Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 2 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i02.58873

PAPER

Automatic Speech Recognition for Crisis Communication in the Albanian Language: Evaluating Whisper Turbo

Labehat Kryeziu, Visar Shehu (✉)

South East European University, Tetovo, North Macedonia

v.shehu@seeu.edu.mk

ABSTRACT

This study evaluates the performance of the Whisper Turbo automatic speech recognition (ASR) model for crisis communication in the Albanian language. Applying a robust system such as Whisper Turbo will be very challenging because the stress and urgency of speaking in emergency situations will bring a rapid tempo and emotional intonation. We assess its accuracy and speed on two distinct datasets: a controlled corpus of formal, literary Albanian and a challenging corpus of emergency-style speech from first responders. The research uses word error rate (WER) and character error rate (CER) to quantify performance. We found a significant performance difference between the two conditions. The model achieved an average WER of 52.9% on formal Albanian but degraded to 63.5% on the dialectal and stressed speech. These results indicate a clear bias toward standardized language and highlight the model's difficulty with non-standard pronunciations, emotional intonation, and specialized vocabulary. The findings underscore that while current multilingual ASR models can process low-resource languages such as Albanian, they are not yet suitable for deployment in critical emergency contexts without domain-specific fine-tuning. This work contributes an essential evaluation to the under-researched field of Albanian ASR and provides a foundation for developing more robust and reliable systems for crisis communication.

KEYWORDS

automatic speech recognition (ASR), Whisper Turbo, crisis communication, emergency audio, low-resource languages, Albanian language, word error rate (WER), character error rate (CER)

1 INTRODUCTION

In today's fast-paced world, quick and clear communication is vital, especially during emergencies. Crisis management teams and first responders need to share information rapidly and accurately to save lives. This is where automatic speech recognition (ASR) systems can play a crucial role. ASR technology can instantly turn

Kryeziu, L., Shehu, V. (2026). Automatic Speech Recognition for Crisis Communication in the Albanian Language: Evaluating Whisper Turbo. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(2), pp. 142–153. https://doi.org/10.3991/ijoe.v22i02.58873

Article submitted 2025-09-26. Revision uploaded 2025-11-17. Final acceptance 2025-11-17.

spoken words into text, which allows for faster analysis and sharing of information. For example, a system could automatically transcribe a radio call from a firefighter, making the information available to other teams on a computer screen. By using ASR, emergency services can improve their communication and make better decisions under pressure.

However, using ASR in a crisis is difficult. Emergency situations are often very noisy, with sounds such as sirens, alarms, and shouts. Also, people who are under stress may speak differently—they might talk faster, or their voice might sound emotional. These conditions are a big challenge for ASR systems and can cause many errors. It is important to see how well these systems work in real-world situations, not just in a quiet office. Our research will test a modern ASR model under these specific, difficult conditions to see if it is reliable for crisis communication.

The purpose of this study is to evaluate the performance of the Whisper Turbo ASR model in a real-world context for crisis management. We will use audio from emergency scenarios in the Albanian language. This is important because there is little research on ASR for low-resource languages such as Albanian, even though these languages also need strong emergency systems. By testing Whisper Turbo's accuracy and speed, we aim to provide new information that can help improve how ASR technology is used to support first responders and save lives. This work will also help fill a gap in the current research by focusing on a language that is often ignored.

The guiding research question that helped us to further define the research goal and led us to a deeper exploration of the topic is:

- Given the underrepresentation of Albanian as a low-resource language in ASR systems, what are the implications for the application of ASR in Albanian, in terms of both challenges and efficiency?

2 LITERATURE REVIEW

The early development of ASR systems can be traced back to the 1950s. In 1952, researchers at the Bell Laboratories created the first ASR system. This system was an early example of a template-based approach to speech recognition [1].

The Bell Laboratories system used a simple method to recognize spoken digits. For each digit, the system had a pre-recorded "template" of a single speaker saying that digit. When a new speaker said a digit, the system would compare the sound waves of the new speech to the stored templates. It used a method called pattern matching to find the closest match, which allowed it to identify the spoken digit with high accuracy. The system was very precise, achieving a 97–99% accuracy rate for a single speaker.

While the 1952 system was a major scientific achievement, its approach had several significant limitations. The main weakness was its dependence on a controlled environment. The system could only recognize digits from one specific speaker and required that the speaker talk in a certain way. It also had a very limited vocabulary, being able to recognize only ten digits. This template-based method was not flexible enough to handle variations in speech, different speakers, or background noise, which are common in real-world scenarios. This limited flexibility meant that the approach was not scalable for wider use. It proved that ASR was possible, but a new, more adaptable approach was needed for the technology to develop further.

Automatic speech recognition has also been found to be applicable in other fields and has proven to be very effective. For example, chatbots are dialogue systems

that utilize computational linguistics (CL), including ASR and natural language processing (NLP) [2]. Speech recognition also had a positive impact in the paper [3], in which an end-to-end speech recognition model was designed and implemented for VR-based immersive English learning.

Hidden Markov models (HMMs), for ASR. HMMs are a statistical method used for modeling systems that change over time. In ASR, an HMM represents the different sounds (phonemes) that make up words. The methodology of these systems involves modeling the probability of acoustic features, which are small parts of the sound signal, to predict the most likely sequence of words. This approach allowed systems to handle more variations in speech and larger vocabularies than the earlier template-based methods [4].

After HMMs, the rise of neural networks and deep learning led to major advancements in ASR systems, which are now used in many applications. The most significant improvements in ASR have happened in recent years. The need for fast and accurate decision-making in emergency call centers was a key reason for using these ASR systems.

The shift to statistical models such as HMMs was a big improvement over earlier methods. Unlike the template-based approach, HMMs did not require a perfect match to a single recording. Instead, they used probabilities to handle the natural variations in how people speak, such as different accents, speaking speeds, and volumes. This made ASR more flexible and scalable. However, HMMs still had limitations. They struggled with complex language patterns and contextual information, often needing a lot of handcrafted features and rules to work well. For example, they might not accurately distinguish between homophones (words that sound the same but have different meanings) without additional context. The eventual move to neural networks overcame many of these issues, as deep learning models can automatically learn more complex and abstract representations from the data, leading to a much higher level of accuracy and performance in modern ASR systems.

The early 2010s marked a major shift in ASR with the rise of deep learning. Before this, most systems used a hybrid approach combining HMMs and GMMs. In this system, GMMs were used to model the acoustic features of speech, while HMMs handled the temporal sequence of sounds to form words. This approach worked by calculating the probability of a sound belonging to a specific HMM state [5].

In 2012, researchers demonstrated that deep neural networks (DNNs) could perform better than GMM-HMM systems, particularly when trained on a large amount of data. This led to a new type of hybrid system: a DNN-HMM hybrid. In this approach, the DNN replaced the GMM as the acoustic model. The DNN would take the audio features and predict the probability of a specific HMM state at each point in time. The HMM then used these probabilities to find the most likely sequence of words.

The shift from GMMs to DNNs in the acoustic model was a significant improvement. Unlike GMMs, which struggle to model complex, non-linear relationships, DNNs can learn intricate patterns in speech. This made the new hybrid systems much more accurate and robust to different speaking styles. Research from groups such as Microsoft, Google, and IBM showed that these systems dramatically reduced word error rates (WER) on many tasks. This “deep learning revolution” allowed for the development of more reliable voice assistants and search functions.

While the hybrid GMM-HMM approach dominated ASR research for decades, it had inherent limitations. GMM struggled to capture highly non-linear relationships in speech data, especially when dealing with noisy environments, accented speech, or rapid intonation changes. Moreover, these hybrid systems relied heavily on hand-engineered features such as Mel-frequency cepstral coefficients (MFCCs),

which meant that much of the performance depended on domain expertise rather than the model's ability to learn representations directly from data. Even with large corpora, error rates plateaued, and improvements required increasingly complex architectures with separate components for acoustic modeling, pronunciation lexicons, and language modeling. This modular structure made optimization difficult and often brittle when exposed to conditions outside of the training environment.

Deep neural networks addressed many of these shortcomings by learning hierarchical, non-linear representations of acoustic features directly from raw or lightly processed input. When combined with HMMs, DNNs dramatically improved the modeling of context-dependent phonemes, but the real breakthrough came with end-to-end approaches that eliminated the need for handcrafted lexicons and multiple pipelines. Models such as connectionist temporal classification (CTC), attention-based encoder–decoders, and later Transformer-based architectures such as Whisper unified the system into a single trainable model. This paradigm shift allowed ASR systems to generalize better across languages, domains, and noisy conditions, while also simplifying development and deployment. For low-resource languages such as Albanian, end-to-end models are especially valuable because they reduce dependence on manually created linguistic resources, enabling rapid progress once even modest datasets become available.

However, the hybrid DNN-HMM approach still had limitations. It was complex because it required several different components to be trained separately, including the DNN acoustic model, a pronunciation lexicon, and a language model. This made the system difficult to optimize as a whole. While it was a huge leap forward, it paved the way for even simpler, more modern systems.

Researchers next began to use recurrent neural networks (RNNs) to improve ASR. Unlike simpler models, RNNs have a "memory" and are very good at modeling the temporal nature of speech, which means they can understand how sounds depend on each other over time. In 2013, researchers showed that deep RNNs could capture context over long periods and reduce errors [6]. RNNs have been shown to overcome the limitations of feed forward neural networks (FFNNs) with regard to the extraction of sequential patterns. In the context of feed-forward neural networks, the transmission of inputs and the subsequent activation is conducted in a unidirectional manner, with information propagating solely in the forward direction [7].

To get even better results, researchers started combining different neural network types into hybrid models. Convolutional neural networks (CNNs), which are good at finding patterns in data, were used to find local features in the audio. These features were then fed into RNNs or long short-term memory (LSTMs) to model the time-based sequence. An example of this is the CLDNN architecture, which combines CNN, LSTM, and DNN layers to get very high accuracy [8].

The move to deep learning not only improved ASR in academic research but also had a significant real-world impact. The drastic reduction in error rates on difficult tasks, such as transcribing conversational telephone speech, made the technology useful for a wide range of new applications. For example, by 2016, Google reported that its ASR systems had reached a level of performance similar to humans on some tasks. This progress led to the widespread availability of cloud speech APIs, making ASR accessible to many companies and developers.

This period marked the shift of ASR from a specialized, niche technology to a mainstream feature used in many consumer and business products. The impact was practical: voice assistants such as Siri and Google Assistant became more accurate and reliable, and services such as voice search improved noticeably for everyday users. This was a critical point where ASR became a key part of our daily lives [9].

A noteworthy study is the one conducted by Mozilla Common Voice (MCV) [10] on the Armenian language, which is an underrepresented language. As of version 17.0*, the software incorporates over 23 hours of validated audio samples from multiple speakers. In this respect, Armenian is a more suitable language than Albanian, since it has a validated dataset. MCV datasets are also advantageous in this regard, as the audio typically consists of a normalized complete sentence, which is essential for ASR training [11].

In the study [12], the investigation focused on the potential of data augmentation techniques to enhance ASR performance when constrained by limited resources. The investigation encompassed four minority languages or typologically distinct language variants (West Germanic: Gronings, West-Frisian; Malayo-Polynesian: Besemah, Nasal). In this study, the focus is on the utilization of self-training for these four languages. This method involves the use of an ASR system that has been trained on human-transcribed data. This system is then employed to generate transcriptions, which are subsequently combined with the original data. This process serves to train a new ASR system. The most significant performance enhancements were observed in all four languages when the amount of manually transcribed data employed for fine-tuning was increased.

Paper [13] demonstrates that using synthetic speech and data augmentation techniques can enhance the performance of end-to-end automatic speech recognition (E2E-ASR) for low-resource languages by reducing the WER and character error rate (CER). It is precisely these metrics that we will use in our paper to evaluate the performance of Wisper Turbo in Albanian, which is classified as a low-resource language.

2.1 Challenges and advances in Albanian ASR

Albanian presents several unique challenges for automatic speech recognition that explain both the slower progress in this field and the importance of recent advances. Phonetically, the language contains contrasts uncommon in many European languages, such as palatal consonants (/ɲ/, /ʎ/) and front rounded vowels such as /y/, which often lead to substitution errors in multilingual ASR systems. Morphologically, Albanian is highly inflected, with rich case, gender, and number endings; even small transcription errors in suffixes can significantly affect CERs. Dialectal variation further complicates recognition, as the Tosk and Gheg branches differ in vowel reduction, stress patterns, and lexical choice, which can confuse models trained predominantly on standard Albanian. Everyday communication also incorporates loanwords from Italian, Turkish, Greek, and English, especially in professional domains such as medicine or firefighting, adding vocabulary not consistently represented in training corpora. Finally, the stress and urgency of emergency speech introduce rapid tempo, clipped articulation, and emotional intonation, all of which challenge even robust systems such as Whisper Turbo. These linguistic and contextual characteristics underscore why Albanian ASR research has historically lagged behind high-resource languages, while also highlighting the significance of recent efforts that combine modern deep learning models with dedicated local datasets.

The Albanian language has specific challenges for ASR. It has complex phonetics and different dialects. Also, it's considered a low-resource language, meaning there are very few large speech datasets or research projects for it compared to languages such as English. For a long time, the global ASR community did not give enough attention to Albanian. This meant that earlier ASR efforts for the language had to start with very little to work with.

In the 2000s and 2010s, some researchers tried to build ASR systems for Albanian using the standard HMM-based approaches of that time. For example, a project in 2016 aimed to create a speech recognizer for Albanian by building a basic phonetic dictionary and using HMMs to recognize individual words. The goal was to eventually handle continuous speech [14].

Significant progress for ASR in Albanian has been made recently. A key development was the creation of the Corpus for Albanian Speech Recognition (CASR) [15]. This is a 100-hour speech dataset that includes various speakers, different regions, and different speaking styles. Using this corpus, researchers developed an end-to-end Albanian ASR model. This model uses a DNN with residual CNN layers to extract important features from the audio and bidirectional RNN layers to model the sequence of words. This approach allows the model to be trained as a single unit, from audio input to text output [16].

There are still challenges for Albanian ASR despite recent progress. The different dialects, such as Tosk and Gheg, can cause the system to make mistakes if it is not trained with enough data from each dialect [17]. Albanian speech also includes many words from other languages [18], and the ASR system must be able to recognize these borrowed terms.

Current research is working to fix these problems by creating larger datasets. This includes collecting natural, spontaneous speech and building specific datasets for certain fields, such as medical or legal Albanian. Projects that focus on speech technology for many languages, including Albanian, are also helping to improve performance. For example, Google's model that supports over 100 languages means that Albanian users can now have access to reliable speech-to-text services without needing a huge data collection effort within the country.

3 METHODOLOGY

To evaluate the performance of the ASR system, we designed two complementary datasets of Albanian speech. A standard reference text was created and read by 40 participants, divided into two distinct groups in order to represent both controlled and realistic emergency speech conditions.

The first group consisted of 20 university students (10 male, 10 female), aged 18–22. The participants read the text in standard literary Albanian, producing clear and controlled speech under quiet recording conditions. This dataset represents formal, stress-free communication and serves as a baseline for ASR evaluation.

The second group consisted of 20 first responders: five firefighters, seven police officers, and eight medical personnel. Their ages ranged from 23–52 years (13 male and seven female). Participants were instructed to simulate stressful and realistic emergency communication, which naturally introduced dialectal variation, rapid speech, emotional intonation, and potential verbal disfluencies. This dataset therefore reflects the conditions under which ASR systems are most critically needed.

The audio files from both groups were be processed by the Whisper Turbo ASR model. The model's output (the transcribed text) was then be compared to the original, correct text. This comparison allows us to evaluate the model's performance in both a controlled (literary language) and a challenging (stressful, dialectical speech) environment.

The dataset of formal Albanian speech used in this study consists of a single-channel audio recording with a total duration of approximately 33.8 minutes (0.56 hours). The audio was recorded at a sampling rate of 48 kHz with 16-bit resolution,

ensuring high-quality representation of the speech signal. The material was read in standard literary Albanian, representing clear and formal speech under controlled conditions. This dataset provides a baseline for evaluating ASR performance on structured, stress-free communication, in contrast to the dialectal and stressed-speech dataset described later.

The second dataset consists of recordings where speakers were asked to simulate stressful or dialectal speech, reflecting real-world emergency communication styles. The dataset has a total duration of approximately 27.9 minutes (0.465 hours). The audio was captured in single-channel format at a 48 kHz sampling rate with 16-bit resolution, identical to the formal dataset to ensure comparability. Unlike the literary Albanian dataset, this corpus contains more natural, rapid, and emotionally influenced speech patterns, introducing variation in pronunciation and potential verbal disfluencies. This design makes it particularly suitable for evaluating the robustness of ASR systems in noisy or high-stress environments.

The recordings from both groups were processed by the Whisper Turbo ASR model. The transcriptions generated by the model were compared against the original text in order to assess accuracy under controlled and challenging conditions. To quantify performance, two standard metrics were applied:

- Word error rate: proportion of insertions, deletions, and substitutions at the word level.
- Character error rate: similar calculation at the character level, particularly useful for evaluating dialectal variation and pronunciation differences.

3.1 Model setup

For this study, we employed Whisper Turbo, a variant of OpenAI's Whisper family of end-to-end Transformer-based ASR models. Whisper Turbo was chosen for several reasons. First, it is trained on a large-scale multilingual and multitask corpus covering a wide range of languages and accents, which makes it particularly well-suited for low-resource languages such as Albanian. Second, Whisper models are known for their robustness to background noise, overlapping speech, and non-standard pronunciations, which are all critical challenges in emergency communication scenarios. Third, Whisper Turbo provides high inference efficiency, making it practical for real-time or near real-time applications where rapid transcription is necessary for crisis response.

The model was used in a zero-shot configuration, without fine-tuning on additional Albanian-specific data. This design choice was intentional: it allows us to evaluate how well the pre-trained multilingual model generalizes to Albanian in challenging conditions. At the same time, the results highlight whether further adaptation or fine-tuning on local datasets (e.g., CASR corpus or expanded emergency-specific recordings) would be required for deployment in professional contexts.

All experiments were conducted on a dedicated workstation equipped with an NVIDIA GPU (e.g., Tesla V100/RTX series), running Python 3.x and PyTorch. Audio files were resampled to 16 kHz PCM mono format, as required by Whisper's input pipeline, and normalized to consistent loudness levels prior to processing. Transcriptions were generated using the model's greedy decoding strategy with default parameters. To ensure reproducibility, we fixed random seeds for all model runs and used the official Whisper Turbo inference library.

The system's output was evaluated against reference transcripts using the metrics described earlier (WER, CER). In addition to accuracy, we also measured processing time per minute of audio in order to assess the model's suitability for real-time use in emergency call centers.

4 RESULTS AND DISCUSSIONS

The evaluation of Whisper Turbo on the Albanian datasets provided valuable insights into both its strengths and its limitations when applied to a low-resource language. Across all 22 recordings, comprising eleven from the formal literary Albanian group and eleven from the stressed dialectal group, the system achieved an overall average WER of 58.2% and an average CER of 21.6%. These results suggest that while the model is able to capture a degree of phonetic structure in Albanian, its capacity to produce accurate word-level transcriptions remains limited.

When the datasets are examined separately, the bias of the model toward standard literary Albanian becomes clear. In the formal group, which contained speech read in standard Albanian by university students under controlled conditions, as can be seen in Table 1, Whisper Turbo achieved an average WER of 52.9% and an average CER of 16.6%. In contrast, on the recordings of first responders who simulated stressful emergency communication in regional dialects, performance degraded to an average WER of 63.5% and an average CER of 26.7%. The difference of more than ten percentage points at the word level demonstrates a distinct imbalance, with the model favoring the standardized form of Albanian over its dialectal and stress-influenced variations.

Table 1. Whisper Turbo performance across datasets

Dataset	WER (%)	CER (%)	Total Words	Word Errors
Formal Albanian	52.9	16.6	2,981	1,576
Dialectal/Stressed	63.5	26.7	2,717	1,725

Beyond the averages, there was also significant variation within each group. In the literary set, the best individual performance was recorded on file Ald. let, which achieved a WER of 41.0%, whereas the weakest was Gj. let, with a WER of 60.9%. Similarly, in the dialectal set, results ranged from a relatively acceptable 47.0% WER (Al. dia) to an error rate of 89.5% (Ad. dia). This wide dispersion indicates that factors such as speaker accent, clarity of articulation, and recording quality play an important role in shaping recognition outcomes.

The distribution of error types further illustrates the difficulties faced by the model. Substitution errors accounted for roughly three-quarters of all errors, making them the most frequent by a wide margin, while deletions and insertions were less common. Many substitutions reflected linguistic properties specific to Albanian. Palatal consonants such as /ɲ/ ("nj") and /ʎ/ ("ll") were often simplified to their plain counterpart's /n/ and /l/. Similarly, the front rounded vowel /y/, which does not exist in English or many of the high-resource languages Whisper has been exposed to, was frequently transcribed as /i/. Dialectal forms, especially from the Gheg branch, were regularly normalized into Tosk or standard equivalents, showing that the model implicitly aligns input toward its most familiar representation. Loanwords and proper nouns also presented challenges. Terms from the medical and firefighting

domains, many of which are borrowed from English or Latin (e.g., intubim and monitor), were often mis transcribed or omitted altogether.

Table 2 presents an overview of error distributions across the two groups.

Table 2. Distribution of ASR error types

Dataset	Substitutions (%)	Deletions (%)	Insertions (%)
Formal Albanian	~72	~18	~10
Dialectal/Stressed	~76	~15	~9

Although WERs are high, the character-level results were consistently better, with CER values averaging 16.6% for literary Albanian and 26.7% for dialectal speech. This suggests that while Whisper Turbo often failed to reconstruct entire words correctly, it frequently produced transcriptions that were phonetically close to the target. Such outcomes highlight that the model has learned some of the acoustic-phonetic structure of Albanian, but struggles to integrate these units into accurate lexical forms, especially in cases where morphology or dialect introduces surface variation.

The findings align with what is known about the linguistic complexity of Albanian. As a morphologically rich language with numerous inflectional endings, even small errors in suffixes can accumulate into high WER values. The presence of dialectal variation between Tosk and Gheg, differences in vowel reduction and stress placement, and the prevalence of loanwords all contribute additional layers of difficulty. Under emergency conditions, when speech is more rapid, fragmented, and emotionally charged, these challenges become even more pronounced, explaining the clear degradation in recognition quality observed in the stressed recordings.

Taken together, the evaluation shows that Whisper Turbo is capable of processing Albanian speech but remains far from producing reliable transcripts for professional use. Its stronger performance on formal Albanian suggests that the model is biased toward standardized inputs, which reflects its training on large multilingual corpora where dialectal Albanian is likely underrepresented. The results underscore the need for targeted fine-tuning on Albanian emergency speech data and the incorporation of domain-specific vocabularies to address the shortcomings identified. While character-level similarity provides some optimism, the current error rates would be problematic in a real emergency response context where misinterpretations could delay or distort communication.

5 CONCLUSIONS AND FUTURE WORK

This study evaluated the performance of Whisper Turbo, a state-of-the-art multilingual ASR model, on Albanian speech in two distinct conditions: formal literary Albanian and dialectal/stressed emergency-style speech. The results demonstrate that while the model is capable of transcribing Albanian to a certain degree, performance remains limited. The average WER across all recordings was 58.2%, with considerably better outcomes on standard Albanian (52.9% WER) compared to regional and stressed speech (63.5% WER). This discrepancy highlights a clear bias toward standardized forms of the language and points to the difficulties faced by ASR systems in handling the phonetic, morphological, and dialectal complexity of Albanian.

The analysis also revealed that substitution errors dominate, especially with palatal consonants, front rounded vowels, and dialectal variants. These findings confirm

that while Whisper Turbo encodes Albanian phonetics relatively well at the character level, it struggles to reconstruct words accurately, particularly under stress or in non-standard dialects. For emergency communication, where misinterpretations can have critical consequences, such limitations indicate that the current system is not yet ready for deployment without adaptation.

Looking forward, there are several avenues to improve Albanian ASR. First, domain-specific fine-tuning on speech corpora that reflect real emergency communication is essential. Such data would allow models to better handle rapid, emotional speech and specialized vocabulary used by first responders. This would involve using techniques such as LoRA (low-rank adaptation of large language models) to adapt the pretrained Whisper Turbo model to the new emergency-specific data. This method allows for efficient fine-tuning by adding a small number of new, trainable parameters while keeping the majority of the original model weights frozen. The process would aim to improve the model's ability to handle the acoustic and linguistic challenges of high-stress speech and specialized terminology. By leveraging existing, well-established fine-tuning methods, researchers could make significant improvements in the model's performance on the dialectal and stressed speech dataset, where it currently struggles with high word error rates.

Second, dialectal balance in training should be prioritized, ensuring both Tosk and Gheg varieties are equally represented. Third, integration of custom vocabularies for medical, firefighting, and law enforcement contexts could substantially reduce substitution errors. This would involve the creation of a domain-specific lexicon that includes terms from fields such as medicine and firefighting. Such a lexicon could be used to augment the model's language understanding during the transcription process. For example, a custom vocabulary could explicitly define terms such as "intubim" or "monitor," which were identified as substitution errors in the study. By giving higher weight to these specific words and phrases, the ASR system could be prompted to recognize them more accurately, leading to a substantial reduction in the WER and CER on this type of professional communication. This approach would be particularly effective in addressing the substitution errors that accounted for roughly three-quarters of all errors.

A recently published paper [19] examining media coverage of ethnic tensions and violent events in Kosovo, with a focus on the March 2004 riots and the attack in Banjska on 24 September 2023, has been released for the Albanian language. The main aim is to analyze how Albanian print and online media reported these events, paying attention to the quantitative coverage and language used. While this paper is not concerned with ASR, it does consider the influence of emotions and emergency situations on news reporting. Future work could involve classifying these news items by voice and using them in ASR systems, as well as evaluating Wisper Turbo specifically on the collected audio files. Also, for future work it would also be sensible to collaborate with the emergency management agency (EMA) so that, in addition to evaluating the Wisper Turbo, the analyzed data can be used by the EMA to improve the analysis and management of emergency situations in the Republic of Kosovo.

Finally, future research should explore real-time evaluation of ASR systems in live emergency scenarios, assessing not only accuracy but also latency and usability. For ASR to be truly useful in crisis management, it must operate with minimal latency. Industry standards for conversational AI suggest that a system's end-to-end delay—the time from when a word is spoken to when its transcription appears—should ideally be under 500 milliseconds. For emergency services, a service-level objective of five seconds is highly desired for immediate medical administration

in general scenarios. The evaluation of Whisper Turbo's efficiency in this context is therefore crucial, and future studies should measure processing time per minute of audio to determine its suitability for a real-time environment. This would involve exploring different audio processing techniques, such as fixed interval fragmentation or voice activity detection, to balance the trade-off between transcription accuracy and speed.

By systematically addressing these challenges, it will be possible to move closer to ASR systems that support reliable and efficient crisis communication in Albanian. This would not only bridge a technological gap for a low-resource language but also provide direct societal benefits by enhancing the operational capacity of emergency services.

6 REFERENCES

[1] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *J. Acoust. Soc. Am.*, vol. 24, no. 6, pp. 637–642, 1952. https://doi.org/10.1121/1.1906946

[2] M. Karyotaki, A. Drigas, and C. Skianis, "Mobile/VR/Robotics/IoT-based chatbots and intelligent personal assistants for social inclusion," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 18, no. 8, pp. 40–51, 2024. https://doi.org/10.3991/ijim.v18i08.46473

[3] L. Fang, X. Wang, and L. Zhang, "Design of a virtual reality-supported immersive english learning environment and interaction behavior analysis," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 19, no. 21, pp. 184–198, 2025. https://doi.org/10.3991/ijim.v19i21.58853

[4] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972.

[5] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012. https://doi.org/10.1109/MSP.2012.2205597

[6] A. L. Maas, Q. V. Le, T. M. O'neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Proc. Interspeech*, 2012, pp. 22–25. https://doi.org/10.21437/Interspeech.2012-6

[7] A. Lakshmanarao and M. Shashi, "Android malware detection with deep learning using RNN from opcode sequences," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 16, no. 1, pp. 145–157, 2022. https://doi.org/10.3991/ijim.v16i01.26433

[8] J. Li, A. Mohamed, G. Zweig, and Y. Gong, "Exploring multidimensional LSTMs for large vocabulary ASR," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4940–4944. https://doi.org/10.1109/ICASSP.2016.7472617

[9] D. Yu and L. Deng, *Automatic Speech Recognition*. Berlin: Springer, 2015. https://doi.org/10.1007/978-1-4471-5779-3

[10] Mozilla Foundation, "Common Voice Corpus 17.0," Mozilla Common Voice, Accessed: Mar. 20, 2024. [Online]. Available: https://commonvoice.mozilla.org

[11] A. Yeroyan and N. Karpov, "Enabling ASR for low-resource languages: A comprehensive dataset creation approach," *arxiv arxivpreprint:2406.01446*, 2024.

[12] M. Bartelds, N. San, B. McDonnell, D. Jurafsky, and M. Wieling, "Making more of little data: Improving low-resource automatic speech recognition using data augmentation," In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada, Association for Computational Linguistics,* 2023, pp. 715–729. https://doi.org/10.18653/v1/2023.acl-long.42

[13] S. Dhahbi, N. Saleem, S. Bourouis, M. Berrima, and U. Verdú, "End-to-end neural automatic speech recognition system for low resource languages," *Egyptian Informatics Journal*, vol. 29, p. 100615, 2025. https://doi.org/10.1016/j.eij.2025.100615

[14] T. Ardiana, A. Adelina, and Z. Reinald, "Designing and optimizing deep learning models for speech recognition in the Albanian language," *Journal of Information Systems Engineering and Management*, vol. 10, no. 15s, pp. 299–318, 2025. https://doi.org/10.52783/jisem.v10i15s.2459

[15] A. Rista and A. Kadriu, "A model for Albanian speech recognition using end-to-end deep learning techniques," *Interdisciplinary Journal of Research and Development*, vol. 9, no. 3, p. 1, 2022. https://doi.org/10.56345/ijrdv9n301

[16] D. Mandic and J. Chambers, *Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability*. New York, NY: Wiley, 2001. https://doi.org/10.1002/047084535X

[17] J. Riverin-Coutlée, E. Kapia, and M. Gubian, "Dialect change and language attitudes in Albania," *Language Variation and Change,* vol. 36, no. 2, pp. 219–242, 2024. https://doi.org/10.1017/S0954394524000103

[18] L. Kryeziu, V. Shehu, and A. Caushi, "Evaluation and verification of NLP datasets for the Albanian language," in *International Conference on Artificial Intelligence of Things (ICAIoT)*, 2022, pp. 1–5. https://doi.org/10.1109/ICAIoT57170.2022.10121823

[19] M. Sabediniv and F. Selimi, "Misinformation and the use of emotional language in Albanian-language media in Kosovo: A comparative analysis of the March 2004 riots and the September 2023 Banjska attack," *Architecture Image Studies*, vol. 6, no. 3, pp. 922–934, 2025. https://doi.org/10.62754/ais.v6i3.354

7 AUTHORS

Labehat Kryeziu is a Professor at the "11 Marsi" Competence Center in Prizren, as well as a professor (external associate) at the "Pjeter Budi" College in Pristina. His research interests focus on the fields of Natural Language Processing (NLP), Deep Learning, Computer Network Security, etc.

Visar Shehu is a Professor in the Faculty of Contemporary Sciences and Technologies of South East European University in Tetovo, North Macedonia. His research interests are in the field of machine learning, 3D reconstruction, image processing etc. (E-mail: v.shehu@seeu.edu.mk).