seamless: Multilingual expressive and streaming speech translation. arXiv preprint arXiv:2312.05187
18. Nachmani E, Levkovitch A, Ding Y, Asawaroengchai C, Zen H, Ramanovich MT (2024) Translatotron 3: Speech to speech translation with monolingual data. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 10686–10690. IEEE
19. Themistocleous C (2025) Linguistic and emotional prosody: A systematic review and ale meta-analysis. Neuroscience & Biobehavioral Reviews 175:106210. https://doi.org/10.1016/j.neubiorev.2025.106210
20. Mohamed A-R, Lee H-Y, Borgholt L, Havtorn JD, Edin J, Igel C, Kirchhoff K, Li S-W, Livescu K, Maaløe L, Sainath TN, Watanabe S (2022) Self-supervised speech representation learning: A review. IEEE Journal of Selected Topics in Signal Processing 16:1179–1210
21. Srijith P, Pahwa K, Mamidi R, Surampudi BR, Gupta M, Oota SR (2025) Aligning text/speech representations from multimodal models with MEG brain activity during listening. In: Christodoulopoulos C, Chakraborty T, Rose C, Peng V (eds) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp 34469–34486. Association for Computational Linguistics, Suzhou, China. https://doi.org/10.18653/v1/2025.emnlp-main.1748, https://aclanthology.org/2025.emnlp-main.1748/
22. Nguyen NM, Nguyen TT, Tran P-N, Lim CP, Pham NT, Dang DNM (2026) Multimodal fusion in speech emotion recognition: A comprehensive review of methods and technologies. Eng Appl Artif Intell 163:112624. https://doi.org/10.1016/j.engappai.2025.112624
23. Vlasenko B, Prasad R, Magimai.-Doss M (2021) Fusion of acoustic and linguistic information using supervised autoencoder for improved emotion recognition. In: Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge. MuSe '21, pp 51–59. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3475957.3484448
24. Moon A-S, Kim H, Park Y-C, Lee J (2026) A survey on multimodal emotion recognition: Methods, datasets, and future directions. Computers, Materials and Continua 87(2). https://doi.org/10.32604/cmc.2026.076411
25. Yadav R, Yadav DK, Jain A, Prasad L (2026) Cross-lingual sentiment analysis via multimodal transformer fusion and lightweight deep ensemble learning framework. Knowl Inf Syst 68(1):151. https://doi.org/10.1007/s10115-026-02768-4

26. Sredojev B, Samardzija D, Posarac D (2015) WebRTC technology overview and signaling solution design and implementation. In: 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp 1006–1009. https://doi.org/10.1109/MIPRO.2015.7160422
27. Tesma0704 (2020) Video/audio conferencing using WebRTC. International Journal of Engineering Applied Sciences and Technology 5(4):276–280
28. Rodríguez C, Baez M, Daniel F, Casati F, Trabucco JC, Canali L, Percannella G (2016) Rest APIs: A large-scale analysis of compliance with principles and best practices. In: International conference on web engineering, pp 21–39. Springer
29. Subramanian H, Raj P (2019) Hands-On RESTful API Design Patterns and Best Practices: Design, Develop, and Deploy Highly Adaptable, Scalable, and Secure RESTful Web APIs. Packt Publishing Ltd, Birmingham, UK
30. Hodoshima N (2019) Effects of urgent speech and congruent/incongruent text on speech intelligibility in noise and reverberation. In: INTERSPEECH, pp 3113–3117
31. Kobayashi M, Hamada Y, Akagi M (2022) Acoustic features correlated to perceived urgency in evacuation announcements. Speech Commun 139:22–34
32. Hellier E, Edworthy J, Weedon B, Walters K, Adams A (2002) The perceived urgency of speech warnings: Semantics versus acoustics. Hum Factors 44(1):1–17
33. Jadoul Y, Thompson B, Boer B (2018) Introducing Parselmouth: A Python interface to Praat. J Phon 71:1–15. https://doi.org/10.1016/j.wocn.2018.07.001
34. Cyders MA, Smith GT (2007) The role of negative urgency and positive urgency in hazardous alcohol consumption. Addict Behav 32(12):2657–2666
35. Smith GT, Fischer S, Cyders MA, Annus AM, Spillane NS, McCarthy DM (2007) On the integration of personality structure and process: Emotional experience as a basis for the structure of traits. J Pers Soc Psychol 93(1):130–146
36. Lerner JS, Li Y, Valdesolo P, Kassam KS (2015) Emotion and decision making. Annu Rev Psychol 66(1):799–823
37. Casanova E, Davis K, Gölge E, Göknar G, Gulea I, Hart L, Aljafari A, Meyer J, Morais R, Olayemi S et al (2024) Xtts: a massively multilingual zero-shot text-to-speech model. arXiv preprint arXiv:2406.04904
38. Morris AC, Maier V, Green P (2004) From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In: Interspeech 2004, pp 2765–2768. https://doi.org/10.21437/Interspeech.2004-668
39. Ali A, Renals S (2018) Word error rate estimation for speech recognition: e-WER. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp 20–24 (2018). Association for Computational Linguistics (ACL)
40. Papineni K, Roukos S, Ward T, Zhu W-J (2002) Bleu: a method for automatic evaluation of machine translation. In: Isabelle P, Charniak E, Lin D (eds) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA. https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040/
41. Streijl RC, Winkler S, Hands DS (2016) Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. Multimedia Syst 22(2):213–227
42. Ballier N, Arnold T, Méli A, Thurston T, Yunès J-B (2024) Whisper for L2 speech scoring. Int J Speech Technol 27(4):923–934
43. Chen T, Wang Q, He R, Gollakota S (2025) Spatial speech translation: Translating across space with binaural hearables. In: Proceedings of the 2025 CHI conference on human factors in computing systems, pp 1–19
44. MacFarland TW, Yates JM (2016) Mann–whitney u test. In: Introduction to Nonparametric Statistics for the Biological Sciences Using R, pp 103–132. Springer, New York
45. Goswami M (2026) Whisper-RIR-Mega: A paired clean-reverberant speech benchmark for ASR robustness to room acoustics. arXiv preprint arXiv:2603.02252
46. Zhang S, Feng Y (2021) Universal simultaneous machine translation with mixture-of-experts wait-k policy. arXiv:2109.05238
47. Gu J, Neubig G, Cho K, Li VOK (2017) Learning to translate in real-time with neural machine translation. In: Lapata M, Blunsom P, Koller A (eds) Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp 1053–1062. Association for Computational Linguistics, Valencia, Spain. https://aclanthology.org/E17-1099/
48. Wilmet V, Du J (2024) What does it take to get state of the art in simultaneous speech-to-speech translation?. arXiv preprint arXiv:2409.00965
49. Papi S, Polak P, Macháček D, Bojar O (2025) How real is your real-time simultaneous speech-to-text translation system? Transactions of the Association for Computational Linguistics 13:281–313

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Simone Pinna[1] · Francesca Maridina Malloci[1] · Mirko Marras[1] · Diego Reforgiato Recupero[1,2] · Daniele Riboni[1] · Giuseppe Scarpi[1,2]

✉ Diego Reforgiato Recupero
diego.reforgiato@unica.it

Simone Pinna
simone.pinna2@unica.it

Francesca Maridina Malloci
francescam.malloci@unica.it

Mirko Marras
mirko.marras@unica.it

Daniele Riboni
riboni@unica.it

Giuseppe Scarpi
giuseppe.scarpi@r2msolution.com

[1] Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale, 72, Cagliari, Cagliari 09124, Italy

[2] Innovation Division, R2M Solution s.r.l., Via Fratelli Cuzio, 42, Pavia, Pavia 27100, Italy