Acta Infologica
acin 2024, 8 (1), 1–10
DOI: 10.26650/acin.1338604

Research Article / Araştırma Makalesi

Transformer-Based Turkish Automatic Speech Recognition

Davut Emre Taşar[1], Kutan Koruyan[2], Cihan Çılgın[3]

[1]Dokuz Eylül University, Graduate School of Social Sciences, Department of Management Information Systems, İzmir, Türkiye
[2]Dokuz Eylül University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, İzmir, Türkiye
[3]Bolu Abant İzzet Baysal University, Gerede Faculty of Applied Sciences, Department of Management Information Systems, Bolu, Türkiye

Corresponding author : Kutan Koruyan
E-mail : kutan.koruyan@deu.edu.tr

ABSTRACT

Today, businesses use Automatic Speech Recognition (ASR) technology more frequently to increase efficiency and productivity while performing many business functions. Due to the increased prevalence of online meetings in remote working and learning environments after the COVID-19 pandemic, speech recognition systems have seen more frequent utilization, exhibiting the significance of these systems. While English, Spanish or French languages have a lot of labeled data, there is very little labeled data for the Turkish language. This directly affects the accuracy of the ASR system negatively. Therefore, this study utilizes unlabeled audio data by learning general data representations with self-supervised learning end-to-end modeling. This study employed a transformer-based machine learning model with improved performance through transfer learning to convert speech recordings to text. The model adopted within the scope of the study is the Wav2Vec 2.0 architecture, which masks the audio inputs and solves the related task. The XLSR-Wav2Vec 2.0 model was pre-trained on speech data in 53 languages and fine-tuned with the Mozilla Common Voice Turkish data set. According to the empirical results obtained within the scope of the study, a 0.23 word error rate was reached in the test set of the same data set.

Keywords: Wav2vec2, automatic speech recognition, speech-to-text transcription, natural language processing, transformer architecture

Submitted	: 06.08.2023
Revision Requested	: 30.10.2023
Last Revision Received	: 29.11.2023
Accepted	: 30.11.2023
Published Online	: 29.02.2024

This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

Speech is the most basic and efficient form of communication between people (Malik, Malik, Mehmood, & Makhdoom, 2021; Padmanabhan & Johnson Premkumar, 2015). However, today's world is developing from an environment where people only communicate with each other to one where they can communicate with sensors and machines. Therefore, in today's world, especially with the development of artificial intelligence in every field, the need for Automatic Speech Recognition (ASR), which aims to provide human-machine or machine-human interaction, has both increased and developed rapidly in parallel with these developments. Yu & Deng, 2016). Although ASR has been a very interesting area for researchers for a long time, it has become much more suitable for use today, although it does not provide effective results compared to many other input tools (keyboard, mouse, touch screen, etc.) (Yu & Deng, 2016). Seen as a science fiction scenario in the past (Ghai & Singh, 2012), ASRs that can only respond to fluent natural language but still have technological barriers to user satisfaction, as pointed out in 2007 by Benzeghiba et al. (2007) and even by Cutajar, Gatt, Grech, Casha, & Micallef (2013) in 2013, are now a reality. In support of this situation, Malik et al. (2021) today described ASRs as the basic communication tools between humans and machines. Recently, the performance of ASR technologies has almost reached that of human transcribers (Amodei et al., 2016; Chiu et al., 2018; Xiong et al., 2017). Despite these developments, ASR is a multidisciplinary field of study, which requires knowledge from disciplines such as linguistics, computer science, signal processing, acoustics, communication theory, statistics, physiology, and psychology (Levis & Suvorov, 2012).

ASR is the process of turning speech into text after the machine recognizes and understands the speech signal, and includes extraction and determination of acoustic features, and acoustic and language models (Shi, 2021). ASR is a critical research area and has a wide range of applications, such as phone applications (Kurian & Balakrishnan, 2009; Tran, Truong, Le, Huh, & Huh, 2023), toys, games, language translations, educational applications such as language learning (Dai & Wu, 2023; Levis & Suvorov, 2012), air control, security and home automation (Cutajar et al., 2013; Annam, Neelima, Parasa & Chinamuttevi, 2023), customer service (Zekveld, Kramer, Kessens, Vlaming, & Houtgast, 2009; Pragati, Kolli, Jain, Sunethra & Nagarathna, 2023), business management (Danis & Karat, 1995; Xie, 2023), evidence gathering (Negrão & Domingues, 2021; Vásquez-Correa & Álvarez Muniain, 2023), healthcare and virtual assistants (Akhilesh, Brinda, Keerthana, Gupta, & Vekkot, 2022).

Parallel to the development of traditional ASR methods, the applications of these ASRs have also become critical for organizational development. Therefore, end-to-end models (E2E) have started to be developed, and more innovative methods have begun to be developed in this area (Yu & Deng, 2016). Although E2E modeling based on machine learning requires large amounts of labelled data (Jain et al., 2023; Yi, Wang, Cheng, Zhou, & Xu, 2021), it does not need complex modeling processes or a manually designed dictionary by humans (Yi, Wang, Cheng, Zhou, & Xu, 2020). Although machine learning based E2E learning offers significant advantages, it faces problems when it is used for ASR purposes, especially in languages with few resources, due to the need for a high number of labeled training sets. Therefore, for machine learning based E2E learning, it is necessary to pre-train the model in a partially supervised or self-supervised learning (SSL) way (Inaguma, Cho, Baskar, Kawahara, & Watanabe, 2019; Schneider, Baevski, Collobert, & Auli, 2019; Yi et al., 2020). In today's modern approaches to ASR, self-supervised learning is a learning method that does not require human annotated data. It is sometimes considered a form of unsupervised learning. In self-supervised learning, some supervised learning tasks are automatically created from unlabeled data. It aims to recover hidden or missing parts or features of input data given an invisible part of the same input. To solve such tasks, the machine is forced to learn strong representations that convey the meaning or structure of the data. These learned representations are expected to be useful in a variety of downstream tasks, usually after fine-tuning with a few tags.

The majority of the research on the Turkish language has adopted classical methods. However, considering the significant success of the Wav2Vec2 architecture in other languages and the lack of any studies for Turkish with the Wav2Vec2 architecture, the potential benefits of analyzing the study results have constituted the study's point of origin. In this study, we developed a transformer-based Turkish speech-to-text conversion model utilizing an XLSR-Wav2Vec2 model (Cross-Lingual Representation Learning for Speech Recognition) pre-trained with speech data from 53 languages and fine-tuned with data from the Mozilla Common Voice Turkish data set with transfer learning. This study has demonstrated the applicability of self-supervised pre-training techniques to acoustic data and open-source E2E ASR systems developed with the Wav2Vec 2.0 (Baevski, Zhou, Mohamed, & Auli, 2020) architecture by using a large corpus for multiple languages or a single language, and language-specific models were constructed through fine-tuning on the target language. We conducted a performance analysis for the model developed, identified its limitations, and discussed potential areas of usage.

In the following parts of this study, the current literature on ASR is given in Chapter 2, the details of the data set and model used in the study are provided in Chapter 3, the findings are presented in Chapter 4, and the discussion is in the last chapter.

2. LITERATURE REVIEW

Since deep learning techniques, which are frequently used for ASR today, are very data-dependent (Malik et al., 2021), the studies carried out within the scope of the literature are both very diverse and can reveal quite different results. In the literature, the amount of ASR research on the Turkish language is limited. Koruyan (2015) proposed a method for automatic caption generation with Google's Web Speech-to-Text API for live online broadcasts, discussing the effects of the speaker's distance to the microphone, ambient noise, and rapid or continuous speech on the model's performance. Yakar (2016) trained speech recognition models for Turkish using a Hidden Markov Model and analyzed their performance. When analyzed with the word error rate (WER), which is used to evaluate the performance of speech recognition systems, the model achieved a success rate of 17.4% with a limited test data set. In addition, to enhance the model's performance, they combined the phonemes predicted by the model and adopted a post-processing algorithm for Turkish word searches. Oyucu, Polat, & Sever (2020) employed the difference in Hirsch histograms to detect noise in speech, designating a threshold value and using sample frequencies to identify noise and silence. After the removal of noise and silence from the data set, the Kaldi library coded in C++ (Povey et al., 2011) was used to develop a speech recognition system. The speech recognition model trained with this approach had a 7.41% higher WER than the model trained without any pre-processing. In more recent studies, Oyucu & Polat (2023) suggested a Language Model (LM) optimization method for the Turkish language with limited resources. As a result of the findings they obtained, lower WER values were obtained in the ASRs applied with the proposed optimized LMs, and the performance of the ASR was improved. Mussakhojayeva, Dauletbek, Yeshpanov, & Varol (2023), in contrast, developed a multilingual ASR by considering ten languages, not only Turkish but also Azerbaijani, Bashkir, Chuvash, Kazakh, Kyrgyz, Sahaca, Tatar, Uyghur and Uzbek. Tombaloğlu & Erdem (2020) developed a Turkish ASR system based on the Deep Belief Network (DBN). As a result of the empirical findings, they found that the DBN-based ASR system outperformed the traditional Gaussian Mixture Method-based Hidden Markov Model. In addition, they supported the study findings that the DBN-based ASR showed superior results when compared to the existing studies in the literature.

Baevski et al. (2020) demonstrated how effective Wav2Vec 2.0 can be compared to other current approaches in speech recognition with a limited amount of labeled data. Schneider et al. (2019) had success with Wav2Vec with a WER of 2.43%, while Baevski et al. (2020) uses 100 times less labeled data with Wav2Vec 2.0, outperforming Wav2Vec2 in a subset of 100 hours. Liu, Yang, Chi, Hsu, & Lee (2020) presented a new speech representation learning approach in which bidirectional transformer encoders are pre-trained on a large amount of unlabeled speech. Chi et al. (2021) proposed Audio ALBERT, a lightweight version of the self-supervised learning speech representation model. Yi et al. (2020) focused on pre-trained Wav2Vec2 implementation using English speaking for the low-resource ASR task in many languages. According to the empirical findings of the study, a relative improvement of more than 20% was achieved in all six languages compared to previous studies. Pham, Waibel, & Niehues (2022) achieved a 44% improvement over fully supervised learning using Wav2Vec2 with a Common Voice corpus. Coto-Solano et al. (2022) tested Wav2Vec2 and alternative models on Cook Islands Maori, an indigenous language spoken by only about 22,000 people in the South Pacific. Their results show that Wav2Vec2 can yield promising results even for extremely low-resource languages such as Cook Islands Maori. Showrav (2022) demonstrated the success of automatic speech recognition with Wav2Vec2 for the Bengali language, similar to the results of Coto-Solano et al. (2022). Shahgir, Sayeed, & Zaman (2022) achieved better performance with Wav2Vec2 than Showrav (2022) for the Bengali language. Akhilesh et al. (2022) produced a simple and computationally cheaper ASR with Wav2Vec2 for the Tamil language. Olev & Alumae (2022) achieved a 6.9% WER for Estonian with an E2E Wav2Vec2 model. Jain et al. (2023), furthermore, obtained successful results by using Wav2Vec2 in children's speech recognition task, which is more difficult than adults. Wills, Bai, Tejedor-Garcia, Cucchiarini, & Strik (2023) went a step further and used the Wav2Vec2 model for a speech recognition task for non-native Dutch children. Although they found that alternative approaches produced better results, they showed that Wav2Vec2 can also be successful in Dutch non-native language children's speech. Hu et al. (2023) used adapted Wav2Vec2 in the task of automatic recognition of elder speech. In addition to these studies, Vaessen & Van Leeuwen's (2022) empirical findings supported that Wav2Vec2 showed successful results in the speaker recognition task besides the speech recognition task. In addition, the study findings revealed that pre-trained weights used to fine-tune the speech recognition task are also useful for fine-tuning speaker recognition.

As a result of the successful results of many studies, Wav2Vec 2.0 has become one of the most preferred neural-based models for ASR today (Vásquez-Correa et al., 2023). Therefore, in this study, a transformer-based ASR is designed using a fine-tuned XLSR-Wav2Vec2 model with data from the Mozilla Common Voice Turkish dataset.

3. MATERIALS AND METHOD

3.1. Data

The study adopted Mozilla Common Voice as the data set. The Common Voice data set is a multilingual transcribed speech collection for speech technology research and development. Common Voice is designed for ASR algorithm development but can also be utilized in other areas, such as language recognition. The Common Voice project employed crowdsourcing for both data collection and data validation, where voluntary users read out and recorded specific texts. The study adopted version 6.1, which contained 76 languages and was later expanded to encompass 85 languages by November 2021. So far, more than 80,000 volunteers have participated in the development of this data set, producing 13,905 and 11,192 hours of recorded and validated data, respectively. The Turkish language set comprises 44 hours of recorded and 39 hours of validated data. It offers one of the most voice-text matches among open-source databases in the field of speech recognition with respect to the number of both hours and languages. Table 1 shows the features of the data set compiled over time. There is a concentration of the voices of male volunteers and people aged 19-39, which creates a bias where the data set could attain the highest possible accuracy with male voices in this age range. However, the study employed the Common Voice data set as it is the most useful data set due to ease of access, being open-source and the high number of hours.

Table 1. Mozilla Common Voice corpus Turkish data set features (Özden, 2021)

Date	Records			Gender		
21.07.21	Validated Hours	Total Hours	Distinct Voices	Male	Female	N/A
	30	37	960	68%	6%	26%
Age Range						
19-29	30-39	40-49	50-59	60-69	Greater than 70	N/A
47%	17%	2%	4%	1%	0%	26%

The Common Voice project was launched in July 2017 with a focus on the English language, and then in June 2018, it was made available for other languages as well. Common Voice, by its nature, has a sustainable data collection pipeline and the collected data is checked by cross-validation. This control system is performed simply by voting whether the text read by one user has been read correctly by other users. Up to three participants listen to each audio clip. Text readings with at least 2 more than the number of correct votes are archived as unverified data, while other data are archived as verified data. Validated training, validation and testing data sets for each language are obtained using at least 2 positive validation voted data. While creating these data sets, repetitive texts are also removed from these data sets, preventing multiplexing.

3.2. Method

Wav2Vec, which is widely used in E2E ASR models, consists of multiple convolution and attention layers (Baevski, Schneider, & Auli, 2019). Convolution layers take the speech input to the algorithm as a sample and produce more compressed hidden representations, and the attention layers allow a more tailored analysis of the input. Strong context dependency modeling capabilities enable the model to make accurate choices during comparative training via inputs. The Wav2Vec2 training procedure consists of two stages: Initially, the model learns the acoustic representations during self-supervised pre-training by using a significant amount of unlabeled speech data. In the second stage, in supervised fine-tuning, the model is trained on labeled speech data to accurately predict sequences of raw audio graphs or characters.

The Wav2Vec2 model structure is presented in Figure 1. Raw speech waveform $x_i \in$ X is normalized to mean and unit variance, transferred to a feature encoder f: X → Z and converted into hidden representations $z_i \in$ Z. The feature encoders in Wav2Vec2 follow the original Wav2Vec and VQ-Wav2Vec design, however, the activation functions are substituted (Hendrycks & Gimpel, 2016).

The Wav2Vec2 network is transformer-based. In other words, it follows the BERT architecture, except for the changes in positional encoding. Hereby, fixed positional embeddings are replaced with relative positional embeddings to learn relative positional information. To that end, a convolution layer is added to the transformer network, similar to Mohamed, Okhonko, & Zettlemoyer (2019). There are two model configurations with different context network setups: Base and large-base models consist of 12 transformer blocks and have 768 hidden sizes and 8 attention heads, while, in large models, hidden sizes increased to 1,024, with 16 attention heads.

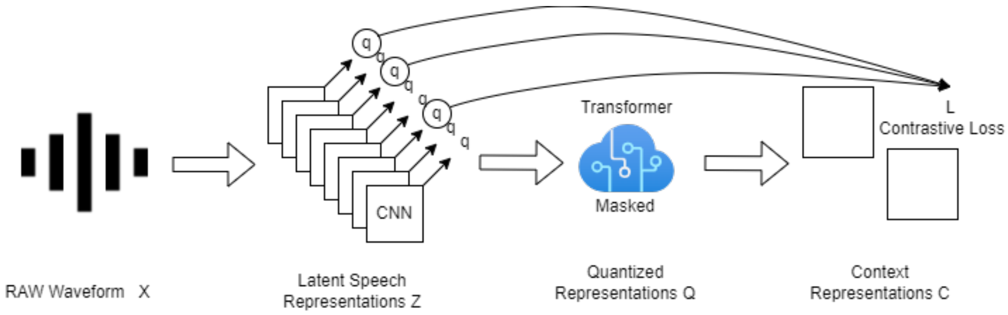
q
q
q
q
q
q
Transformer
L
Contrastive Loss
CNN
Masked
RAW Waveform X
Latent Speech Representations Z
Quantized Representations Q
Context Representations C

Figure 1. Wav2Vec 2.0 architecture (Baevski et al., 2020)

The second training stage for Wav2Vec2, fine-tuning for downstream tasks (ASR), begins with the random initiation of a classifier or the projection of a linear layer to C classes on the transformer network. The classes represent the vocabulary comprising characters and a word limit variable. The classifier is trained on labeled speech data and is optimized with a standard Connectionist Temporal Classification (CTC) loss (Graves, Fernández, Gomez, & Schmidhuber, 2006). Particularly, encoder weights are frozen during fine-tuning and therefore not updated. In addition, the quantification module is deactivated.

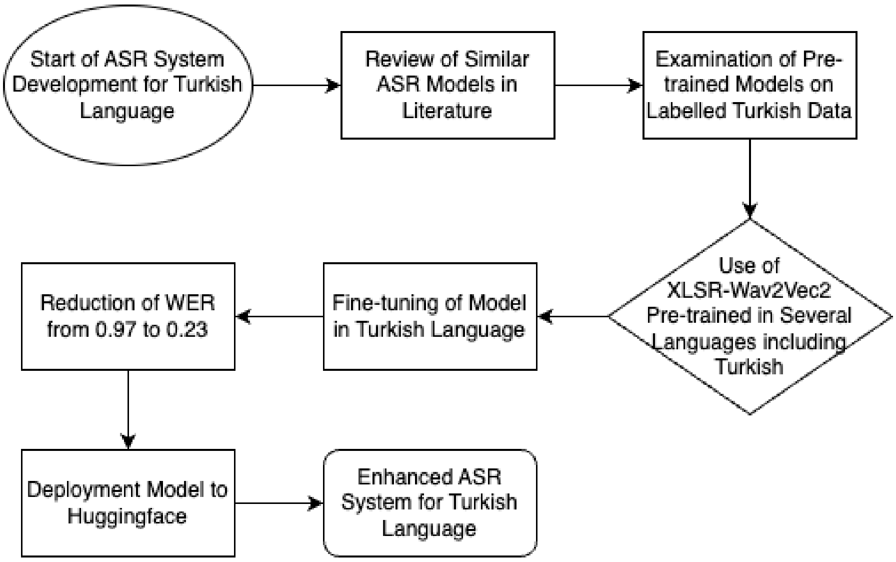
Start of ASR System Development for Turkish Language
Review of Similar ASR Models in Literature
Examination of Pre-trained Models on Labelled Turkish Data
Use of XLSR-Wav2Vec2 Pre-trained in Several Languages including Turkish
Fine-tuning of Model in Turkish Language
Reduction of WER from 0.97 to 0.23
Deployment Model to Huggingface
Enhanced ASR System for Turkish Language

Figure 2. Flow chart of the study

The flow chart in Figure 2 above summarizes the systematic process carried out in the development of the Turkish-specific Automatic Speech Recognition (ASR) system developed in this study. Commencing with a thorough review of extant ASR models in the literature, the research methodically progresses to assess the performance of these models with labeled Turkish data. Central to this endeavor is the employment and fine-tuning of the XLSR-Wav2Vec2 model, which has been pre-trained in multiple languages, including Turkish. This step is pivotal in refining the model's efficacy, as evidenced by the significant reduction in the Word Error Rate (WER) from 0.97 to 0.23. The culmination of this process is an enhanced ASR system, demonstrably more adept at Turkish language recognition, thereby marking a significant advancement in the field of speech recognition technology.

3.3. Word Error Rate

WER is a common measure of the performance of a speech recognition or machine translation system (Klakow & Peters, 2002). WER is a metric used to compute the difference between word-level predicted output and the current output. It can be computed as:

$$WER = (S + D + I) / N,$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of measured words.

4. FINDINGS

ASR models convert speech to text; this means that a feature extractor that processes the speech signal into the model's input format, a feature vector, and a specifier that converts the model's output format to text is needed. For this purpose, a token module called Wav2Vec2 CTC Tokenizer was used to split the data into tokens for the Wav2Vec2 model, and a feature extractor called Wav2Vec2 Feature Extractor was used for feature extraction. The Mozilla Common Voice dataset does not only contain speech data and text equivalents. Other than these two features, all fields (e.g., accent, age, client ID, gender, etc.) were removed from the data set within the scope of this study. Afterwards, the characters of the text data in the data set were determined and arranged. In its final form, our dictionary has 40 unique characters. Then, lowercase conversions were made for all text data. The sampling rate needs to be adjusted so that speech signals can be handled by computers. It is pre-trained on the audio data of the XLSR-Wav2Vec2. Since most of these datasets are sampled at 16 kHz, Common Voice sampled at 48 kHz should be downgraded to 16 kHz for training. Therefore, our fine-tuning data was reduced to 16 kHz. Finally, "*Wav2Vec2Processor*" is used to process the data in the format expected by "*Wav2Vec2ForCTC*" for training. For this, the "*map*" function was used. First, by simply calling "*batch['audio']*", the audio data is loaded and resampled, secondly, "*input_values*" values are extracted from the loaded audio file. This only includes normalization, but for other speech models such as Log-Mel filters, this step may correspond to subtraction.

Python programing language was preferred for the entirety of the application, and programing was conducted in Google Colaboratory (Colab). A Google Colab notebook explaining every programing stage in detail that was kept for experimental reproducibility has been provided as an appendix to this manuscript. In contrast to most Natural Language Processing (NLP) models, XLSR-Wav2Vec2 is an open source, multilingual speech-to-text model with a much larger input length than output length. For instance, the output length of a sample with an input length of 50,000 is no more than 100. Due to the large input size, dynamic padding of training groups was preferred to improve model efficiency, which involves padding all the training samples only with the longest sample. Therefore, fine-tuning XLSR-Wav2Vec2 requires a special padding data collator.

Initially, a data collator must be defined. Contrary to common data collators, this data collator should handle "*input_values*" and "*labels*" differently. Subsequently, the XLSR-Wav2Vec2 checkpoint is loaded, and all training parameters are defined. Here, "*group_by_length*" promotes training efficiency via grouping training samples of similar input length as one batch, which can significantly expedite training by vigorously decreasing the total number of useless padding tokens passing through the model. "*learning_rate*" and "*weight_decay*" are heuristically set up until fine-tuning stabilizes. These parameters can be adopted in the present study as they are mostly dependent on the Common Voice data set. The parameters used for training this model are presented in Table 2.

Table 2. Model training hyperparameters

Hyperparameter	**Selected Value**
learning_rate	0.0005
train_batch_size	2
eval_batch_size	8
seed	42
distributed_type	multi-GPU
num_devices	8
total_train_batch_size	32
total_eval_batch_size	16
optimizer	*Adam with betas* = (0.9,0.999)
epsilon	1e-08
lr_scheduler_type	linear lr
scheduler_warmup_steps	500
num_epochs	30.0
mixed_precision_training	Native AMP

Table 3 shows the performance scores and overall performance charts of the model after training for 30 iterations. The trained model was shared by the authors as “wav2vec-tr-lite-AG” on the website huggingface.co, where the performance of the model can be tested through a web interface if desired. The model’s WER was computed as 0.23 with the test data set. Individual inquiries yielded results that correlated with these rates. However, as discussed in the Data section of the manuscript, the speech recordings used in model training mainly comprised the voices of male volunteers and people aged 19-39, which might cause the model to produce the result given above in a limited framework.

Table 3. Model performance scores

Training Loss	Iterations	Steps	Validation Loss	Word Error Rate
0.439	3.70	400	1.366	0.970
0.377	7.40	800	0.492	0.537
0.230	11.11	1200	0.393	0.413
0.112	14.81	1600	0.327	0.291
0.147	18.51	2000	0.310	0.267
0.101	22.22	2400	0.259	0.232
0.070	25.92	2800	0.287	0.234
0.054	29.63	3200	0.270	0.231

Table 4 presents sample results of the ASR application developed for the Turkish language within the scope of this study. As can be seen, the audio inputs given as input to the application and the text outputs obtained are exemplified. The examples presented here are randomly selected for testing data. The results obtained appear to be quite successful. In addition, the word errors that occur are generally caused by word suffixes, as seen in Table 4.

Table 4. Model sample results of the ASR application developed for the Turkish language

Input (Turkish and English Translation)	Model Transcription
Bugün hava nasıl? (How is the weather today?)	Bugün hava nasıl?
Akşam yemeği için ne düşünüyorsun? (What are you thinking of making for dinner?)	Akşam yemek için ne düşünüyorsun?
Bu kitabı okudun mu? (Have you read this book?)	Bu kitabı okudun mu?
En yakın hastane nerede? (Where is the nearest hospital?)	En yakın hastane nerde?
Çay mı kahve mi tercih edersin? (Do you prefer tea or coffee?)	Çay mı kave mi tercih edersin?
Pazar günü buluşalım mı? (Shall we meet on Sunday?)	Pazar günü buluşalım mi?
Telefonum nerede acaba? (I wonder where my phone is?)	Telefonum nereye acaba?

Table 4 shows how well our ASR system can transcribe Turkish sentences that people might use in everyday life. These sentences are a bit longer to show the system’s ability to handle more than just short phrases. The overall Word Error Rate (WER) of about 0.23 means that most of the time, the system gets the words right, but there are still some small mistakes. For example, it might miss a letter in a word or slightly change a word. These small errors show where we can still make the system better. Even with these mistakes, the table shows that our system does a good job of understanding and writing down Turkish sentences, which is a big step forward for this kind of technology. In addition, all source codes and applications regarding the developed model architecture and the results obtained are presented in the appendix in order to serve to vividly demonstrate the capabilities of the model.

5. DISCUSSION AND CONCLUSION

In the study, an ASR system for Turkish based on a small data set was developed, with Self-Supervised Learning as the primary instrument. Firstly, similar ASR models in the literature were reviewed, the performance of these pre-trained models on labeled Turkish data was examined, and our research was conducted to improve upon the previous results. Considering that ASR systems play an important role in the development of organizations, this study provides a resource for the Turkish language for this need. Especially in recent years, the increase in the use of these systems and the fact that organizations can do the managed work more efficiently and quickly thanks to these systems further emphasizes this importance. In many studies, it has been shown that ASR systems increase business efficiency and speed up business processes in organizations. Researchers such as Filippidou & Moussiades (2020) and Pallett (2003) discussed the increase in work efficiency and the acceleration of work processes with the use of ASR systems. In addition, thanks to ASR systems, organizations can also improve their customer service and customers can easily and quickly solve their problems. As Song et al. (2022) revealed in their study, customer satisfaction increases with the use of ASR systems. As a result, ASR systems play an important role in the development of organizations. Thus, work efficiency increases, and business processes accelerate. Therefore, in parallel with the increase in the use of ASR systems, this study makes significant contributions to the field of ASR for the Turkish language.

The study employed an unsupervised cross-lingual speech representation (XLSR-Wav2Vec2) pre-trained in several languages including Turkish. Subsequently, this pre-trained model was fine-tuned with transfer learning in the Turkish language. Fine-tuning in Turkish was implemented with Wav2Vec 2.0, and the initially higher WER of XLSR-Wav2Vec2 (0.97) was reduced to 0.23 after fine-tuning in the model parameters.

This study showed that Wav2Vec2 models with labeled data outperformed traditional ASR systems. Furthermore, Wav2Vec2 performs an E2E matching of a raw audio file with a sequence of graphs or words, thus eliminating the need to train separate components. However, it must be noted that these models are quite sizable and comprise approximately 317 million parameters, ruling out the possibility of real-time speech-to-text conversion, and require additional hardware, such as a GPU, to ensure an acceptable decoding speed. Therefore, we recommend future studies to adopt larger data sets for the model, reduce the model size or the number of model parameters, or compress Wav2Vec2 models to decrease model training time.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- D.E.T., K.K., C.Ç.; Data Acquisition- D.E.T.; Data Analysis/Interpretation- D.E.T., K.K., C.Ç.; Drafting Manuscript- D.E.T., K.K., C.Ç.; Critical Revision of Manuscript- D.E.T., K.K., C.Ç.; Final Approval and Accountability- D.E.T., K.K., C.Ç.; Supervision- K.K.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: The authors declared that this study has received no financial support.

ORCID IDs of the authors

Davut Emre Taşar	0000-0002-7788-0478
Kutan Koruyan	0000-0002-3115-5676
Cihan Çılgın	0000-0002-8983-118X

REFERENCES

Akhilesh, A., Brinda, P., Keerthana, S., Gupta, D., & Vekkot, S. (2022). Tamil speech recognition using XLSR Wav2Vec2.0 & CTC algorithm. *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1-6. https://doi.org/10.1109/ICCCNT54827.2022.9984422

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Zhu, Z. (2016). *Deep speech 2: End-to-end speech recognition in English and Mandarin. ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning, Volume 48*, 173-182. https://dl.acm.org/doi/10.5555/3045390.3045410

Annam, S. V., Neelima, N., Parasa, N., & Chinamuttevi, D. (2023, March). Automated Home Life using IoT and Speech Recognition. *In 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)* (pp. 809-813). IEEE.

Baevski, A., Schneider, S., & Auli, M. (2019). vq-wav2vec: Self-supervised learning of discrete speech representations. arXiv. https://doi.org/10.48550/arXiv.1910.05453

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems: 34th conference on neural information processing systems (NeurIPS 2020)*, https://proceedings.neurips.cc/paper_files/paper/2020

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., ... & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech communication, 49*(10-11), 763-786. https://doi.org/10.1016/j.specom.2007.02.006

Chi, P. H., Chung, P. H., Wu, T. H., Hsieh, C. C., Chen, Y. H., Li, S. W., & Lee, H. Y. (2021). Audio albert: A lite bert for self-supervised learning of audio representation. *2021 IEEE Spoken Language Technology Workshop (SLT)*, 344-350. https://doi.org/10.1109/SLT48900.2021.9383575

Chiu, C. C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., ... & Bacchiani, M. (2018). State-of-the-art speech recognition with sequence-to-sequence models. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4774-4778. https://doi.org/10.1109/ICASSP.2018.8462105

Coto-Solano, R., Nicholas, S. A., Datta, S., Quint, V., Wills, P., Powell, E. N., ... & Feldman, I. (2022). Development of automatic speech recognition for the documentation of Cook Islands Māori. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3872–3882. https://aclanthology.org/volumes/2022.lrec-1/

Cutajar, M., Gatt, E., Grech, I., Casha, O., & Micallef, J. (2013). Comparative study of automatic speech recognition techniques. *IET Signal Processing, 7*(1), 25-46. https://doi.org/10.1049/iet-spr.2012.0151

Danis, C., & Karat, J. (1995). Technology-driven design of speech recognition systems. D*IS '95: Proceedings of the 1st conference on Designing interactive systems: processes, practices, methods, & techniques*, 17-24. https://doi.org/10.1145/225434.225437

Dai, Y., & Wu, Z. (2023). Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study. *Computer Assisted Language Learning*, 36(5-6), 861-884.

Filippidou, F., & Moussiades, L. (2020). A benchmarking of IBM, Google and Wit automatic speech recognition systems. *IFIP Advances in Information and Communication Technology*, 73-82. https://doi.org/10.1007/978-3-030-49161-1_7

Ghai, W., & Singh, N. (2012). Literature review on automatic speech recognition. *International Journal of Computer Applications, 41*(8), 42-50. http://dx.doi.org/10.5120/5565-7646

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd international conference on Machine learning - ICML '06,* 369-376. http://dx.doi.org/10.1145/1143844.1143891

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). arXiv. https://doi.org/10.48550/arXiv.1606.08415

Hu, S., Xie, X., Jin, Z., Geng, M., Wang, Y., Cui, M., ... & Meng, H. (2023). Exploring self-supervised pre-trained ASR models for dysarthric and elderly speech recognition. ICASSP 2023 - 2023 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1-5. https://doi.org/10.1109/ICASSP49357.2023.10097275

Inaguma, H., Cho, J., Baskar, M. K., Kawahara, T., & Watanabe, S. (2019). Transfer learning of language-independent end-to-end ASR with language model fusion. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6096-6100). https://doi.org/10.1109/ICASSP.2019.8682918

Jain, R., Barcovschi, A., Yiwere, M., Bigioi, D., Corcoran, P., & Cucu, H. (2023). A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition. *IEEE Access, 11*, 46938-46948. https://doi.org/10.1109/ACCESS.2023.3275106

Klakow, D., & Peters, J. (2002). Testing the correlation of word error rate and perplexity. *Speech Communication, 38*(1-2), 19-28. https://doi.org/10.1016/S0167-6393(01)00041-3

Koruyan, K. (2015). Canlı internet yayınları için otomatik konuşma tanıma tekniği kullanılarak alt yazı oluşturulması [Generating captions using automatic speech recognition technique for live webcasts]. *Bilişim Teknolojileri Dergisi, 8*(2), 111-116. https://doi.org/10.17671/btd.31441

Kurian, C., & Balakrishnan, K. (2009). Speech recognition of Malayalam numbers. *2009 World Congress on Nature & Biologically Inspired Computing (NaBIC*), 1475-1479. https://doi.org/10.1109/NABIC.2009.5393692

Levis, J., & Suvorov, R. (2012). Automatic speech recognition. In *The encyclopedia of applied linguistics.* Retrieved from https://onlinelibrary.wiley.com

Liu, A. T., Yang, S. W., Chi, P. H., Hsu, P. C., & Lee, H. Y. (2020). Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. *ICASSP 2020 - 2020 IEEE International Conference* on Acoustics, Speech and Signal Processing (ICASSP), 6419-6423. https://doi.org/10.1109/ICASSP40776.2020.9054458

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: A survey. *Multimedia Tools and Applications, 80*, 9411-9457. https://doi.org/10.1007/s11042-020-10073-7

Mohamed, A., Okhonko, D., & Zettlemoyer, L. (2019). Transformers with convolutional context for ASR. arXiv. https://doi.org/10.48550/arXiv.1904.11660

Mussakhojayeva, S., Dauletbek, K., Yeshpanov, R., & Varol, H. A. (2023). Multilingual speech recognition for Turkic languages. *Information, 14*(2), 74. https://doi.org/10.3390/info14020074

Negrão, M., & Domingues, P. (2021). SpeechToText: An open-source software for automatic detection and transcription of voice recordings in digital forensics. Forensic Science International: *Digital Investigation, 38*, 301223. https://doi.org/10.1016/j.fsidi.2021.301223

Olev, A., & Alumae, T. (2022). Estonian speech recognition and transcription editing service. *Baltic Journal of Modern Computing, 10*(3), 409-421. https://doi.org/10.22364/bjmc.2022.10.3.14

Oyucu, S., & Polat, H. (2023). A language model optimization method for Turkish automatic speech recognition system. *Politeknik Dergisi*, (Early Access). https://doi.org/10.2339/politeknik.1085512

Oyucu, S., Polat, H., & Sever, H. (2020). Sessizliğin kaldırılması ve konuşmanın parçalara ayrılması işleminin Türkçe otomatik konuşma tanıma üzerindeki etkisi [The effect of removal the silence and speech parsing processes on Turkish automatic speech recognition]. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 8*(1), 334-346. https://doi.org/10.29130/dubited.560135

Özden, B. (2021, September 14). Common voice Türkçe'nin durumu [Web blog post]. Retrieved from https://discourse.mozilla.org/t/common-voice-turkcenin-durumu/85895

Padmanabhan, J., & Johnson Premkumar, M. J. (2015). Machine learning in automatic speech recognition: A survey. *IETE Technical Review, 32*(4), 240-251. https://doi.org/10.1080/02564602.2015.1010611

Pallett, D. S. (2003). A look at NIST's benchmark ASR tests: Past, present, and future. 2003 *IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*, 483-488. https://doi.org/10.1109/ASRU.2003.1318488

Pham, N. Q., Waibel, A., & Niehues, J. (2022). Adaptive multilingual speech recognition with pretrained models. arXiv. https://doi.org/10.48550/arXiv.2205.12304

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., . . . Vesely, K. (2011). The Kaldi speech recognition toolkit. *IEEE 2011 workshop on automatic speech recognition and understanding*, https://www.fit.vut.cz/research/publication/11196/.en

Pragati, B., Kolli, C., Jain, D., Sunethra, A. V., & Nagarathna, N. (2023, January). Evaluation of Customer Care Executives Using Speech

Emotion Recognition. *In Machine Learning, Image Processing, Network Security and Data Sciences: Select Proceedings of 3rd International Conference on MIND 2021* (pp. 187-198). Singapore: Springer Nature Singapore.

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. arXiv. https://doi.org/10.48550/arXiv.1904.05862

Shahgir, H. A. Z. S., Sayeed, K. S., & Zaman, T. A. (2022). Applying wav2vec2 for speech recognition on Bengali common voices dataset. arXiv. https://doi.org/10.48550/arXiv.2209.06581

Shi, Z. (2021). *Intelligence science: Leading the age of intelligence.* Elsevier.

Showrav, T. T. (2022). An automatic speech recognition system for Bengali language based on wav2vec2 and transfer learning. arXiv. https://doi.org/10.48550/arXiv.2209.08119

Song, Y., Lian, R., Chen, Y., Jiang, D., Zhao, X., Tan, C., ... & Wong, R. C. W. (2022). A platform for deploying the TFE ecosystem of automatic speech recognition. *Proceedings of the 30th ACM International Conference on Multimedia*, 6952-6954. https://doi.org/10.1145/3503161.3547731

Tombaloğlu, B., & Erdem, H. (2020). Deep learning based automatic speech recognition for Turkish. *Sakarya University Journal of Science, 24*(4), 725-739. https://doi.org/10.16984/saufenbilder.711888

Tran, D. T., Truong, D. H., Le, H. S., & Huh, J. H. (2023). Mobile robot: automatic speech recognition application for automation and STEM education. *Soft Computing,* 1-17.

Vaessen, N., & Van Leeuwen, D. A. (2022). Fine-tuning wav2vec2 for speaker recognition. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7967-7971. https://doi.org/10.1109/ICASSP43922.2022.9746952

Vásquez-Correa, J. C., & Álvarez Muniain, A. (2023). Novel speech recognition systems applied to forensics within child exploitation: Wav2vec2. 0 vs. whisper. *Sensors, 23*(4), 1843. https://doi.org/10.3390/s23041843

Wills, S., Bai, Y., Tejedor-Garcia, C., Cucchiarini, C., & Strik, H. (2023). Automatic speech recognition of non-native child speech for language learning applications. arXiv. https://doi.org/10.48550/arXiv.2306.16710

Xie, T. (2023). Artificial intelligence and automatic recognition application in B2C e-commerce platform consumer behavior recognition. *Soft Computing, 27*(11), 7627-7637.

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The Microsoft 2017 conversational speech recognition system. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP).* 5934-5938. https://doi.org/10.1109/ICASSP.2018.8461870

Yakar, Ö. (2016). *Sözcük ve hece tabanlı konuşma tanıma sistemlerinin karşılaştırılması* (Master's thesis). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2020). Applying wav2vec2.0 to speech recognition in various low-resource languages. arXiv. https://doi.org/10.48550/arXiv.2012.12121

Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2021). Transfer ability of monolingual wav2vec2.0 for low-resource speech recognition. *2021 International Joint Conference on Neural Networks* (IJCNN), 1-6. https://doi.org/10.1109/IJCNN52387.2021.9533587

Yu, D., & Deng, L. (2016). *Automatic speech recognition (Vol. 1).* Berlin: Springer.

Zekveld, A. A., Kramer, S. E., Kessens, J. M., Vlaming, M. S., & Houtgast, T. (2009). The influence of age, hearing, and working memory on the speech comprehension benefit derived from an automatic speech recognition system. *Ear and Hearing, 30*(2), 262-272. https://doi.org/10.1097/aud.0b013e3181987063

How cite this article

Taşar, D. E., Koruyan, K., & Çılgın, C. (2024). Transformer-based Turkish automatic speech recognition. *Acta Infologica, 8*(1), 1-10. https://doi.org/10.26650/acin.1338604

Appendix:

Notebook URL: https://colab.research.google.com/drive/1xcfnBFdtT6HcW6MDLgImj7tJcYafJSJU