Neurorecognition visualization in multitask end-to-end speech

Orken Mamyrbayev[a], Sergii Pavlov[b], Akbayan Bekarystankyzy[c,d] , Dina Oralbekova[e], Bagashar Zhumazhanov[a] , Larysa Azarova[b], Dinara Mussayeva[f], Tetiana Koval[g], Konrad Gromaszek[h]*, Nurdaulet Issimov[i], Kadrzhan Shiyapov[j]

[a]Institute of Information and Computational Technologies, 28 Shevchenko St, 050010 Almaty, Kazakhstan; [b]Vinnytsia National Technical University, Khmel'nyts'ke Hwy, 95, 21000 Vinnytsia, Ukraine; [c]Satbayev University, Almaty, Kazakhstan, Satpaev St 22, 050000 Almaty, Kazakhstan; [d]Narxoz University, Almaty, Kazakhstan, Zhandossov St 55, 050035 Almaty, Kazakhstan; [e]Almaty University of Power Engineering and Telecommunications, Baytursynuli St 126/1, 050013 Almaty, Kazakhstan; [f]Institute of Economics CS MES RK, Almaty, Kazakhstan, Kurmangazy St 29, A25K1B0 Almaty, Kazakhstan; [g]Vinnytsia Mykhailo Kotsiubynskyi State Pedagogical University, Ostroz'koho street 32, 21100 Vinnytsia, Ukraine; [h]Lublin University of Technology, ul. Nadbystrzycka 38D, 20-618 Lublin, Poland; [i]Turan University, Satpayeva St. 16a, 050013, Almaty, Kazakhstan; [j]Abai Kazakh National Pedagogical University, Dostyk Ave 13, 050010 Almaty, Kazakhstan

ABSTRACT

Nowadays, speech-processing technologies with different language systems are successfully used in mobile and stationary devices. Kazakh is considered a low-resource language, which poses various challenges for conventional speech recognition methods. This paper presents a proposed model capable of multitasking and handling concurrent speech recognition, dialect identification, and speaker identification, all in an end-to-end framework. The developed multitask model enables training three different tasks within a single model. A multitask recognition system is created based on the WaveNet-CTC model. Experiments show that for the concrete task end-to-end multitask model has better performance than other models..

Keywords: end-to-end, multitask training, speech recognition, speaker identification, dialect identification

1. INTRODUCTION

Beginning machine learning (ML) methods are effectively used in recognition technologies. Real-time speech recognition systems that use ML models like CTC Listen, Attend, and Spell end-to-end significantly improved the performance of systems. Due to the system performance improvement, it was possible to use devices with speech translation, speech diarization, speaker identification, and speech recognition systems. Nowadays, speech-processing technologies with different languages are successfully integrated into mobile and stationary applications. Nonetheless, there is still a shortage of research and a lack of availability of systems for speech processing and its applications for multiple low-resource languages.

Kazakh language is one of the low-resource languages and belongs to the group of agglutinative languages. Researchers are increasingly interested in processing Kazakh speech due to its high demand and potential applications in various real-life scenarios, indicating that speech recognition in Kazakh has significant practical implications.

The last improvements in computing technologies allowed the introduction of end-to-end models for speech recognition, which show better results than Hidden Markov Models (HMM). End-to-end automatic speech recognition (ASR) systems have demonstrated faster and more efficient performance than traditional Deep Neural Networks (DNNs), particularly for low-resource languages1. Khassanov et al. in[2] present a 335-hour dataset for the Kazakh language. The experiments have shown that using a sufficiently large training dataset results in better metrics for speech recognition with the use of end-

*k.gromaszek@pollub.pl; www.pollub.pl

Optical Fibers and Their Applications 2023, edited by Waldemar Wójcik, Zbigniew Omiotek,
Andrzej Smolarz, Proc. of SPIE Vol. 12985, 129850G · © 2023 SPIE
0277-786X · doi: 10.1117/12.3022727

to-end models in comparison with hybrid models. The study described in reference` examined the "listen, attend, spell" model and found that it performed well in recognizing seven different dialects of the English language. Watanabe et al. [4] introduced ESPnet, an open-source speech processing platform. ESPnet is primarily designed for automatic speech recognition (ASR) and employs popular deep neural network (DNN) frameworks such as Chainer and PyTorch. The core technology of ESPnet is based on deep learning. In a related study[5], a similar platform was used to significantly improve end-to-end multitasking training for a group of Indian languages. This approach involved linking the encoder to the language identity of the speech, resulting in improved performance across all languages. The research findings suggest that an end-to-end model can effectively address language differences by training and optimizing a single neural network.

This work is structured the following way: part 2 observes advanced research in an appropriate scientific direction, part 3 shows the model of end-to-end multitask recognition, experiments are described in part 4, part 5 contains a discussion of experimental results, and conclusions are made in part 6.

2. RELATED WORK

The development of machine learning significantly impacted the quality of speech recognition. Orken et al.[9] shows the usage of DNN in speech recognition. Scientists have used different models of neural networks like ANN, CNN, RNN, and LSTM 10, 11, and [12] to improve speech recognition performance.

In recent years, E2E systems based on deep learning have shown impressive results in ASR and dialect identification[13]. The efficacy of the primary end-to-end systems is contingent on the amount of training data, and to address specific tasks, multiple databases are often amalgamated. Combined databases are extensively used in multitasking systems that work with multi-condition modeling, such as feature extraction, speech styling, and language modelling[14].

One research investigates the ASR system implemented in the wild nature, which can recognize the voices of intruders. The system relies on multi-task learning (MTL) and has been effective in accent classification and speech recognition tasks. Based on MTL, the approach showed significant improvements in WER, ranging from 17.25 to 59.90, over single-task baseline models[15,16,17].

Based on the research analysis, three main systems for multitasking recognition can be identified. The initial system integrates various subsystems to overcome the challenge of running multiple recognizers, which can be computationally intensive[18,19,20]. The second system is designed with a recursive architecture, where a network of each task is trained using specific features, and the resulting output of one task is leveraged as supplementary resource data for different tasks to regulate the training of both tasks. The third system utilizes shared models and model parameters that are applicable to multiple tasks, with different classifiers being modeled at the output for each specific task. This includes a hierarchical multitasking model, where distinct classifiers are modeled at different layers[21,22,23].

Prepared work focuses on using a single model with common features and a classifier to perform speech neurorecognition in three tasks[24,25,26].

3. THE SYSTEM OF MULTI-TASK SPEECH NEURORECOGNITION

Our methodology is organized as follows:

3.1 Dialects of Kazakh language

Kazakh language is classified as an agglutinative language and is a member of the Turkic language family, along with languages like Kyrgyz, Turkish, and Uzbek. Dialects of the Kazakh language were formed due to territorial location and historical past. There are three main dialects in the Kazakh language: western, northeastern, and southern. There are dialects in which the words are pronounced differently but written the same way.

It is a known fact that the performance of identification or recognition tasks can degrade when dialects are mixed with standard Kazakh speech due to their unique characteristics and differences from standard Kazakh. Recognizing the dialects used in the Kazakh language is crucial for improving the quality of Kazakh speech recognition (see Table 1).

Table 1. Dialects of Kazakh language.

Dialect pronunciation	Dialect spelling	English translation
уақиға	оқиға	event
бәйге	бәйгі	race
палуан	балуан	wrestler
бөлегі	бөлігі	part of
әйтеу	әйтеуір	somehow
тұтқидан	тұтқиылдан	ambush
Алың	алғың	Take it
...	...	

The differences between the Kazakh dialects are mainly expressed in phonetics but are insignificant in vocabulary and grammar. The Kazakhs had no significant barriers to communicating through the written language[27,28,29].

3.2 Multi-task learning for Kazakh language

In our experiment, each task has a limited amount of labeled data, and the knowledge or features between tasks overlap, which is the basis of multi-task learning. In the considered problem, training an E2E multi-task speech recognition model consists of the following parts: expanding the sequence of Kazakh characters, dialect characters, and speaker identification as output targets. In the training, we use speaker and dialect identifiers[30,31,32].

In our approach, we paid attention to the influence of speaker parameters or dialect parameters on the accuracy and efficiency of recognition in specific sectors of speech[7, 32,33].

The Connectionist Temporal Classification model was additionally applied to improve the performance of the WaveNet model for the joint tasks dedicated to recognizing speech content and dialect identification. The label format used in this model remained the same but was also used for the joint tasks, which consist of the components, like recognition of speech content and speaker recognition, by replacing the identifier of the dialect with the identifier of the speaker. Figure 3 summarizes the E2E multi-task model based on the entire WaveNet-CTC for recognition of the Kazakh language.

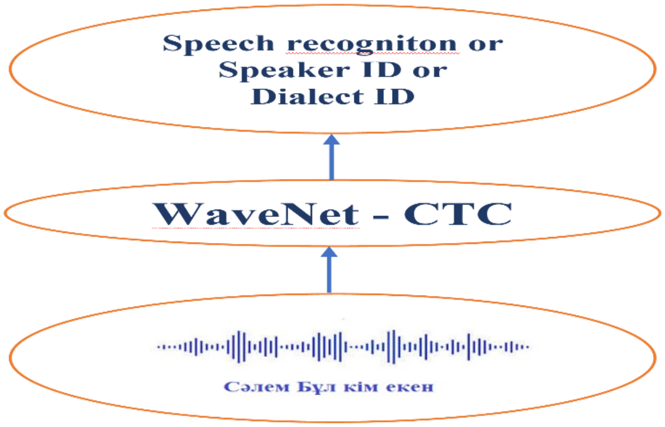

Figure 1. E2E structure based on WaveNet-CTC for simultaneous recognition of Kazakh speech tasks.

3.3 WaveNet-CTC Model

The WaveNet model addresses the challenge of obtaining distributed speech data. The model inputs a speech signal $=\{x_1,\ldots,x_N\}$ and the probability of producing label d for a dialect from the sequence of speech given during training.

If we apply automatic speech recognition to the speech signal *X*, we can obtain the corresponding text, which we'll refer to as *Text*(*X*). The speech signal *X* has *N* acoustic features. The Model predicts the distribution at each point in time, taking all previous predictions as input. We use dialect and speaker data during the training to improve recognition performance for different dialects for label output. The WaveNet model can be described as:

$$P(X) = \prod_{n=1}^{N} p\ (x_n |\ x_1, \dots, x_{n-1}) \tag{1}$$

In the scenario where the WaveNet model takes dialect speech as an input, the probability of identifying the dialect label *d* is denoted by (*d* | X;*DialectID*), where *DialectID* represents the model parameters for dialect identification. On the other hand, if the input is text, the probability of identifying the dialect label d is expressed as *P*(*d* | *Text*(*X*);*DialectID*). Here, *N* represents the number of speech's acoustic features.

The identification model has a defined objective function as:

$$\alpha(DialectID) =- \sum_{s=1}^{S} \log P(d^S | \boldsymbol{X}^S; DialectID) \tag{2}$$

where the variable *S* represents the total number of spoken sentences in the training dataset.

Causal convolutional layers are used in the WaveNet model to calculate probabilities of waveforms, which are based on the product of conditional probabilities by stacking building blocks. The convolutional network structure is presented in Fig. 2.

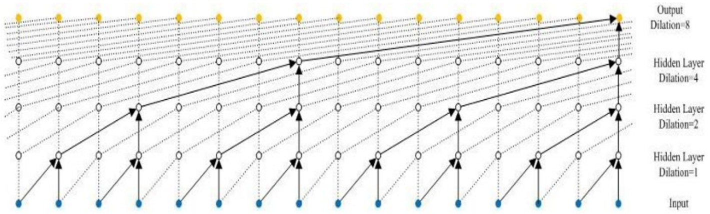

Figure 2. Figure Causal convolutional layers operation principle [8].

The WaveNet model comprises four convolutional layers: 1, 2, 4, and 8. Following the initial convolution, extended convolution layers extract features from the speech signal. Adding multiple convolutional blocks makes the receptive field size very large and crucial for context-sensitive sequential speech recognition tasks. The outputs of these layers are summed and further processed through a series of 1x1 convolutions and activations, culminating in a SoftMax level with 256 outputs.

The activation function of this model is written as follows:

$$z_j = \tanh\left(W_{f,j} * x_j\right) \odot \sigma\left(W_{g,j} * x_j\right) \tag{3}$$

where σ(*) - sigmoid function, j - layer index, f and g denote filters, W - trainable weight. The considered model uses parameterized skipping connections to enable much deeper models to be trained. To speed up the process of recognizing the dialect of the Kazakh language after the WaveNet model, the CTC (Connectionist temporal classification) model is used and the system architecture can be seen in Fig. 3.

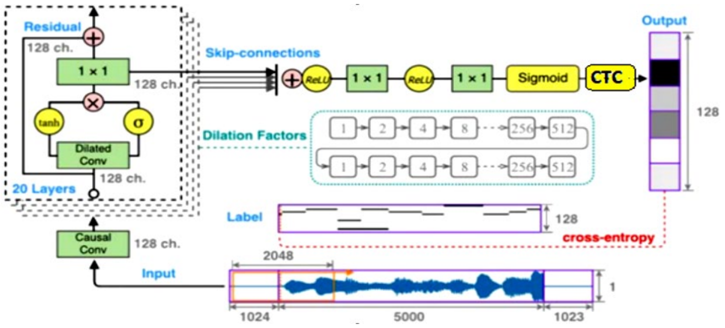

Figure 3. WaveNet-CTC architecture.

4. EXPERIMENTS AND RESULTS

Our experiment consists of two parts. The first part is speech recognition and speaker identification. The second part is dialect recognition. In the process of multi-task recognition, we use identifiers to identify the speaker and the dialect.

4.1 Data for the experiment

Our research team collected the experimental data for the study. For data collection, 53 speakers from different regions of Kazakhstan were involved to cover more dialects of the Kazakh language. The speech that the speaker will deliver consists of everyday Kazakh colloquial phrases. The collected speech and text data were generalized in an automatic recognition system. In the speakers' text, different colloquial speech dialects were used. Voice data is stored in WAV format with 16 kHz and 16 bits' characteristics.

Table 2. Experimental data.

Dialect	Data(hours)	Pronunciation	Number of speakers
West Region	3,11	1056	8
Northeast Region	5,22	3389	18
South Region	6,46	4657	27
Total	14,79	9102	53

According to Table 2. we used 90% of the available data to train our model, while 10% was reserved for testing. To extract features from the speech data, we used a 21 ms window and a 10 ms overlap, resulting in 27 Mel-frequency cepstral coefficients (MFCC) features for each observation frame.

4.2 Model's basic settings

In our experiments, WaveNet-CTC consists of 20 layers grouped into 4 extended stacks, each containing 5 layers of residual blocks. The original input in each layer has been added with the result taken from the residual block. In each stack for each layer, the expansion rate increases by 2 times. In the last layer, the expansion rate reaches the maximum value, increasing 16 times. Input layers (gating layers) have 128 hidden blocks. The number of hidden blocks in the input layers (gating layers) is 128. The learning rate for our experiments was three chosen as 2×10^{-3}.

The system includes two layers of LSTM, with each layer containing 250 hidden units. The system employs a SoftMax layer to categorize the speaker or dialect label and uses cross-entropy as the loss function. The weights of the SoftMax layer are randomly initialized from a uniform distribution with values ranging from 0 to 1. The multitasking model mainly takes into account the task of recognizing speech content. To improve recognition quality, we chose a more considerable weight (0.9 0.5) for the loss. In our case, recognition consists of three tasks, and for estimation we choose the following weights [0.9 0.04 0.04], [0.8 0.3 0.3], [0.7 0.25 0.25], [0.6 0.2 0.2], [0.5 0.01 0.01]. The gradient clipping technique is applied to ensure numerical stability during the training process by limiting the maximum value of the gradients to a fixed threshold, usually set to 1. This prevents the gradients from becoming too large and causing the training process to diverge. The training process of the WaveNet-CTC model is visualized in Figure 4.

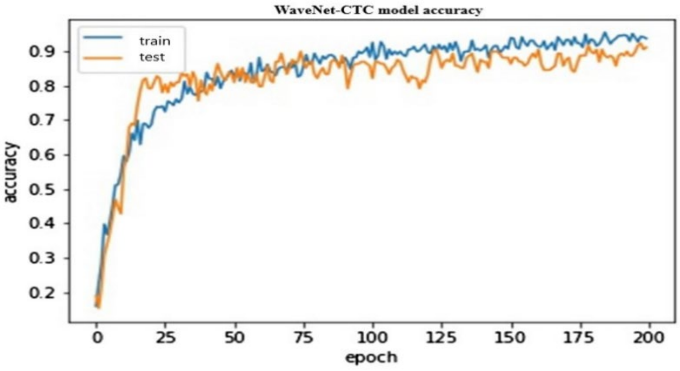

Figure 4. Process of training WaveNet-CTC model.

The quality of training is determined as follows:

$$ACC = \frac{\neq of\ correct\ files}{\neq\ of\ total\ files} \times 100\%. \tag{4}$$

To calculate ACC (accuracy), the highest probability of output data from the SoftMax layer of the multi-task recognition all models were trained using the graphical processor GeForce RTX3090 on the server AMD Ryzen9. 1000 GB of SSD memory was used to save the data during training.

4.3 Results of experiments

In the initial phase of our study, we trained a basic WaveNet model exclusively for dialect recognition. Subsequently, we proceeded to train a single model for the concurrent recognition of speech, speaker, and dialect.

During the second phase of our experiment, we conducted various assessments to gauge the effectiveness of our WaveNet-CTC model in accomplishing multiple tasks concurrently, such as speech recognition, speaker identification, and dialect identification. Table 3 summarizes the results of the experiment.

Table 3. Experimental data.

Architecture	Model	ACC
WaveNet	DialectID	96.61
	DialectID- SpeakerID-Speech	97.13
	DialectID-Speech-SpeakerID	97.56
	Speech-SpeakerID-DialectID	97.89
WaveNet-CTC	DialectID	96.72
	DialectID- SpeakerID-Speech	98.26
	DialectID-Speech-SpeakerID	98.83
	Speech-SpeakerID-DialectID	99.29

According to the results of the study, it can be seen that the simple WaveNet model for recognizing one task and three tasks shows the worst result. The result of the experiment shows in more detail that the recognition of speech content, the identification of the speaker, and the dialect with different training settings give different results. Despite other models with three tasks, the Speech-SpeakerID-DialectID model using WaveNet-CTC has demonstrated relatively better recognition quality than other models. The results have indicated that models based on the relevant features in section 4.2 effectively improve the recognition quality. Table 9 demonstrates that the WaveNet-CTC model achieves better results regarding speaker recognition compared to the joint loss-based and single-task models. This indicates that the multitasking mechanism is effective in improving speaker identification performance while also addressing the problem of imbalanced training data. The proposed WaveNet-CTC model effectively models dialect identification, speaker, and speech recognition. Compared to the single-task and three-task WaveNet models, the WaveNet-CTC model with three simultaneous tasks achieved higher accuracy in speaker recognition and dialect identification while experiencing a slight decrease in speech-to-text accuracy. The results indicate that when more tasks are trained together, the overall performance of the multitasking model improves due to the shared representation of features and model parameters, which enables better utilization of internal information between tasks while reducing model complexity. On the other hand, a model that relies on separate features for each task cannot take advantage of any possible connections between tasks to enhance its performance. Speech, speaker, and dialect recognition are similar tasks involving processing one input given as speech signals, but each has a different objective. They share some common features used as input for all three tasks. Training a single model to perform all three tasks simultaneously is analogous to how our brain processes speech signals - decoding the content and extracting other information, such as language, speaker characteristics, and emotions. This study demonstrates the effectiveness of such multitasking training on Kazakh language data and highlights the interrelated nature of these speech-processing tasks.

5. DISCUSSION

This article presents a multitask recognition approach that uses the E2E model to combine speech recognition, speaker identification, and dialect identification into a single neural network. The article compares the performance of this architecture with that of another multitasking model. The experiment results demonstrate that the proposed multitasking model can enhance the accuracy of speech content recognition and achieve good performance for speaker and dialect recognition. By leveraging the interdependence among tasks, the proposed approach avoids constructing separate models for each task, reducing the costs associated with model development and parameter tuning. Nevertheless, the interaction mechanism among the speech content recognizer, speaker recognizer, and dialect identifier needs more research and experimental approval.

6. CONCLUSION

This study presents a novel approach for multitask recognition of Kazakh language dialects applying an E2E model. The proposed model provides a simple and efficient way to develop a dialect model for the Kazakh language, eliminating the need for specialized resources such as pronunciation dictionaries. The proposed model in this study is designed for multitask recognition of Kazakh language dialects. It offers a simple and effective solution for building a Kazakh language dialect model without needing pronunciation dictionaries or other specific resources. The model is optimized to predict a sequence of Kazakh characters, dialect characters, and a speaker ID, which helps the model learn shared latent representations suitable for dialect, speaker, and syllable prediction. The experimental results indicate that the proposed multitasking approach enhances the performance of individual tasks and could apply to other languages. To improve the existing model, future research will incorporate an attention mechanism based on previous studies that have successfully implemented this approach to enhance end-to-end models.

REFERENCES

[1] Palaz, D., Magimai-Doss, M. and Collobert, R., "End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition," Speech Commun. 108, 15-32 (2019):

[2] Khassanov, Y., Mussakhojayeva, S., Mirzakhmetov, A., Adiyev, A., Nurpeiissov, M. and Varol, H.A., "A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline," Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, April 2021, 697-706 (2020).

[3] Li, B., Sainath, T.N., Sim, K.C., Bacchiani, M., Weinstein, E., Nguyen, P., Chen, Z., Wu, Y. and Rao K., "Multi-dialect speech recognition with a single sequence-to-sequence model," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Calgary, AB, Canada, Apr., 4749-4753 (2018).

[4] Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A. and Ochiai, T., "ESPnet: End-to-End Speech Processing Toolkit," Proc. Interspeech 2018, 2207-2211, DOI:10.21437/Interspeech, 2018-1456 (2018)

[5] Zou, W., Jiang, D., Zhao, S., & Li, X. "Comparable Study Of Modeling Units For End-To-End Mandarin Speech Recognition". 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), 369-373 (2018).

[6] Tang, Z., Li, L., and Wang, D., "Multi-task recurrent model for speech and speaker recognition," Computer Science: Sound, arXiv:1603.09643, 1-9 (2016).

[7] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K. "WaveNet: Agenerative model for raw audio," Computer Science: Sound, arXiv:1609.03499, 1-15 (2016).

[8] Mamyrbayev, O., Turdalyuly, M., Mekebayev, N., Alimhan, K., Kydyrbekova, A. and Turdalykyzy, T., "Automatic Recognition of Kazakh Speech Using Deep Neural Networks," Asian Conference on Intelligent Information and Database Systems, 07 March, 465-474 (2019).

[9] Tian, X., Zhang, J., Ma. Z., , He, Y., Wei, J., Wu, P., Situ, W., Li, S. and Zhang, Y., "Deep LSTM for large vocabulary continuous speech recognition", Computer Science: Computation and Languagea, rXiv:1703.07090, 1-8 (2017).

[10] Pan, Y. and Zhang, W.Q., "Multi-task learning based end-to-end speaker recognition,"" in Proceedings of the 2019 2nd International Conference on Signal Processing and Machine Learning, ser. SPML 19. New York, NY, USA: Association for Computing Machinery, 5661, 234-241 (2019).

[11] Imaizumi, R., Masumura, R., Shiota, S., & Kiya, H., "End-to-end Japanese Multi-dialect Speech Recognition and Dialect Identification with Multi-task Learning," APSIPA Transactions on Signal and Information Processing, 11(1), 341-349 (2022).

[12] Kim, S., Hori, T., and Watanabe, S., "Joint CTC-attention based End-to-end Speech Recognition using Multi-task Learning," in Proc. ICASSP, 4835–4839, (2017).

[13] Moriya, T., Ochiai, T., Karita, S., Sato, H., Tanaka, T., Ashihara, T., Masumura, R., Shinohara, Y., Delcroix, M., "Self-Distillation for Improving CTC-Transformer-Based ASR Systems," in Proc. INTERSPEECH, 546–50, (2020).

[14] Imaizumi, R., Masumura, R., Shiota S. and Kiya H., "End-to-end Japanese Multi-dialect Speech Recognition and Dialect Identification with Multi-task Learning", APSIPA Transactions on Signal and Information Processing: Vol. 11: No. 1, e4, (2022).

[15] Narisetty, C., Tsunoo, E., Chang, X., Kashiwagi, Y., Hentschel, M., and Watanabe, S., "Joint Speech Recognition and Audio Captioning," arXiv preprint arXiv:2202.01405, 1405-1413 (2022).

[16] Cai, L. and Zhao, C., "Method and implementation of endpoint detection in Ando Tibetan language," Gansu Sci. Technol., vol. 24, no. 5, 46-47 (2008).

[17] Han, Q. and Yu, H., "Research on speech recognition for Ando Tibetan based on HMM," Softw. Guide, vol. 09, no. 7, 173-175 (2010).

[18] Li, G., Yu, H., Zheng, T.F., Yan, J., Xu, S., "Free linguistic and speech resources for Tibetan," in Proc. AsiaPacic Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC), Kuala Lumpur, Malaysia, 733-736 (2017).

[19] Ruder, S., "An overview of multi-task learning in deep neural networks," arXiv:1706.05098, 17-25 (2017).

[20] Kisała, P., Kalizhanova, A., Kozbakova, A., Yeraliyeva, B., "Identification of cladding modes in SMF-28 fibers with TFBG structures," Metrology and Measurement Systems 30(3), 507–518, (2023).

[21] Wang, Q., Guo, W., Chen, P., and Song, Y., "Tibetan-Mandarin bilingual speech recognition based on end-to-end framework," in Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC), Kuala Lumpur, Malaysia, 1214-1217 (2017).

[22] Chen, D. and Mak, B., "Multitask learning of deep neural networks for low-resource speech recognition," IEEE/ACM Trans. Audio, Speech, Language Process., vol. 23, no. 7, pp. 1172-1183 (2015).

[23] Siohan, O. and Rybach, D., "Multitask learning and system combination for automatic speech recognition," in Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU), Scottsdale, AZ, USA, 589-595 (2015).

[24] Tang, Z., Li, L., and Wang, D., "Multi-task recurrent model for speech and speaker recognition," arXiv:1603.09643, 123-131 (2016).

[25] Qian, Y., Yin, M., You, Y., and Yu, K., "Multi-task joint-learning of deep neural networks for robust speech recognition," in Proc. IEEE Workshop Automatic Speech Recognition Understand (ASRU), 310-316 (2015).

[26] Krishna, K., Toshniwal, S. and Livescu, K., "Hierarchical multitask learning for CTC-based speech recognition," 543-551 (2018).

[27] Mamyrbayev, O., Oralbekova, D., Alimhan, K., Othman, M. and Turdalykyzy, T., "A study of transformer-based end-to-end speech recognition system for kazakh language," Scientific reports 12, 8337-8345 (2022).

[28] Mamyrbayev, O.Z., Oralbekova, D., Alimhan, K., and Nuranbayeva, B.M., "Hybrid end-to-end model for Kazakh speech recognition," International Journal of Speech Technology 26, 261–270 (2023).

[29] Oralbekova, D., Mamyrbayev, O., Othman, M., Alimhan, K., Zhumazhanov, B., Nuranbayeva, B., "Development of CRF and CTC Based End-To-End Kazakh Speech Recognition System. Intelligent Information and Database Systems," ACIIDS. Lecture Notes in Computer Science, vol 13757. Springer, Cham, (2022).

[30] Du, W., Maimaitiyiming, Y., Nijat, M., Li, L., Hamdulla, A., Wang, D., "Automatic Speech Recognition for Uyghur, Kazakh, and Kyrgyz: An Overview," Appl. Sci. 13, 326-339 (2023).

[31] Mukhamadiyev, A., Khujayarov, I., Djuraev, O., Cho, J., "Automatic Speech Recognition Method Based on Deep Learning Approaches for Uzbek Language," Sensors 22, 3683-3695 (2022).

[32] Ren, Z., Yolwas, N., Slamu, W., Cao, R., Wang, H., "Improving Hybrid CTC/Attention Architecture for Agglutinative Language Speech Recognition," Sensors 2022, 22, 7319-7329 (2022).

[33] Bisikalo, O., Kharchenko, V., Kovtun, V., Krak, I. and Pavlov, S., "Parameterization of the Stochastic Model for Evaluating Variable Small Data in the Shannon Entropy Basis. Entropy, 25, 1-18 (2023).