

Language Identification of Similar Short-time Speech in Russian and Ukrainian Based on Attention Mechanism and Convolutional Recurrent Neural Network

Xinyu Hu^{1*}

¹ Luoyang Campus, Former Information Engineering University, Luoyang, China

Email Address

huxinyu1002015@163.com (Xinyu Hu)

*Correspondence: huxinyu1002015@163.com

Received: 8 July 2023; **Accepted:** 24 August 2023; **Published:** 13 September 2023

Abstract:

Language identification is often used as a pre-processing technique for various speech recognition tasks. However, the current language identification model exhibits low accuracy when faced with confusable languages and short-time speech. Currently, there is a lack of research on Russian and Ukrainian, which share a high degree of similarity. To address these issues, we propose a language identification model based on Attention Mechanism (AM) and Convolutional Recurrent Neural Network (CRNN). Specifically, we first summarized the similarities and differences between Russian and Ukrainian, and represented the language features using Log-Mel spectrogram. Secondly, we designed a CRNN structure, which includes a CNN that captures spatial information and a Bidirectional Long Short-term Memory (BiLSTM) that captures temporal information, enabling the learning of language features from both dimensions. Finally, a module based on AM was designed to weight hidden states of BiLSTM, enhancing the learning performance of language characteristics. Experimental results on the collected dataset showed that the proposed model can identify similar short-time speech in Russian and Ukrainian with an accuracy of 95.38%, surpassing other existing models. Additionally, we also verified the effectiveness of Log-Mel spectrogram features through comparative experiments and confirmed the similarity of Russian and Ukrainian through confusion matrix.

Keywords:

Language Identification, Russian, Ukrainian, Short-time Speech

1. Introduction

Language Identification (LID) is a technology used to identify the language category of a given audio sample, which is commonly used as a preprocessing technique for various speech recognition tasks [1]. For instance, in an Automatic Speech Recognition (ASR) system, automatic language detection is essential as it allows for the correct parsing of speech corpus and the application of grammar rules, preventing

failures in subsequent speech recognition steps. Furthermore, with the advancement of deep learning and artificial intelligence, language identification plays an increasingly important role in multilingual speech recognition tasks such as machine translation, speech assistants, and sentiment analysis.

At present, several deep learning models have been proposed by researchers to identify the language type of speech [2], including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), etc. However, these studies often focus on classifying multilingual datasets with limited language associations, overlooking the linguistic features of each language and the similarities and dissimilarities between languages [3]. Therefore, it is meaningful to improve the accuracy of language identification performed on easily-confusable languages dataset or investigating the similarities and dissimilarities between a specific language and other languages. The presence of many interrelated words and pronunciations among these languages makes it challenging for existing language identification models to achieve accurate results, particularly for languages belonging to the same language branch [4], for example, it is difficult to distinguish between Russian and Ukraine. Their differences and similarities between the two languages are described in detail in Section 1.2.

To address the aforementioned issues of limited research on language identification for Russian and Ukrainian similar speech, as well as the poor performance of existing language identification models, this paper proposes a language identification method for Russian and Ukrainian similar short-time speech. This method builds a language identification model based on the Attention Mechanism (AM) and Convolutional Recurrent Neural Network (CRNN), where the CNN is responsible for capturing the spatial information, RNN is responsible for capturing the temporal information, and the Attention Mechanism enables the model to pay more attention to the features related to language itself and ignore non-language information, so as to enhance the learning of language information and improve the classification performance of the model. Additionally, this paper summarizes the similarities and differences between Russian and Ukrainian and conducts comparative experiments to validate the effectiveness of the Attention Mechanism and Convolutional Recurrent Neural Network-based language identification model for identifying similar short-time speech in Russian and Ukrainian.

This paper is organized as follows: Section 1 introduces the related work of language identification research as well as the similarities and differences between Russian and Ukrainian languages. Section 2 introduces the language recognition model based on AM-CRNN proposed in this paper. Section 3 describes the comparison experiment setup; Section 4 analyzes the experimental results; finally, summarizes and provides prospects.

2. Related Work

2.1. Language Identification Method

Since 2011, research of Deep Neural Networks (DNN) has achieved tremendous success in the field of speech recognition, prompting researchers to introduce deep learning methods into language identification tasks. In 2014, researchers first applied DNN models to the language task of short-term speech segments on a large scale, and realized feature extraction, using neural networks for feature extraction, feature trans-

formation, and classification [5]. Also in 2014, Ganapathy et al. [6] proposed using bottleneck features from CNN, training with labeled data of a target language, can provide important additional knowledge for convolutional acoustic features, including languages not used in its training. In 2016, an end-to-end language identification framework based on Attention Mechanisms was proposed [7]. In 2017, Bartz et al. [8] introduced a hybrid Convolutional Recurrent Neural Network that used spectrograms of input speech segments as features, achieving an average accuracy of 92% on a self-built corpus of six languages. In 2018, David et al. [9] proposed the x-vector method, which significantly improved language identification performance for short speech segments. Heracleous et al. [10] conducted a comparative study on language identification systems using DNN, CNN, and a combination of both. The system was trained and simulated on the NIST LRE 2015 dataset, and experimental results showed that the recognition cost of 10, 20, 30, 40 and 50 input languages was the lowest on CNN. In 2020, Draghici et al. [11] used Mel-spectrogram images as input features and modified the training strategy for language identification. In 2021, Kaiyr et al. [12] proposed a method combined with CNN and Long short-term Memory (LSTM) as an end-to-end classifier for language identification, achieving a test accuracy of 94.28% on seven languages: English, Kazakh, French, German, Russian, and Spanish. In 2022, Jianchuan Zhao et al. [13] proposed a multi-task learning network of Emphasized Channel Attention, Propagation and Aggregation-Time Delay Neural Network (ECAPA-TDNN) contrast prediction coding model for language recognition.

Differentiating highly confusable languages is a challenge in current language identification, and there is limited research in this area both domestically and internationally, with most of it focusing on the identification of dialects. In 2007, a study [14] discussed the language identification problem of Croatian, distinguishing it from Serbian, Slovenian, and Slovak. In 2012, Zampieri et al. [15] proposed the use of character n-gram models and word n-gram language models to identify and classify European and Brazilian Portuguese. In 2018, Suwon Shon et al. [16] extracted three acoustic features, Mel Frequency Cepstral Coefficient (MFCC), Filter Bank (Fbank), and spectrograms, to identify Arabic dialects. In 2019, Xie Kexin [17] proposed a method combining Gated Recurrent Units (GRU) and Hidden Markov Models (HMM) for the identification of Hunan dialect. In 2020, Aitor Arronte Alvarez et al. [18] proposed the Res-BiLSTM end-to-end network, which combines Residual Blocks and Bidirectional Long Short-term Memory (BiLSTM), for Arabic dialect identification. In 2022, Fu Ying built a CNN-BiGRU network consisting of CNN and Bidirectional Gated Recurrent Units (BiGRU) [19] to identify eight dialects: Changsha, Hebei, Hefei, Nanchang, Ningxia, Shaanxi, Shanghai, and Sichuan.

When conducting language identification, we pay closer attention to the differences between two languages, mainly in terms of pronunciation and vocabulary, as shown in next section.

2.2. Similarities and Dissimilarity between Russian and Ukrainian

Russian is the official language of Russia, belonging to the East Slavic branch of the Indo-European language family. Currently, approximately 285 million people worldwide use Russian as their first or second language, making it the most widely spoken language among the Slavic language branch and it is also one of the six working languages of the United Nations. Ukrainian, as the official language of Ukraine, is used by approximately 45 million people and belongs to the same

language family and branch as Russian. Therefore, Russian and Ukrainian exhibit a high degree of similarity, which can be observed in the following aspects:

Firstly, 62% of the words in their vocabulary are the same, such as basic words like numbers, family members, and animals. Table 1 provides relevant examples.

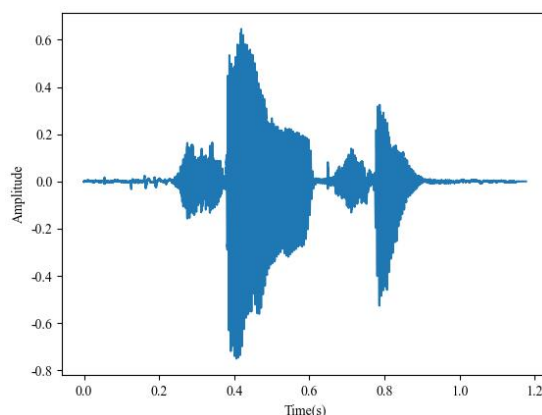


Figure 1 (a)

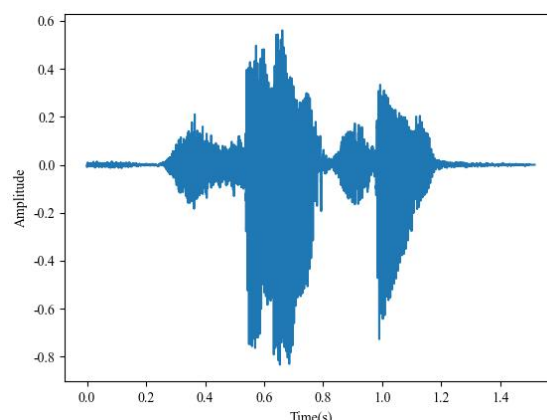


Figure 1 (b)

Figure 1. The oscillogram of “sun” in Russian is shown as (a); the same in Ukrainian is shown as (b).

And the audio frequency oscillogram of the word “sun” is shown in Figure 1 (Words in it default to the nominative singular form).

Table 1. Examples of similar pronunciation words and transliterations in Russian and Ukrainian.

Word	Russian	Ukrainian	Pronunciation in Russian	Pronunciation in Ukrainian
Sun	солнце	сонце	solnce	sontse
Word	свет	світ	svet	svit
River	речка	річка	rečka	richka
Mother	мать	мати	mat'	máty
Bird	птишка	пташка	ptaška	ptashka

Secondly, they have similar grammatical structures, both featuring six cases (nominative, genitive, dative, accusative, indirect accusative, and prepositional) and three genders (masculine, feminine, and neuter). Thirdly, their phonetic system is similar, utilizing the Cyrillic alphabet and sharing similarities in their phonetic systems. The differences between the two languages are outlined in Table 2.

Table 2. Differences between Russian and Ukrainian from three level.

Level	Russian (Ru)	Ukrainian (Ukr)
Letter	There are 33 letters, missing two letters <i>ï</i> and <i>r</i> in Ukrainian language.	There are 31 letters. It lacks four letters <i>ѐ</i> , <i>ы</i> , <i>э</i> and <i>ъ</i> , compared to Russian.
Pronunciation	1. There are three Russian consonants, <i>ж</i> , <i>ш</i> , and <i>ц</i> , that are always hard (also called non-palatalized) and they have no soft (palatalized) counterpart. 2. In Russian, “a” and “o” are weakened in unstressed syllables.	1. They also have soft counterpart in Ukrainian: <i>шя</i> , <i>жя</i> , <i>ця</i> 2. They aren’t weakened in unstressed syllables.
Vocabulary	1. They have some totally different words: “Strawberry” (Ru) клубника-(Ukr) полуниця; “Eye” (Ru) глаз-(Ukr) око etc. 2. They also have some words with the same meaning that differ by just one letter change. (1) о - і: (Ru) соль, переход - (Ukr) сіль, перехід. (2) е - і: (Ru) хлеб, вера, зверь - (Ukr) хліб, віра, звірь. (3) е - о: (Ru) человек, его, четыре - (Ukr) чоловік, його, чотири. (4) л – в: (Ru) волк, шелк, желтый - (Ukr) вовк, шовк, жовтий (In Ukrainian, the letter “л” changes to “в” when it appears after a vowel.)	

3. AM-CRNN Model

In this paper, a language identification model based on the Attention Mechanism and CRNN is proposed in this paper, comprising three components: CNN for spatial information capture, BiLSTM (a type of RNN) for temporal information capture, and a self-attention module to enhance the learning of language information and improve the model’s classification performance. Traditionally, language identification based on BiLSTM involves inputting features into the BiLSTM layer and connecting it to the Softmax layer for output. The final output sequence from the BiLSTM’s last node is used as the embedding to express the language. This paper introduces an innovative approach by combining the AM with CNN and BiLSTM. The output of each node in the BiLSTM is recorded, and Attention is applied by weighting each frame’s features, thereby enhancing the algorithm’s performance. The model’s structure, as illustrated in Figure 2, begins with the extraction of features from the input speech, producing a Log-Mel spectrogram as the speech feature. Subsequently, these features are processed by CNN to extract spatial information, followed by inputting the convolutional results into the BiLSTM network. The output of each node in the BiLSTM is recorded and serves as input for the Attention block. Finally, the language is determined through the Softmax classifier.

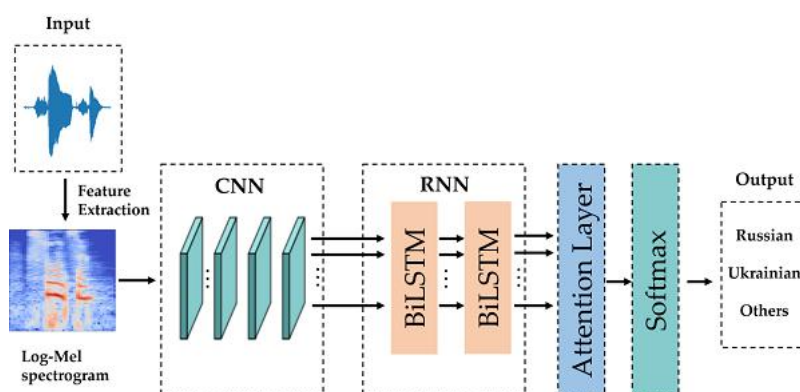


Figure 2. Language identification model structure based on Attention Mechanism and CRNN.

3.1. CNN

Within the architecture of the CNN, there are 4 convolutional layers and 2 max pooling layers. Each convolutional layer is 2-Dimensional Convolution (Con2D) activated by the non-linear function Rectified Linear Unit (ReLU), to reduce the number of parameters and complexity. Each convolutional kernel size is $3 * 3$. The first and second convolutional layers have 512 output channels. The pooling layers have a size of $3 * 3$ with the step size of 3. The third convolutional layer has 256 output channels, and the fourth has 128. The CNN extracts local visual features from the input spectrogram through a series of convolutional steps, resulting in feature maps with a height of 1. The final convolutional layer with a size of $t * 1 * c$ (c is the number of convolution kernel) is divided into t time steps along the time axis. Each time step represents a frequency feature extracted and serves as the input for the RNN.

3.2. RNN

In order to obtain the temporal features of the input spectrogram, we employ BiLSTM of RNN structure for time-series feature extraction of the output results from the convolutional module. In this paper, BLSTM is composed of two separate LSTMs with opposite directions. The LSTM structure includes a forget gate, cell state, hidden state, input gate, output gate and candidate state. Each LSTM has 256 output units, allowing for the storage of contextual information and output.

3.3. Attention Mechanism

After obtaining the output sequence from the BiLSTM, attention network is employed to enhance the learning of language information and improve the classification performance of language identification. Attention weights are applied to all hidden states, enabling the focus to be concentrated on the most important information within the sequence. The vector h represents the LSTM operation, with h vectors from different directions representing the bidirectional LSTM.

$$\begin{aligned}\vec{a}_n &= h(\vec{a}_n), n \in [1, N] \\ \overleftarrow{a}_n &= h(\overleftarrow{a}_n), n \in [N, 1] \\ a_i &= [\vec{a}_n, \overleftarrow{a}_n]\end{aligned}\quad (1)$$

In Formula (1), N represents the total number of audio samples. The vector a_i serves as the foundation of the Attention Mechanism. The objective of the AM is to train the model by randomly initializing weight values and focusing on key information, and then making the network continuous through the $\tanh$ function. The layer with the $\tanh$ function is multiplied again with the trainable context vector u_i , which is randomly initialized and jointly learned during the training process. The improved vector is denoted as a'_i , as shown in Formula (2):

$$a'_i = \tanh(a_i \cdot w_i + b_i) \cdot u_i. \quad (2)$$

By dividing the exponential values of the previously generated vectors by the sum of all the exponentiated values, and finally providing weights for each W_i to calculate the context vector, as shown in Formula (3). To avoid a denominator of zero, ϵ is added to the denominator.

$$W_i = \frac{\exp(a'_i)}{\sum_i \exp(a'_i) + \epsilon}. \quad (3)$$

The structure of the attention module is shown in Figure 3, where a_i represents the input of the module and the output of the BiLSTM layer.

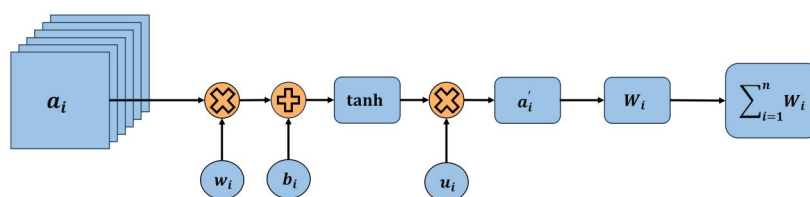


Figure 3. Attention module structure.

Afterwards, the obtained importance weights are combined with the previously computed context vector and fed into a Softmax layer for the backend classification of language identification.

4. Experiment

Firstly, 62% of the words in their vocabulary are the same, such as basic words like numbers, family members, and animals. Table 1 provides relevant examples, and the audio frequency oscillogram of the word “sun” is shown in Figure 1 (Words in it default to the nominative singular form). Secondly, they have similar grammatical structures, both featuring six cases (nominative, genitive, dative, accusative, indirect accusative, and prepositional) and three genders (masculine, feminine, and neuter). Thirdly, their phonetic system is similar, utilizing the Cyrillic alphabet and sharing similarities in their phonetic systems.

4.1. Datasets

To assess the performance of the language identification model proposed in this paper, pure Russian and Ukrainian speech data sets are collected from public sources such as television and radio, with each set consisting of 21 hours of data. Some of the speech data is acquired through microphone recordings, with a sampling rate of 16kHz, a precision of 16 bits, and in mono channel format. The datasets for each language comprise recordings from multiple speakers, including both males and females. The entire dataset is approximately 4.5 GB in size and stored in “.wav” format.

Both language data sets are divided into two non-overlapping subsets: a training set of 20 hours, which includes approximately 20,000 speech samples, and a test set of 500 speech samples. Since this study primarily focuses on the differentiation between Russian and Ukrainian at the phonetic and lexical levels, it is crucial to evaluate the model's identification performance on a similar short-duration speech dataset containing phrases in both languages. Therefore, the maximum length of the testing utterances was limited to no more than 5 seconds.

Additionally, a separate dataset of 20 hours of mixed Slavic languages, including Belarusian, Bulgarian, Czech, and Polish, are also prepared for training the model. The label for this mixed dataset is “Others”. Similarly, 500 utterances are randomly selected from this dataset for testing purposes.

4.2. Feature Extraction

The spectrogram is a two-dimensional visualization of the frequency-based changes in a signal over time. The Log-Mel spectrogram is a two-dimensional visual description of the frequency of a signal at different times, which can highlight

fluctuations in energy levels. The process of obtaining a Log-Mel spectrogram mainly includes the following steps:

Firstly, Windowing operation is performed. Since the characteristics of speech signals remain relatively stable over short time intervals, the signal can be considered as stationary within these short frames. The Hamming window, known for its high frequency resolution and better performance in capturing the frequency characteristics of short-duration signals, is chosen as the window function, shown as Formula (4):

$$w(n) = \begin{cases} 0.54 - \cos\left(\frac{2\pi n}{N-1}\right), & 0 \leq n \leq N-1 \\ 0, & \text{Others} \end{cases} \quad (4)$$

As the features of a signal are often difficult to discern in the time domain, each frame, after windowing with the Hamming window, needs to undergo Fast Fourier Transform (FFT) to obtain the signal's energy distribution in the frequency domain. The originally non-periodic speech signal in the window is transformed into a periodic signal, and different energy distributions represent different speech features.

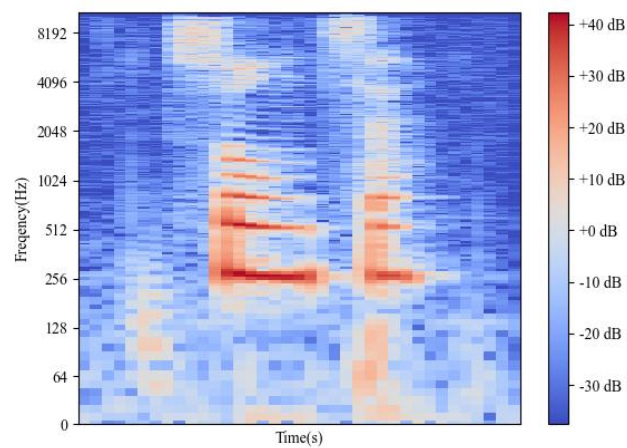


Figure 4 (a)

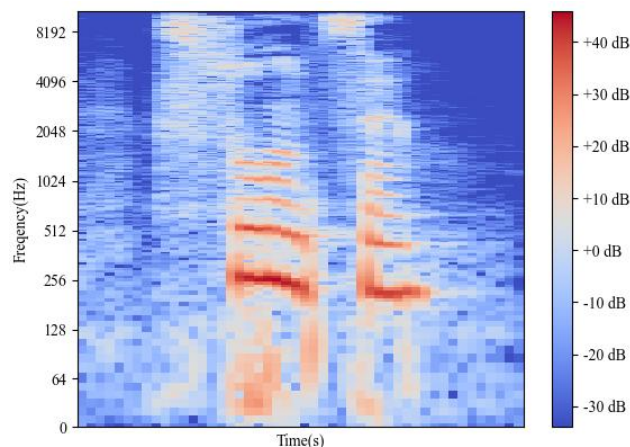


Figure 4(b)

Figure 4. Log-Mel spectrogram of the “sun” speech signal in Russian and Ukrainian are represented by (a) and (b) respectively.

After obtaining the processed signal, feature extraction is performed. While Linear Predictive Coding (LPC) is a commonly used method, but human perception of sound is nonlinear. Therefore, to process speech in a manner in a way similar to human

perception, this study firstly applies a Mel Filterbank to map the linear spectrogram into a nonlinear representation based on Mel frequency. The energy of the Mel Filterbank is then transformed to the logarithmic scale to generate the Log-Mel spectrogram using either Formula (5), enabling conversion between Hertz (f) and Mel (F_{mel}):

$$F_{mel} = 3322.23 \lg (1 + 0.001) f_{Hz} \quad (5)$$

The resulting Log-Mel spectrogram, as shown in Figure 4, provides a more intuitive visualization of the high similarity between the word in the two languages.

4.3. Evaluation Indicator

In this study, the performance of the experimental model is evaluated using accuracy, precision, recall, and F1 score. For all metrics, Tp represents true positives, Tn represents true negatives, Fp represents false positives, and Fn represents false negatives.

Accuracy (Acc) is the proportion of correctly predicted samples to all samples, and it is calculated as:

$$Accuracy = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \quad (7)$$

Positive Predictive Value (PPV) is the probability of true positive predictions among all samples predicted as positive, and it is expressed as:

$$Precision = \frac{Tp}{Tp + Fp} \quad (8)$$

True Positive Rate (TPR) measures the proportion of correctly predicted positive samples to all actual positive samples, and it is defined as:

$$Recall = \frac{Tp}{Tp + Fn} \quad (9)$$

F1-score is the weighted average of precision and recall, and it is calculated as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (10)$$

4.4. Experiment Settings

The present study conducts the following experiments on the same dataset: Firstly, different features are employed for language identification, including the extraction of commonly used Fbank features, MFCC features, and the Log-Mel spectrogram features primarily used in this experiment. A CNN-based discriminative model is employed for language identification, and the accuracy of the results is compared to evaluate the effectiveness of selecting the Log-Mel spectrogram features.

Secondly, experiments are conducted for language identification using CNN, CRNN, AM-CRNN methods. By comparing the identification results, the effectiveness of the AM-CRNN proposed in this paper for the identification of similar

Russian and Ukrainian short-time speech is verified. Furthermore, patterns of mispredicted languages are analyzed and summarized.

We have trained the language identification model in this paper based on the TensorFlow deep learning architecture, and the parameters are shown in Table 3.

Table 3. *Model parameter.*

Parameter	Parameter
Convolution kernel	3×3
Batch size	256
Dropout	0.1
BiLSTM layer	2
Learning rate	0.0001

4.5. Experimental Results and Analysis

The accuracy of the three feature recognition methods is shown in Table 4. The results indicate that, within the CNN architecture employed in this experiment, the Log-Mel spectrogram feature exhibits a higher accuracy in language identification compared to Fbank and MFCC, achieving 86.56%.

Table 4. *Recognition accuracy of Fbank, MFCC and Log-Mel feature.*

Feature	Accuracy
Fbank	84.35%
MFCC	81.60%
Log-Mel	86.56%

Subsequently, the Log-Mel spectrogram feature is tested on the three models: CNN, CRNN, and AM-CRNN, and the experimental accuracy is presented in Table 5.

Table 5. *Recognition accuracy of CNN, CRNN and AM-CRNN model.*

Model	Accuracy
CNN	86.56%
CRNN	90.21%
AM-CRNN	95.38%

Further detailed evaluation metrics are provided in Table 6.

Table 6. *Evaluate metrics of CNN, CRNN and AM-CRNN model in detail*

Language	CNN				CRNN				AM-CRNN			
	PPV	TPR	F1	ACC	PPV	TPR	F1	ACC	PPV	TPR	F1	ACC
Russian	0.8481	0.8707	0.8592	0.8656	0.8942	0.9013	0.8977	0.9024	0.9444	0.9620	0.9531	0.9538
Ukrainian	0.8065	0.8667	0.8355		0.8878	0.9340	0.9103		0.9465	0.9440	0.9453	
Others	0.9562	0.8593	0.9052		0.9277	0.8720	0.8990		0.9709	0.9553	0.9630	

The results demonstrate that, on a testing speech dataset of no more than 5 seconds, the three models used in this study achieve accuracy rankings as follows (from smallest to largest): CNN, CRNN, and AM-CRNN, with the AM-CRNN model designed in this paper achieving the highest recognition accuracy of 95.38%. Similarly, the precision, recall, and F1 score rankings for Russian, Ukrainian, and

other language identification are roughly the same, providing evidence of the effectiveness of the AM-CRNN models employed in this experiment for identifying similar short-time speech in Russian and Ukrainian.

It is worth noting that when using the CNN model for identifying “Others” language category, its precision is higher than that of the CRNN model. An analysis of the reason behind this is conducted. The confusion matrices corresponding to the identification results of the three models are shown in Figures 5a, b and c. PPV represents the accuracy of positive results. Through a comparative analysis of the CNN and CRNN confusion matrices, it can be observed that during CNN identification, more samples with a true value of “Russian” are identified as “Ukrainian” rather than “Others”. Consequently, the total count of predictions as “Others” decreases, thereby increasing precision. Therefore, it can be concluded that when the selected model's recognition performance is not satisfactory, Russian and Ukrainian exhibit higher similarity and confusion.

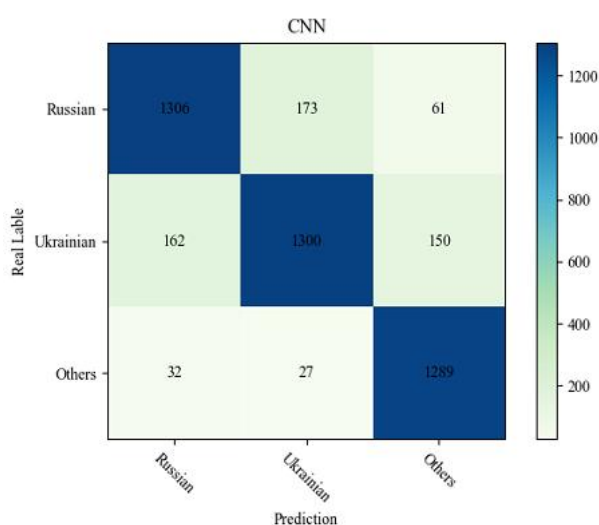


Figure 5 (a)

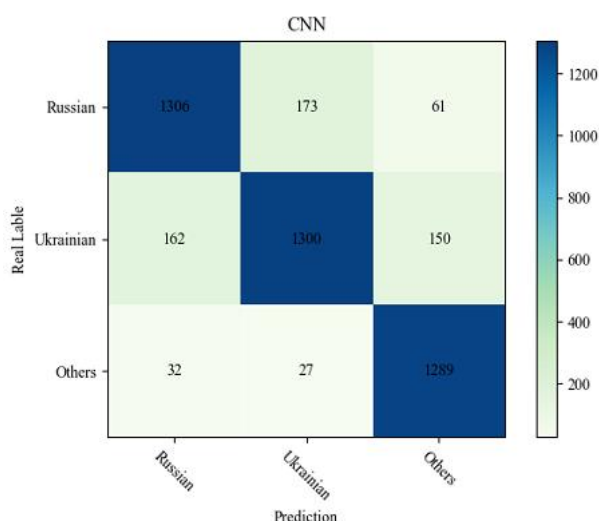


Figure 5 (b)

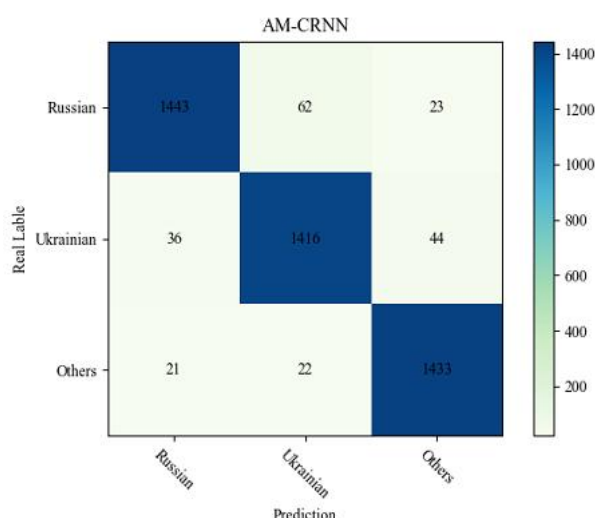


Figure 5(c)

Figure 5. Confusion matrix of CNN, CRNN and AM-CRNN are represented by (a), (b) and (c) respectively

By comparing the confusion matrices of the identification results of the three models, it can be seen that in all three experiments, when samples with a true value of “Russian” are mispredicted, the number of predictions as “Ukrainian” is higher than as “Others.” For samples with a true value of “Ukrainian” being mispredicted, more are predicted as “Russian” than as “Others”. However, regarding mispredictions of “Others” as the true label, no specific pattern is observed. Thus, it can be inferred that, compared to the other Slavic language datasets chosen in this paper, Russian and Ukrainian exhibit the highest similarity and confusion.

5. Conclusions

In this paper, we proposed a language identification model based on Attention Mechanism and Convolutional Recurrent Neural Network (AM-CRNN) for identifying the language of similar short-time speech in Russian and Ukrainian. The research addresses issues of limited studies on language identification of similar Russian and Ukrainian speech and poor performance of existing language identification models. The proposed AM-CRNN model extracts Log-Mel spectrogram features firstly, which are then fed into a CNN to capture spatial features. The convolutional results are subsequently input into a BiLSTM network to extract temporal information. Afterwards, the outputs of each node are recorded and fed into an Attention module. Finally, the language identification results can be obtained through a Softmax layer. In the paper, we collect and establish datasets for Russian, Ukrainian, and other Slavic languages, then conducts comparative experiments based on these datasets. The effectiveness of Log-Mel spectrogram in feature extraction is validated with an accuracy of 86.56%, which is higher than other features. Moreover, the performance of the proposed AM-CRNN model in language identification, with an accuracy of 95.38%, was superior to other models. The results indicate that the Attention Mechanism enables the model to focus more on language-related features, thereby enhancing the learning of language information and improving the classification performance of the model. Furthermore, the paper summarizes the similarities and differences between Russian and Ukrainian speech and verifies their similarity through a confusion matrix. In future research, the differences between these two languages will be further considered to improve identification accuracy.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

- [1] Oro, E.; Ruffolo, M.; Sheikhalishahi, M. Language Identification of Similar Languages Using Recurrent Neural Networks. In Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART), Funchal, Madeira, Portugal, 16-18 January 2018, Volume 2, pp. 635-640.
- [2] Mandal, A.; Pal, S.; Dutta, I. Is Attention Always Needed? A Case Study on Language Identification from Speech. arXiv 2021, arXiv:2110.03427.
- [3] Prasad, K.V.; KP S. A Deep Learning Approach For Similar Languages, Varieties And Dialects. arXiv 2019. arXiv: 1901.00297.
- [4] Thukroo, I.A.; Bashir, R. Spoken Language Identification System for Kashmiri And Related Languages Using Mel-Spectrograms And Deep Learning Approach. In Proceedings of the 2021 7th International Conference on Signal Processing and Communication (ICSC), Noida, India, 25–27 November 2021, pp. 250-255. DOI: 10.1109/ICSC53193.2021.9673212.
- [5] Lopez-Moreno, I.; Gonzalez-Dominguez, J.; Plchot, O. Automatic Language Identification Using Deep Neural Networks. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 04-09 May 2014, pp. 5337-5341. DOI: 10.1109/ICASSP.2014.6854622.
- [6] Ganapathy, S.; Han, K.; Thomas, S. Robust Language Identification Using Convolutional Neural Network Features. In Proceedings of the 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014, pp. 1846-1850.
- [7] Geng, W.; Wang, W.; Zhao, Y. In Proceedings of the INTERSPEECH, San Francisco, CA, America, 08-12 September 2016, pp. 2944-2948.
- [8] Bartz, C.; Herold, T.; Yang, H. Language Identification Using Deep Convolutional Recurrent Neural Networks. In Proceedings of the 24th International Conference on Neural Information Processing (ICONIP), Guangzhou, China, 14–18 November 2017, End-To-End Language Identification Using Attention-Based Recurrent Neural Networks. pp. 880-889. https://doi.org/10.1007/978-3-319-70136-3_93.
- [9] Snyder, D.; Garcia-Romero, D.; McCree, A.; Sell, G.; Povey, D.; Spoken Language Recognition Using X-Vectors. Odyssey 2018, 105-111.
- [10] Heracleous, P.; Takai, K.; Yasuda, K.; Comparative Study on Spoken Language Identification Based on Deep Learning. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3-7 September 2018, pp. 2265-2269.

- [11] Draghici, A.; Abeßer, J.; Lukashevich, H. A Study on Spoken Language Identification Using Deep Neural Networks. In Proceedings Of the 15th International Audio Mostly Conference, Graz Austria, 15-17 September 2020, pp. 253-256. <https://doi.org/10.1145/3411109.3411123>.
- [12] Kaiyr, A.; Kadyrov, S.; Bogdanchikov, A.; Automatic Language Identification from Spectrogram Images. In Proceedings of the 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST), Nur-Sultan, Kazakhstan, 28–30 April 2021, pp. DOI:1-4. 10.1109/SIST50301.2021.9465996.
- [13] Zhao, J.C.; Yang, H.Q.; Xu, Y. Multitask Learning Language Recognition Method Based on Contrast Predictive Coding Model. *Data Acquisition and Processing*. 2022, 37.02, 288-297.
- [14] Ljubesic, N.; Mikelic, N.; Boras, D. Language Identification: How To Distinguish Similar Languages. In Proceedings of the 2007 29th International Conference On Information Technology Interfaces, Cavtat, Croatia, 25-28 June 2007, pp. 541-546.
- [15] Zampieri, M.; Gebre, B.G. Automatic Identification of Language Varieties: The case of Portuguese. In the Proceedings of the 11th Conference on Natural Language Processing (KONVENS), Vienna, Austria, 19-21 September 2012, pp. 233-237. <https://hdl.handle.net/11858/00-001M-0000-000F-EC46-0>.
- [16] Shon, S.; Ali, A.; Glass, J. Convolutional Neural Networks and Language Embeddings for End-To-End Dialect Recognition. *arXiv* 2018, arXiv:1803.04567.
- [17] Xie, K.X.; Hu, D.; Zou, X. Hunan Dialect Identification Based on GRU-HMM Acoustic Model. *Computer and Digital Engineering*, 2019, 47(03), 496+538.
- [18] Alvarez, A.A.; Issa, E.S.A. Learning Intonation Pattern Embeddings For Arabic Dialect Identification. *arXiv* 2020, arXiv:2008.00667.
- [19] Fu, Y.; Li, Z.L.; Tang, H. Dialect Recognition Based on CNN-BiGRU. *Communication Technology*, 2022, 55(06), 712-719.



© 2023 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)