Speech to Text System for Noisy and Quiet Speech

A Kuzmin¹, S Ivanov¹

¹School of Electronic Engineering and Computer Science, South Ural State University (national research university), Chelyabinsk, Russia

E-mail: SazerLife@yandex.ru, saivanov@susu.ru

Abstract. This paper examines one of the available and simple methods to develop speech recognition systems capable of recognizing speech from noisy or silent recordings. Such systems improve the automated operation of call centers, and also bring us closer to creating speech recognition models capable of ignoring the speech deficiencies of speakers.

1. Introduction

For a long time, people have dreamed of having robots do all the routine work for them. Needless to say, today, we are surrounded by works of fiction showing the world of the future and people with access to technology that in former times would have been considered pure magic. Despite this, a good half of it is already available to us today. Much of this is due to machine learning and deep learning techniques. It opened the door for us to create systems capable of detecting human movement [1] and estimating traffic congestion [2]. Also, thanks to these technologies, new cybersecurity standards are being created [3].

In our age, computer science has become an integral part of everyday life. For a long time now, no one writes large texts by hand. Typewriters are relic of the past. Even typing on a computer using modern keyboards and all kinds of clever assistants no longer seem carefree enough. Contemporary people tirelessly invent new technologies designed to simplify their work and make their lives easier. One such technology is speech-to-text translation systems.

Speech-to-text translation systems [4] have multiple applications. They are used for the automatic insertion of subtitles and to make it easier for supervisors to evaluate the quality of voice calls. Many support call classifiers work on their basis, and they also enable voice assistants to understand what we are saying [5-6].

This paper will focus on a system capable of recognizing noisy and quiet speech recorded in low quality. Such systems are very important in call centres, as calls to them are made from different phones, whose dynamics record speech of different quality. As a result, in some audio recordings, the speech is accompanied by various kinds of noise, in some recordings, the voice of the person is pretty quiet, and sometimes the target voice is accompanied by the background communication of other people.

This paper is an attempt to create a system that solves the problem of recognizing noisy and quiet speech.
2. Related works
Speech-to-text systems are widespread in today's world. In paper [7] shows an automatic speech recognition system with word error rate (WER) of 27.2% when transcribing English voice messages by non-native speakers.

In this paper [8] shows the application of modern state-of-the-art automatic speech recognition models on the largest open Russian-language dataset Open_STT [9] with the best WER from phone calls of 33.5%.

The system presented in this paper builds on the work of other authors. First of all, we would like to mention the tool for working with speech recognition models – ESPnet [10]. The developers of this library have done a great job in simplifying the process of creating ASR models. It provides users with access to predefined implementations of popular speech recognition architectures, as well as trained models based on some of these architectures.

Russian-speaking models in ESPnet are presented based on a large open Russian speech corpus [9]. It includes more than ten thousand hours of speech, open for download to all researchers. This set contains the voices of numerous Russian speakers from different domain areas and even a separate corpus of addresses voiced by a synthesized voice.

3. System architecture
Usually, speech recognition systems consist of two important components represented by different neural networks. In particular, they are an acoustic model and a linguistic model. It consists of three key components: encoder, decoder and attention layer.

3.1. Acoustic model
The acoustic model is the heart of a speech recognition system. It usually represents by some type of "sequence to sequence" architecture. Its task is to convert the input sound, presented as a spectrogram, into a sequence of symbols of the respective language.

![Transformer architecture model](image)

**Figure 1.** Transformer architecture model.
In our work, sound is represented by spectrograms using the filter-bank provided by the Kaldi library \cite{11} on which ESPnet was built. Such a filter allows us to consider the peculiarities of human ear perception at different frequency levels. It is important because the model's process of "understanding" human speech is very similar to the way our ears do.

The acoustic model in our work presented by the transformer architecture \cite{12} shown in figure 1.

The encoder's role is to create an internal representation of the input sequence. The decoder's task is to translate this representation into a sequence of characters of the target language. A feature of this architecture is the attention layer. It allows the model to take into account the context of the time interval in question. Unlike recurrent neural networks, the context is built not only on previous occurrences but also on future occurrences.

3.2. Linguistic model
The linguistic model is an assistant model for a speech recognition system. Its task, given the context, is to correct errors and inaccuracies in the recognition results of the acoustic model. The linguistic model can be trained on a specific domain to improve the recognition quality of the final system on highly specialized data. In our work, the Linguistic Model represented by the LSTM network in figure 2.

![LSTM architecture model](image)

**Figure 2. LSTM architecture model.**

LSTM is a kind of recurrent neural network architecture \cite{13-14}. LSTM pretty well adapted to tasks in which important events separated by time lags of indefinite duration and boundaries. Its relative immunity to the duration of time gaps gives LSTM an advantage over alternative recurrent neural networks, hidden Markov models \cite{15}, and other training methods for sequences in various fields of application.

4. Implementation
To begin with, it is worth noting that all the data on which the experiments were carried out was provided to us by Intersvyaz company. These are the audio messages of the call centre subscribers with the corresponding transcripts. To train the linguistic model, the company also provided us with a
dataset of user's chat dialogues. All this data belongs to Intersvyaz company and was provided without disclosing personal data for research purposes only.

To implement a system capable of recognizing speech from noisy or quiet audio recordings was decided to use the pre-trained on data [9] acoustic model provided in the ESPnet library. The original model showed a recognition quality, measured by the WER metric, of 25.5%.

Our idea was that if we continued to train the acoustic model on noisy audio recordings, we could get a gain in recognition quality on them. By conducting the corresponding experiment, we obtained a WER quality metric of 24.6%. By performing some more experiments with frozen encoder and decoder layers, we were not able to get a significant gain in quality of recognition.

Then we decided to improve the system by adding a linguistic model to it. By training LSTM on a mixed dataset of audio transcripts and chat support calls, we were able to reduce the recognition error of the final system to 23.4%. Based on the measurements, the average time taken for a ten-second recording was 2 seconds. An Nvidia RTX 2060 video card was used for training and speed measurement.

5. Results
As a result, the best result of the model was 23.4%.

When reviewing the recognition results, we noted that the final system did indeed become better at handling words related to the subject of the calls. However, the decoding of words not related to it was still distorted by extraneous noise.

6. Conclusion
This article has demonstrated an approach to pre-training the ESPnet model. This method improved the performance of the original model, and also demonstrated higher quality than the state-of-the-art models trained on the same data.

7. References
[1] Golodov V, Zavei-Boroda A, Ivanov S and Nikolskaya K 2017 Development of a deep learning neural network for human movements analysis RPC 2017 - Proceedings of the 2nd Russian-Pacific Conference on Computer Technology and Applications vol 2017-Decem pp 72–4
[2] Kataev G, Varkentin V, Nikolskaia K 2020 Method to estimate pedestrian traffic using convolutional neural network Transportation Research Procedia 50 234-41
[3] Nikolskaia K, Minbaleev A 2020 Legal Support of Cybersecurity in the Field of Application of Artificial Intelligence Technology International Conference Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS) pp 59-62
[4] Graves A, Mohamed A-R, Hinton G 2013 Speech recognition with deep recurrent neural networks ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 6638947 6645-9
[5] Litvinov D M 2021 Speech analytics architecture for banking contact centers 10th Annual International Scientific and Practical Conference named after A. I. Kitov Information Technologies and Mathematical Methods in Economics and Management, IT and MM - CEUR Workshop Proceedings 2830-oct 2020 pp 227–39
[6] Bartz C, Herold T, Yang H and Meinel C 2017 Language identification using deep convolutional recurrent neural networks Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) vol 10639 LNCS pp 880–9
[7] Negrao M, Domingues P 2021 SpeechToText: An open-source software for automatic detection and transcription of voice recordings in digital forensics Forensic Science International: Digital Investigation 38
[8] Andrusenko A, Laptev A, Medennikov I 2020 Exploration of End-to-End ASR for OpenSTT – Russian Open Speech-to-Text Dataset Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) vol 12335 LNAI 35-44

[9] Slizhikova A, Veysov A, Nurtdinova D, Voronin D Open STT URL (Accessed 30 May 2021) https://github.com/snakers4/open_stt

[10] Watanabe S, Hori T, Karita S, Hayashi T, Nishitoba J, Unno Y, Soplin N E Y, Heymann J, Wiesner M, Chen N, Renduchintala A, Ochiai T 2018 ESPnet: End-to-End Speech Processing Toolkit Proc. Interspeech’18 pp. 2207-11

[11] Kaldi-asr 2019 URL (Accessed 30 May 2021) https://kaldi-asr.org/

[12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I Attention Is All You Need arXiv preprint arXiv:1706.03762

[13] Hochreiter S and Schmidhuber J 1997 Long Short-Term Memory Neural Comput. 9 1735–80

[14] Cheng J, Dong L, Lapata M 2016 Long short-term memory-networks for machine reading arXiv preprint arXiv:1601.06733

[15] Bahl L R, Brown P F, de Souza, Peter V, Mercer, Robert L Speech recognition with continuous-parameter hidden Markov models 1988 ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings pp 40-43

Acknowledgments
The authors of this article would like to thank Intersvyaz Company for providing a recording of their speaker.

The work was supported by the Ministry of Science and Higher Education of the Russian Federation (government order FENU-2020-0022).