Araştırma Makalesi
Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi
16 (2022) 156-164

DEVELOPMENT OF TEST CORPUS WITH LARGE VOCABULARY FOR TURKISH SPEECH RECOGNITION SYSTEM AND A NEW TEST PROCEDURE

Saadin OYUCU*

[1]Adıyaman University, Faculty of Engineering, Department of Computer Engineering, Adıyaman, 02040, Turkey
Geliş Tarihi/Received Date: 20.12.2021 Kabul Tarihi/Accepted Date: 11.03.2022 DOI: 10.54365/adyumbd.1038766

ABSTRACT

The most fundamental problem in the automatic speech recognition systems is not the development of a domain-specific automatic speech recognition system, but the development of an automatic speech recognition system with a large vocabulary. Developed automatic speech recognition systems should be tested with a large vocabulary test dataset. For this reason, an automatic speech recognition test corpus was prepared within the scope of the study. Prepared automatic speech recognition test corpus includes conversations from 20 different areas and text files of these conversations. The test procedure presented in the study was also tested on Turkish automatic speech recognition systems with a large vocabulary. It has been observed that the word error rate results ranged between 14-21%. The test corpus and test procedure with a large vocabulary prepared are guiding for the success of automatic speech recognition systems in future studies to be revealed more clearly.

Keywords: *Speech recognition, Turkish speech recognition, speech corpus, test corpus, Turkish speech corpus*

TÜRKÇE KONUŞMA TANIMA SİSTEMİ İÇİN GENİŞ KELİME DAĞARCIĞINA SAHİP TEST VERİ KÜMESİNİN GELİŞTİRİLMESİ VE YENİ BİR TEST PROSEDÜRÜ

ÖZET

Otomatik konuşma tanıma sistemlerindeki en temel sorun, alana özgü bir otomatik konuşma tanıma sisteminin geliştirilmesi değil, geniş kelime dağarcığına sahip bir otomatik konuşma tanıma sisteminin geliştirilmesidir. Geniş kelime dağarcığına sahip olacak şekilde geliştirilen otomatik konuşma tanıma sistemleri, geniş kelime dağarcığına sahip bir test veri kümesi ile test edilmelidir. Bu nedenle çalışma kapsamında bir otomatik konuşma tanıma test veri kümesi hazırlanmıştır. Hazırlanan otomatik konuşma tanıma test veri kümesi, 20 farklı alandan konuşmaları ve bu konuşmalara karşılık gelen metin dosyalarını içermektedir. Çalışma kapsamında sunulan test prosedürü, geniş kelime dağarcığına sahip farklı Türkçe otomatik konuşma tanıma sistemleri üzerinde de test edilmiştir. Elde edilen kelime hata oranı sonuçlarının %14-21 arasında değişkenlik gösterdiği görülmüştür. Geniş kelime dağarcığına sahip olacak şekilde hazırlanan test veri kümesi ve test prosedürü, ilerideki çalışmalarda otomatik konuşma tanıma sistemlerinin başarısının daha net ortaya konması için yol göstericidir.

Anahtar Kelimeler: *Konuşma tanıma, Türkçe konuşma tanıma, Konuşma veri seti, Türkçe konuşma veri seti, Test veri seti*

1. Introduction

Systems that convert human-spoken expressions into computer-readable text are called Automatic Speech Recognition (ASR) systems [1]. Many new applications have been developed with speech information converted into a format that can be understood by computers or electronic devices. ASR systems; voice command applications, smart home systems, security systems, education systems, call

* e-posta: saadinoyucu@adiyaman.edu.tr ORCID ID: https://orcid.org/0000-0003-3880-3039

centers, conference systems, automatic speech reporting, dictation software, etc. used in many fields or applications. The expected success from ASR systems, the usage area of which is increasing day by day, is to be able to understand speech with human ear sensitivity and transfer it to text. To achieve this expected success from ASR systems, it is necessary to use different disciplines together. However, this situation complicates the general structure of ASR systems and complicates the development processes. For this reason, although intensive studies have been carried out on ASR, the desired level of success has not been reached yet. In a classical ASR architecture, there are important components such as Feature Extraction, Decoder, AM Acoustic Model, LM: Language Model and Lexicon. The combination of these components increases the complexity of ASR systems but has a great impact on success rates [2]. Speech and text data are used to create AM, LM and Lexicon required for ASR systems. These data must have some features to be used in AM and LM training.

The data must match speech and text equivalents, and records must be obtained from multiple speakers and speakers of different genders. Preparing the required corpus for ASR brings different times, costs and difficulties for different languages. Turkish is a language with rich morphology. The productive morphology of Turkish allows the formation of many unique word forms. For this reason, Turkish is among the languages with a large vocabulary. The large vocabulary increases the size of the recognizable word set required for speech recognition. With the increase in the size of the recognizable word set required for speech recognition, two main problems arise. The first of these problems is that the corpus to be used in the training process must contain a lot of speech and text data. Acoustic information required for AM can be obtained from a large training corpus. Another problem is the modeling of agglutinative, long sentences, and productive language structure.

In the literature, it has been seen that many studies have been done on Turkish ASR. Studies on AM have attempted to develop systems that are not affected by the speaker and accent changes [3,4]. In a different study, models containing more than one accent sample were developed for the training of the acoustic model [5]. However, it is not enough to include more than one accent sample in the training corpus. For this reason, improvements have been made on acoustic modeling within the scope of the study [5]. Other studies in the literature have been carried out on language modeling. Recent work has focused on attention-based model development and the development of language model integration for ASR systems using decoders. Zeineldine at. Al. investigated methods for integrating an external language model trained on unpaired text data into ASR. They proposed a method to predict an implicit internal language model directly from the attention-based encoder-decoder model [6]. Gandhi and Rastrow proposed a system that learns to reassess the output of the ASR system. End-to-end approaches are combined with a traditional structure using an attention-based distinctive language model [7]. These studies have increased the overall success of ASR with the effect of the language model. However, most research on the language model has focused on the effect of LM size on ASR [8-12]. The results revealed the necessity of large-scale language models, especially for agglutinative languages such as Turkish. For a large-sized language model, a large amount of text data is needed.

When the related studies are examined, it is seen that there is a deep relationship between the models used in the creation of ASR systems and the dataset used in the development of these models. Dataset is an important problem in studies on low-resource languages. Speech and text data required for ASR are not available for low-resource languages such as Turkish, apart from a few academic studies. In the development and testing of Turkish ASR systems, METU 1.0 sound corpus, prepared by Bogaziçi University in 2012 [13] and the METU corpus provided by METU, was generally used [14]. The vocabulary of these datasets is insufficient for Turkish. Therefore, an ASR system with a large vocabulary cannot be developed. In addition, testing processes of ASR systems with a large vocabulary cannot be performed successfully. A piece of test data separated from the training dataset is used in the testing of ASR systems. Therefore, an ASR system with a large vocabulary has not been developed. In addition, testing process of ASR systems with a large vocabulary could not have been performed successfully. For example, the dataset prepared by Bogaziçi University only includes news speeches. In this case, the ASR system developed with a dataset containing only news conversations, will not be able to transcribe the conversations in a field such as sports or technology.

In the literature, different studies have been carried out to solve the problem of a large vocabulary. First of all, the creation of a corpus with a large vocabulary was studied and ASR systems with a large vocabulary were developed [15,16]. However, a balanced Turkish dataset of spontaneous conversations and conversations in different fields is not currently available. In addition to academic studies, technology companies such as Google, Amazon and Microsoft have also carried out studies on the problem of large vocabulary in ASR systems. Google, Amazon and Microsoft companies state that they offer their ASR service with a large vocabulary. However, there is currently no test corpus to be used in the large vocabulary tests of Turkish ASR systems, both developed in academic studies and offered as a service by technology companies. Therefore, within the scope of the study, first of all, a test corpus containing conversations from different fields that can be used in large vocabulary testing of Turkish ASR systems was prepared.

The Turkish ASR test corpus prepared is related to science, technology, economy, etc. It includes conversations from 20 different areas and text files of these conversations. For large vocabulary tests, a Word Error Rate (WER) based test procedure has been prepared. This test procedure has been tested on Google Speech to Text, Amazon Transcribe and Azure Speech to Text services. The prepared test procedure was also tested on the Turkish ASR system, which has a large vocabulary presented by Polat and Oyucu in 2020 [15]. When the WER results obtained within the scope of the study were examined, it was observed that the word error rates varied between 14 and 21%. In addition, it has been observed that different ASR services have their advantages and disadvantages. The test dataset and test procedure with a large vocabulary will guide the success of ASR systems in future studies.

2. Preparation of The Turkish ASR Test Corpus

In speech recognition systems developed specifically for the field, the vocabulary of spoken words belongs to a certain set. However, it is not possible to carry out spontaneous conversations with words belonging to a certain topic. Therefore, within the scope of this study, a test corpus was prepared to test Turkish ASR systems with a large vocabulary.

The vocabulary of words used in real life is quite large. Besides, the number of words spoken in rich languages such as Turkish is very high. In addition, conversations in daily life consist of different domains. For example, sports, politics, or science represent a few of these domains. For this reason, a test corpus consisting of different domains has been prepared to test the academic studies in the literature and Turkish ASR systems with a large vocabulary already developed by technology companies. However, it is not enough to just create the test corpus from different domains. Different speakers from different domains should be included in the corpus. For this reason, data were obtained from speakers of different genders in different domains. The domains and speaker information determined for the test corpus prepared are given in Table 1.

As seen in Table 1, 20 different domains spoken in daily life in Turkish were added to the test corpus. The parliamentary speeches, which are one of these domains, were obtained from the minutes of the Turkish Grand National Assembly. Conversations about other fields were selected on YouTube. Since there are text equivalents in the minutes of the parliamentary speeches, the transcripts were used in the speech-text matching. However, the text equivalent of the speeches received on YouTube was obtained by real users by listening to the speech recording. In addition, by keeping the gender distribution balanced, the test corpus was ensured to be more effective. For this reason, speech recordings were obtained from a total of 286 different speakers, 143 male and 143 female speakers. The information about the speeches made by different speakers according to a different domain is given in Table 2.

Table 1. Speaker information by domain

Domain Name	Speaking Time / Seconds	Number of Male Speakers	Number of Female Speakers
Science	335	5	5
Education	327	5	5
Economy	703	9	9
Philosophy	325	5	5
Physics	318	5	5
News	325	5	5
Weather	397	5	5
Law	287	5	5
Business	393	5	5
Culture	1.067	10	10
Magazine	960	12	12
Math	301	5	5
Parliament	2.088	19	19
Humor	227	5	5
Health	838	11	11
Art	805	10	10
Sociology	305	5	5
Sport	425	6	6
Agriculture	355	5	5
Technology	431	6	6
Total	11.212	143	143

The contributions of different speakers, who gave speeches in different domains, to the test corpus according to their gender are shown in Table 2. However, just the number of speakers or speaking time is not enough to test an ASR system with a large vocabulary. The number of words in the conversations is as important as the speaking time. For this reason, the vocabulary of the speeches in the test corpus is given in Table 3. Vocabulary can be easily tested with unique words. Gender difference will reveal accent differences. The gender difference and the high number of speakers in the dataset will enable to encounter more discourse examples in the test set. Besides, the higher the gender and the number of speakers in the training dataset, the more robust the acoustic model will be.

Table 2. Speaker durations by domain

Domain Name	Total Speaking Time of Male Speakers / Seconds	Total Speaking Time of Female Speakers / Seconds	Total Speaking Time / Seconds
Science	184	151	335
Education	160	167	327
Economy	388	315	703
Philosophy	147	178	325
Physics	170	148	318
News	191	134	325
Weather	179	218	397
Law	132	155	287
Business	212	181	393
Culture	667	400	1.067

Magazine	477	483	960
Math	167	134	301
Parliament	1.116	972	2.088
Humor	119	108	227
Health	442	396	838
Art	394	411	805
Sociology	171	134	305
Sport	225	200	425
Agriculture	208	147	355
Technology	250	181	431
Total	5.999	5.213	11.212

Table 2. Continue

Table 3. Number of words by domain

Domain Name	Number of Words	Number of Unique Word
Science	633	424
Education	632	414
Economy	1576	907
Philosophy	651	424
Physics	673	422
News	587	436
Weather	898	476
Law	520	381
Business	789	505
Culture	1440	927
Magazine	1711	970
Math	712	383
Parliament	3.723	1.756
Humor	473	324
Health	1354	832
Art	1470	845
Sociology	573	409
Sport	892	568
Agriculture	694	466
Technology	894	579
Total	20.895	12.448

In Table 3, the number of words in the speeches is given. The test corpus prepared according to the determined features was used in the testing of different Turkish ASR systems with a large vocabulary. Thus, comparative experimental results were obtained on Turkish ASR studies and services with a large vocabulary.

3. Test Procedure for ASR Systems

In this study, firstly, a test corpus was prepared to test Turkish ASR systems with a large vocabulary. After the test corpus was obtained, an experimental procedure was prepared to carry out large vocabulary experiments. For this procedure, the algorithm that performs the WER calculation was coded using the Java language and an experimental procedure application was developed. The developed

application compares the reference text with the text in the ASR output. In addition to the work of Oyucu and Polat, Google Speech to Text, Amazon Transcribe, and Azure Speech to Text services were used to obtain comparative results on more than one Turkish ASR system. The application developed for the test procedure uses the directory structure given in Figure 1.

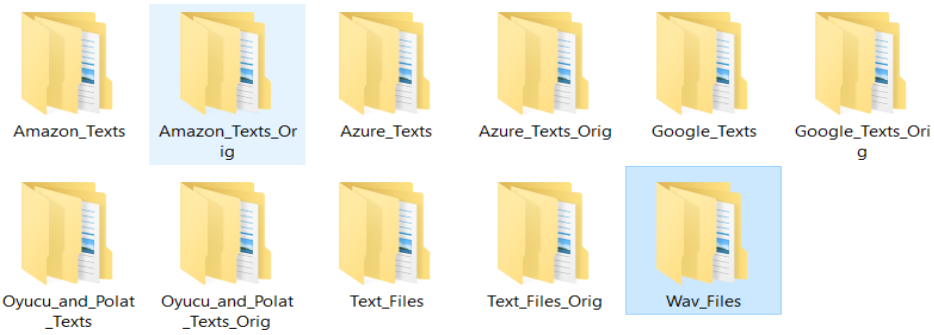

Figure 1. Test procedure directory structure

In the "Text_Files" directory in Figure 1, there are speech expressions transcribed directly by real users. In the "Wav_Files" directory, speech recordings are corresponding to text files. The "_Orig" statement at the end of the names of the folders indicates the directories where the output of the ASR system is stored without any processing on the text files. In the examinations, it has been seen that some ASR services do not only output text, numbers are written with digits, some words are written with capital letters and punctuation marks are added to the text. Therefore, all outputs of ASR systems are preprocessed from the WER calculation.

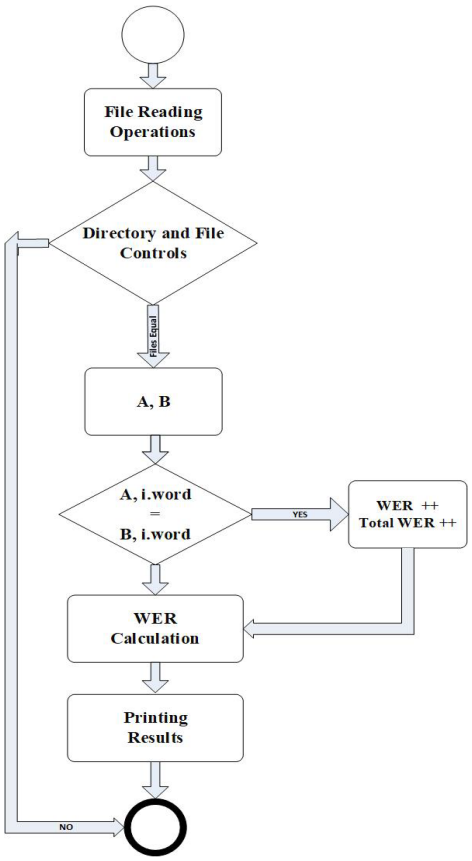

Figure 2. Flow chart for the test procedure

With the pre-processing performed before the WER calculation, the punctuation marks in the results of the Turkish ASR services were removed, uppercasWERe letters were converted to lowercase letters, and the text equivalents of the numbers were written. Thus, a clearer WER calculation was by analogy with one another the writing of all text files. The flow chart of the realized WER account is given in Figure 2.

Necessary calculations were made according to the flow chart shown in Figure 2 and it was printed on the screen to inform the user. The flowchart given in Figure 2 was coded with the Java language and converted into an application. Thanks to the application, average, minimum and maximum WER values were obtained for each speech file. Detailed information about WER is given in detail in the study of Kentsel et al [17]. It has been stated that WER is accepted as an industry standard for measuring speech recognition accuracy.

4. Experimental Results

With the test procedure presented within the scope of the study, different Turkish ASR outputs of each speech file were obtained. Example WER calculation for the Turkish ASR output obtained is given in figure 3. In Figure 3, a transcription of a speech in the field of law is shown.

In Figure 3, two different Turkish ASR outputs of a speech speaking in the field of law are shown. While the first text given as output represents the system developed by Polat and Oyucu, the second output represents the Google Speech to Text service. The original text shown in the figure includes the reference text of the speech transcribed by real users. Thus, each speech file was output according to different ASR systems and WER calculation was performed for each output. The total success of the ASR services is presented to the user by performing the average WER calculation. In addition to the average WER calculation, minimum and maximum WER calculations were performed (Table 4).

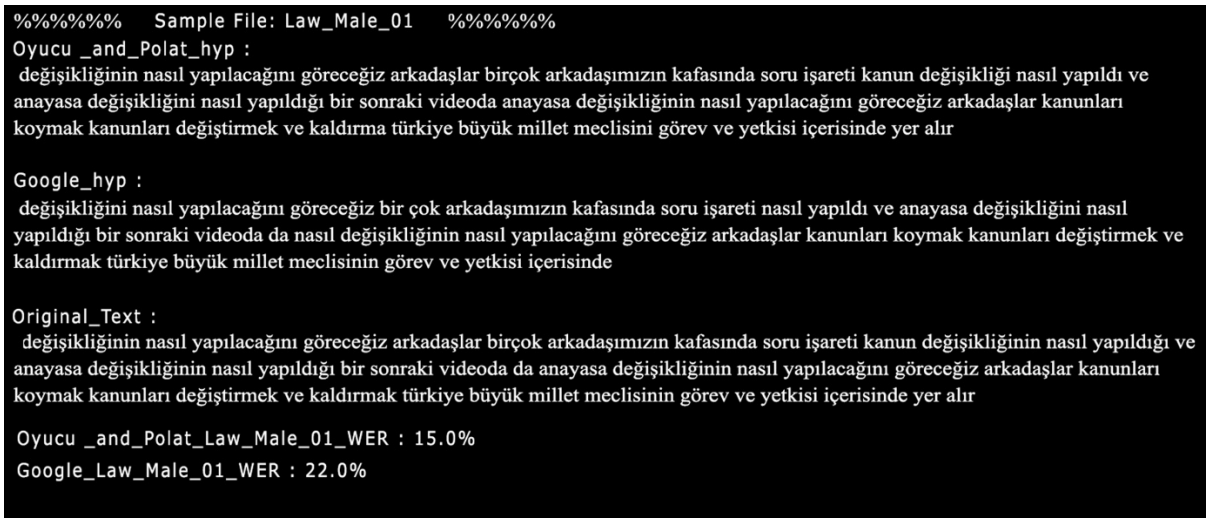

Figure 3. Comparative WER results

Table 4. WER results for different Turkish ASR systems

Name of Turkish ASR System	Minimum WER	Maximum WER	Average WER
Polat and Oyucu	0.0	68.0	20.24
Google Speech to Text	1.0	96.0	20.75
Amazon Transcribe	0.0	77.0	19.25
Azure Speech to Text	0.0	58.0	14.35

When the results in Table 4. were examined, it was seen that the Azure Speech to Text service gave the best average WER result for Turkish. Google Speech to Text service gave the worst average WER result. The difference between the average WER results is not very large. However, the difference between the minimum and maximum WER results is quite large. When the difference between the minimum and maximum values is examined, it is seen that Google has the highest maximum WER rate. The main reason for this high value is that Google preprocesses the speech files before exporting them to the ASR system. Speech files with noise above a certain noise ratio are not processed by Google. In addition, if a high WER rate is obtained, Google does not give results.

5. Conclusion and Recommendations

Due to the significant improvements in the success rates of automatic speech recognition systems, the application areas are increasing day by day. However, even state-of-the-art speech recognition systems that give some comparative results do not give successful results outside the vocabulary area. For this reason, first of all, related studies were reviewed. Then, a test corpus was prepared for Turkish ASR systems, where we can evaluate the large vocabulary problem. Using the prepared test corpus, a test procedure was prepared for the comparative evaluation of future Turkish ASR systems.

The test procedure prepared within the scope of the study has been tested on Google Speech to Text, Amazon Transcribe and Azure Speech to Text services. It has also been tested on the ASR system, which has a large vocabulary prepared by Oyucu and Polat in 2020. It has been observed that the obtained WER results vary between 14-21% and different ASR services have their advantages and disadvantages. Google's preprocessing of speech files and not processing files above a certain signal-to-noise ratio causes the maximum WER value to be high. However, this situation is completely different in Azure and Amazon. Therefore, they have a lower maximum WER ratio. When the existing Turkish ASR services were examined, it was seen that the problem of large vocabulary could not be solved. For this reason, studies should be carried out on the problem of large vocabulary for Turkish in future studies. The test corpus and test procedure presented in this study will guide the studies on large vocabulary. Another problem is the accent difference in rich languages such as Turkish. Turkish has many different regional accent structures. In future studies, systems that will not be affected by speech with different accents should be developed.

Acknowledgments

I would like to thank Associate Professor Hüseyin POLAT for all his efforts.

Conflict Of Interest

The authors declare that they have no conflict of interest.

References

[1] Prakoso H, Ferdiana R, Hartanto R. Indonesian Automatic Speech Recognition system using CMUSphinx toolkit and limited dataset. International Symposium Electronic Smart Devices 2016: 283-286.
[2] Miao Y. Kaldi+PDNN: Building DNN-based ASR Systems with Kaldi and PDNN. arXiv CoRR, 2014;1401.6:1-4, 2014.
[3] Yang X, Audhkhasi K, Rosenberg A, Thomas S, Ramabhadran B, Hasegawa-Johnson M. Joint modeling of accents and acoustics for multi-accent speech recognition. IEEE International Conference Acoustic Speech Signal Processing. 2018:5989-5993.

[4] Rebai I, Benayed Y, Mahdi W, Lorré J.P. Improving speech recognition using data augmentation and acoustic model fusion. Procedia Computer Science. 2017; 112:316-322.
[5] Jain A, Singh V.P, Rath S.P. A multi-accent acoustic model using mixture of experts for speech recognition. Annual Conference International Speech Communication Association. 2019: 779-783.
[6] Zeineldeen M, Glushko A, Michel W, Zeyer A, Schlüter R, Ney H. Investigating methods to improve language model integration for attention-based encoder-decoder ASR models. Annual Conference of the International Speech Communication Association. 2021: 2856-2860.
[7] Gandhe A, Rastrow A. Audio-attention discriminative language model for ASR rescoring. International Conference Acoustic Speech Signal Processing. 2020: 7944-7948.
[8] Anusuya M.A, Katti S.K. Speech recognition by machine, a review. International Journal of Computer Science and Information Security. 2009; 6:181-205.
[9] Dikici E, Saraçlar M. Semi-supervised and unsupervised discriminative language model training for automatic speech recognition. Speech Communication. 2016; 83:54-63.
[10] Irie K, Tüske Z, Alkhouli T, Schlüter R, Ney H. LSTM, GRU, highway and a bit of attention: An empirical overview for language modeling in speech recognition. Annual Conference of the International Speech Communication Association. 2016: 08-12.
[11] Siddharth D, Xinjian L, Florian M, Alan W. Domain robust feature extraction for rapid low resource ASR development. Black Language Technologies Institute. Carnegie Mellon University; Pittsburgh, USA. 2018; 258-265.
[12] Inaguma H, Cho J, Baskar M.K, Kawahara T, Watanabe S. Transfer learning of language-independent end-to-end ASR with language model fusion. IEEE International Conference Acoustic Speech Signal Processing. 2019: 6096-6100.
[13] Arısoy E, Can D, Parlak S, Saraçlar M, Sak H. Turkish broadcast news transcription and retrieval," IEEE Transaction Audio, Speech Language Processing. 2019; 17: 874-883.
[14] Salor Ö, Pellom B.L, Ciloglu T, Demirekler M. Turkish speech corpora and recognition tools developed by porting SONIC: Towards multilingual speech recognition. Computer. Speech Language. 2007; 21:580-593.
[15] Polat H, Oyucu S. Building a speech and text corpus of Turkish: Large corpus collection with initial speech recognition results. Symmetry. 2020; 12: 1-19.
[16] Abate S.T. Large vocabulary read speech corpora for four Ethiopian languages: Amharic, Tigrigna, Oromo and Wolaytta. International Conference Language Resource Evaluation Conference Proceeding. 2020: 4167-4171.
[17] Urban E, Buck A, Farley P, Bullwinkle M. Evaluate and improve Custom Speech accuracy. Microsoft Documents, 2022: 1-5.