CASR: A Corpus for Albanian Speech Recognition

Amarildo Rista *, Arbana Kadriu *
* SEE University/Faculty of Contemporary Sciences and Technologies, Tetovo, North Macedonia
ar29102@seeu.edu.mk;akadriu@seeu.edu.mk

***Abstract* – This research paper introduces a Corpus for Albanian Speech Recognition (CASR), aims for training and evaluating Automatic Speech Recognition models for Albanian language. The corpus is based on bible' audiobook, comprising 20 hours of transcribed audio data, where transcripts and audios are in the Albanian standard language. An end-to-end speech recognition model based on deep learning is implemented to test and evaluate this corpus. It shows that acoustic models trained on CASR gives satisfactory results. The corpus will be freely available for independent research and provides a valuable resource for research on Albanian ASR.**

Keywords - Speech Recognition, Corpus, Albanian Language.

I. INTRODUCTION

Automatic Speech Recognition (ASR) has grown tremendously in recent years, focusing mostly on deep learning techniques [1]. The ASR system is concerned with recognition and translation of spoken language into text by machine [2]. In this field compared to others even though includes general principles serving all languages of the world, has a difficulty to standardize its practical usage for all the natural languages. It happens partially because of the specificity of the different languages (including morphological, syntactical, grammatical and semantical ones), the dynamics (such as the constant language evolution), and language characteristics (mostly of spoken languages) [3]. The major issue for an ASR system relies in the design of the corpus which is determinant factor for its success. The following attributes are to be considered in the design of the corpus: speaker and channel variability which includes attributes such as; phonetics, adverse environment conditions (clean, noisy); speaker attributes such as; age, gender, accents, speed of utterance, dialects; training process and voice recording device [4]. An efficient ASR system should be able to identify all the above mentioned types of attributes to produce the text corresponding to input signal. This paper introduces the CASR corpus, which is a read speech dataset based on bible audio book which will be freely available for independent research. To show the effectiveness of this corpus an end-to-end speech recognition model based on deep learning is implemented. The rest of the paper is structured as follows: Section 2 - an overview of the Albanian Language; Section 3 - the corpus development; Section 4 is focused on the end-to-end speech recognition system; Section 5 relates the discussion to the overall results observed within the experiments; Section 6 - conclusions and recommendations.

II. OVERVIEW OF ALBANIAN LANGUAGE

Albanian language according to linguistic studies [5], is one of the oldest languages belonging to the group of Indo-European languages. The Albanian language is characterized by a rich dictionary with a large number of words, where in total there are about 41'000 words [6]. The grammar of the Albanian language is more complicated compared to other Indo-European languages [7]. Structurally, the Albanian language today is presented as a synthetic-analytical language, with a predominance of synthetic features and a tendency towards analyticism [8]. A good part of its phonetic and grammatical features are inherited from Indo-European languages, and other part are subsequent developments. The Albanian language today has its own phonological system, consisting of 7 vowels phonemes and 29 consonants phonemes. It has an accent with intensity and is generally immobile during flexion. In most cases, especially in the noun system, the emphasis falls on the penultimate syllable. The Albanian language has a developed system of grammatical forms, it has a binary inflection system, prominent and indistinguishable inflection, and still well preserved case forms (there are five cases) as well as the system of three genders, etc. [9]. The noun system has prominent and in-prominent forms and consequently prominent and in-prominent inflection. The plural of nouns is followed by suffixes as well as by changes within the noun. The nouns in Albanian language are inflected by gender (masculine, feminine, neuter) and number (singular and plural), also there are five declensions with five cases (nominative, accusative, genitive, dative and ablative) [10]. The verb in Albanian can be in one of the 6 possible moods: indicative, admirative, conjunctive, conditional, optative, and imperative [11]. In addition to inflection with special endings, the Albanian language also recognizes internal inflection. There are two types of adjective structures; similar and indistinguishable. The numbers are presented mainly by the decimal system, but also is preserve the vigesimal system. The Albanian language has a rich

system of modal and temporal forms, some of which are inherited and other were performed during its historical evolution. The verb system is very diverse where has six ways and three non-elaborate forms. The order of the words in the sentence is generally free. The lexicon consists from words of local origin and borrowed words. As far as spoken language is concerned, the presentation of language is manifested through its standard and dialects. The two main dialects of the Albanian language are the Tosk and Geg dialect. The lack of a digital dictionary as well as the application of some rules on it, encounters great difficulties in the processing of spoken language. Vocabulary is considered as a basic tool for spoken language processing.

III. CORPUS DEVELOPMENT

A. Source Data

The CASR is based on bible audio book, which is freely available[1]. The audio recordings are in MP3 format up to 8 minutes' length. All audio recordings are compressed in 32 bit-float and rate 16kHz. In total there are 20 hours' audio' records in the Albanian standard language. Text data are transcribed and saved in the PDF format and are published in the same website. The speeches are transcribed strictly verbatim and are adapted into standard written form by omitting speech disfluencies, correcting factual errors and slips of the tongue, and adding context to ensure the transcripts reflect the intentions of the speaker clearly and accurately. The speaker speaks in standard Albanian language without using dialects. The linguistic register is strictly standard, and the topics are primarily concerned with religious issues using a rich vocabulary. It is chosen to convert this type of raw data to a corpus suitable for ASR because of : compatibility between the written transcripts and the speeches; low presence of noise in the audio data and small compression loss of the audio data.

B. Audio and Text Preprocessing

Once all the audio recordings were downloaded it has been organized to adapt to the corresponding transcripts. Considering that most acoustic models are trained with relatively short utterances, usually up to a few seconds in length, the Audacity tool has been used [13] to align the audio recordings and split them into short utterances. This exercise is to briefly show how to create an audio file from the beginning to the final step. Each record audio is exported to the Audacity tool where the sound is represented by raw audio waves. The audio records has been trimmed in range 3 to 17 seconds length suitable for ASR systems. To avoid time spaces between words, they are cut leaving no more than 0.1 seconds length. Next, all audio files are exported in *wav* format and then are converted into *flac* format using a 16-bit linear PCM sample encoding (PCM_S16LE) sampled at 22.05 kHz[2]. Each audio file is composed of an average of 22 words. In total are crated 7667 utterances. A very important element is the naming of the files. Each utterance is named with a randomly generated four-digit number that corresponds to the text file names. Regarding to text preprocessing, each book's text is normalized by converting it into upper-case, re-moving the punctuation, and expanding common abbreviations and non-standard words [12]. It has been checked and corrected text-audio mismatch including inaccuracies in source texts, reader-introduced insertions, deletions, substitutions and transpositions, and involuntary disfluencies. Other significant sources of mismatch that were noticed are inaccurate text normalization and grapheme to phoneme errors in the automatically generated pronunciations by speaker. Data segmentation is a very important issue as they have to correspond with audio files. After segmenting the data, each segment is named with the same four-digit number that corresponds to the audio file. This way are created the audio and text files of the CASR corpus.

C. Corpus description and organization

After data normalization, text and audio preprocessing, the finished corpus contains 8300 utterances with a total duration of 20 hours and has a total size 1.3GB. The CASR is split into two subset called CASR_1 and CASR_2. Each subset is separated on two part: CASR_train_1, CASR_test_1 and CASR_train_2, CASR_test_2. Information on the corresponding transcripts, such as the number of sentences, the number of words and the average length of sentences are shown in Table 1.

TABLE 1. CORRESPONDING TRANSCRIPTIONS INFO

	CASR train_1	CAS test_1	CASR train_2	CASR test_2
Number of sentences	4850	867	1615	335
Avg. words per sentence	21.95	24	21.84	20.93
Total word	106493	20812	35274	7014

While for details regarding to speech from the both sets are presented in Table 2, such as number of utterances, duration and average duration per speaker. The splitting of the dataset in this way ,enable all independent researchers to experiment with their ASR models, first with small datasets by overcoming any hardware limitations they may have. Also to show how the size of the dataset affects the performance of ASR systems.

[1] https://albanianorthodox.com/dhjata-e-re-chc/
[2] https://cloudconvert.com/wav-to-flac

TABLE 2. CORRESPONDING SPEECH INFO

	CASR train_1	CASR test_1	CASR train_2	CASR test_2
Duration	12 h	3 h	4 h	1 h
Utterances	4850	867	1615	335
Avg.duration per utterance	8.9 s	12.5 s	8.9 s	10.7 s

IV. END-TO-END-SPEECH RECOGNITION

Recently Deep Neural Networks (DNN) are considered as fundamental part of current ASR [1]. With the introduction of end-to-end models [14, 15] deep learning has made a turn in the field of speech recognition. These models take as input the audio signal and directly output transcriptions. In this paper an end-to-end Speech Recognition model is introduced, suitable for Albanian language which will be trained and evaluated using CASR. The model will be focused in the Deep Speech models that are recurrent neural network (RNN) based architectures [16] which will be built in Pytorch tool [17]. This model has two main neural network modules — N layers of Residual Convolutional Neural Networks (ResCNN) to learn the relevant audio features, and a set of Bidirectional Recurrent Neural Networks (BiRNN) [18] to leverage the learned ResCNN audio features. The audio waves derived from CASR will be processed using the Mel-Frequency Cepstral Coefficients (MFCCs) [19], which are motivated by the nature of the speech signal. To simplify speech recognition pipelines by taking advantage of the capacity of deep learning system to learn from large datasets, the Connectionist Temporal Classification (CTC) loss function [20] will be applied, to predict the speech transcript. For handling the audio data, will be used a tool called *torchaudio* which is a library built by the PyTorch team specifically for audio data. The output of model is a probability matrix for characters which will be feed into the decoder to extract what the model believes are the highest probability characters that were spoken.

V. PERFORMANCE EVALUATION

The evaluation of ASR system is based on word error rate (WER) [21], loss training word and character error rate (CER) as metrics [22].

A. Assessment Criteria

I. Word Error and Word Error Rate

Word error is calculated by computing the Levenshtein distance between reference sequence and hypothesis sequence in word level. Where, Levenshtein distance is a string metric for measuring the difference between two sequences [23]. Informally, the Levenshtein distance is defined as the minimum number of single-character edits (substitutions, insertions or deletions) required to change one word into the other. The edits to word level can be naturally extended when calculate Levenshtein distance for two sentences. While word error rate (WER) compares reference text and hypothesis text in word-level. Mathematically WER is defined as:

$$WER = (Sw + Dw + Iw) / Nw \quad (1)$$

Where:
- Sw is the number of words substituted
- Dw is the number of words deleted
- Iw is the number of words inserted
- Nw is the number of words in the reference

II. Character Error and Character Error Rates

Character error is calculated by computed the Levenshtein distance between reference sequence and hypothesis sequence in char-level. While character error rate (CER) compares reference text and hypothesis text in char-level. Mathematically CER is defined as:

$$CER = (Sc + Dc + Ic) / Nc \quad (2)$$

Where:
- Sc is the number of characters substituted
- Dc is the number of characters deleted
- Ic is the number of characters inserted
- Nc is the number of characters in the reference

B. Results

First, the model was trained and evaluated using the CASR_1 corpus and then the CASR_2 corpus. Because of limited time and resources, our model get trained for 50 epochs only. Fig.1 indicates the WER results in percentage for CASR_1 and CASR_2. In both cases, as shown in Fig.1, a WER of 30% is obtained. Note that the size of the dataset does not affect the WER performance, as in the case when the model is trained with CASR_1 with 4 hours of audio as well as in the case when the model is trained with CASR_2 with 16 hours of audio, the WER achieves almost the same results.

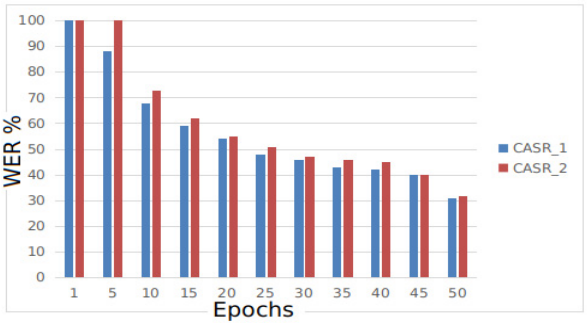

Figure 1. The performance on the development set in terms of word error rate.

Fig. 2 shows the CER results in percentage for CASR_1 and CASR_2. With increasing number of epochs, the error from the model has been sufficiently minimized reaching in 10% in both cases, as shown in Fig.2. The

number of epochs defines the number times that the learning algorithm will work through the entire training dataset. Also, the CER is not affected by the size of the dataset, as in the case when the model is trained with CASR_1 and in the case when the model is trained with CASR_2, the CER achieves almost the same results.

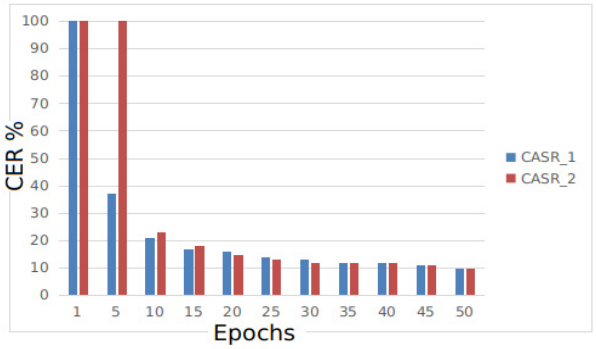

Figure 2. The performance on the development set in terms of character error rate.

Table 3 indicates the loss results. Loss function is used to gauge the error between the prediction output and the provided target value. We may conclude that the loss as in the case of model training with CASR_1 as well as in the case where the model is trained with CASR_2 is negligible, where with increasing number of epochs going to zero.

TABLE 3. LOSS

Epochs	1	10	20	30	40	50
CASR_1	2.91	0.73	0.58	0.54	0.54	0.48
CASR_2	2.93	0.78	0.53	0.51	0.51	0.49

VI. CONCLUSION

This paper introduces CASR, an Albanian speech corpus suitable to train and evaluate ASR systems. The corpus includes about 20 hours speech data from 1 speakers which speaks in standard Albanian language without using dialects. Both the audio and the corresponding text are provided in the corpus. The experimental results shows that the corpus is able to achieve good performances on ASR. The corpus is expected to accelerate research within the speech community and specifically for Albanian language. Design and analysis of heterogeneous and large corpora are expected to be analyzed in the future.

REFERENCES

[1] Kamath, U., Liu, J., & Whitaker, J. (2019). Deep learning for NLP and speech recognition (Vol. 84). Cham: Springer.

[2] Yu, D., & Deng, L. (2016). AUTOMATIC SPEECH RECOGNITION. Springer London limited.

[3] Comrie, B. (1989). Language universals and linguistic typology: Syntax and morphology. University of Chicago press.

[4] Juang, B. H. (1991). Speech recognition in adverse environments. Computer speech & language, 5(3), 275-294.

[5] Indo – European Languages – Evolution and Locale maps. Slocum J., Linguistic Research Center, The University of Texas at Austin & Collage of Liberal Arts. https://lrc.la.utexas.edu/eieol, accessed 30.01.2021.

[6] Fjalor i gjuhës së sotme shqipe. Akademia e Shkencave e RPS të Shqipërisë, Instituti i Gjuhësisë dhe Letërsisë, Tiranë 1980.

[7] Millaku, S., & Topanica, X. (2020). The contrast of Albanian language articles with Balkans and Indo-European languages. IJO-International Journal of Educational Research, 3(11), 37-52.

[8] Orel, V. Ė. (2000). A concise historical grammar of the Albanian language: reconstruction of Proto-Albanian. Brill.

[9] Newmark, L. (1980). Standard Albanian. A Reference Grammar for Students.

[10] Kadriu, A. (2013, June). NLTK tagger for Albanian using iterative approach. In Proceedings of the ITI 2013 35th International Conference on Information Technology Interfaces (pp. 283-288). IEEE.

[11] Kadriu, A. (2010). Modeling a two-level formalism for inflection of nouns and verbs in Albanian. Modeling Simulation and Optimization-Focus on Applications.

[12] Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., & Richards, C. (2001). Normalization of non-standard words. Computer speech & language, 15(3), 287-333.

[13] Audacity, I. (2013). What is Audacity.

[14] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... & Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.

[15] Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016, March). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4960-4964). IEEE.

[16] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.

[17] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703.

[18] Xie, P., Wang, G., Zhang, C., Chen, M., Yang, H., Lv, T., ... & Zhang, P. (2018, July). Bidirectional recurrent neural network and convolutional neural network (BiRCNN) for ECG beat classification. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 2555-2558). IEEE.

[19] Logan, B. (2000, October). Mel frequency cepstral coefficients for music modeling. In Ismir (Vol. 270, pp. 1-11).

[20] Graves, A. (2012). Connectionist temporal classification. In Supervised Sequence Labelling with Recurrent Neural Networks (pp. 61-93). Springer, Berlin, Heidelberg.

[21] Morris, A. C., Maier, V., & Green, P. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In Eighth International Conference on Spoken Language Processing.

[22] https://deepgram.com/blog/what-is-word-error-rate/, accessed 10.01.2021.

[23] Fiscus, J. G., Ajot, J., Radde, N., & Laprun, C. (2006, May). Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech. In LREC (pp. 803-808).