About the possibility of using speech recognition technologies in problems of assessing the protection of acoustic information from the leakage through technical channels

A V Ivanov¹, Sh R Salimov²
¹Novosibirsk State Technical University, 20 Pr. K. Marksa, Novosibirsk, 630073, Russia
²Samara State Technical University, 244 Molodogvardeyskaya street, Samara, 443001, Russia
E-mail: shamil.salimov@mail.ru, andrej.ivanov@corp.nstu.ru

Abstract. This article discusses the possibility of improving the existing instrumental calculation method for assessing the security of meeting rooms by moving from the Pokrovsky formant method to using speech recognition algorithms based on deep neural networks. In our case, it is proposed to use recurrent neural networks that have shown themselves best in terms of information processing with low SNR ratios. Also, using the long-term short-term memory of a recurrent neural network, it is proposed to improve the test signal. In addition, it substantiates the need to take into account the individual characteristics of the room and the possibility of using interference, such as a “speech choir”.

1. Introduction
Nowadays, the information security of the enterprise is one of the leading factors in its effective development. Possible losses from information leakage can cause a specific amount of economic damage.

According to the statistics, about 90% of all necessary information about business rivals can be obtained by legal methods. As a rule, they try to get the most valuable and carefully guarded information by resorting to industrial espionage. The enterprise which is interested in maintaining its trade secrets has to use special rooms for negotiations.

In order to bring these premises into compliance with the requirements for the leakage protection via vibro-acoustic channels, they are guided by the instrumental-calculation method for evaluating speech intelligibility. This method was developed by A. A. Khorev, Yu. K. Makarov, and V. K. Zheleznyak, [1], their research was based on the results of experimental studies designed by N. B. Pokrovsky [2]. Speech intelligibility is some integral assessment of a speech signal and, according to the international standard ISO/TR 4870, it is defined as “the degree to which speech can be understood (transcribed) by listeners”.

It should be noted that the Pokrovsky’s formant method [2] is focused on assessing the quality lines of communication, and the conditions of experiments differ significantly from the actual abilities of the intruder in the field of information security. According to this, we can distinguish a number of disadvantages of this approach for assessing the security of emplacements:

− the conditions under which the results of measuring speech intelligibility were obtained in the method of N. B. Pokrovsky (the distance to the microphone was 0.08 meters) [2] significantly differ from the conditions which are described in the methodology for assessing security (distance to the microphone is 1 meter)[1];
− the choice of an octave partition of the frequency range, as well as the width of the investigated frequency range, is not justified or experimentally confirmed;
− in the experiments on the basis of which the basic dependencies of the method were built, was not considered the possibility of repeatedly listening to a noisy recording, the noise levels were much lower, the possibility of using noise cleaning was not also taken into account;
articulation’s tests were made by using the tables (syllable, word, etc.) elements in which maximum uncorrelated each other, but we study only coherent, meaningful texts;
- the effect of forced speech was not taken into account;
- the individual speakers’ characteristics are not taken into account, everything is based on the average spectrum of speech;
- masking is taken into account only with the use of interference which based on “white noise”, the ability to evaluate security using interference of the “speech choir” type is excluded.

Many of these issues are analyzed in detail and considered in the following papers [3].

2. Formulation of the problem
Thus, the main task is to change the approach of assessing security by moving from the formant method to the use of speech recognition algorithms.
If the proposed approach is implemented, we should take into account:
- the use of masking interference based on the "speech choir", as well as formulate requirements for their creation;
- individual speakers’ characteristics, by creating the test signal which is based on various voices, covering various options in terms of spectral characteristics, formant distribution, amplitude composition of speech;
- the influence of the building’s and enclosure’s characteristics on a speech signal.

3. Theory
At this part of our research we consider the widely developing nowadays methods of speech recognition, based on machine learning. Today, there are five basic methods of machine learning: each of which has its own strengths and weaknesses, each is designed for its own specific purposes [4].

![Machine Learning Diagram](image)

Figure 1. The main types of machine learning.

Since 2006, deep learning of machine studies has become widespread and has been included in hundreds of researches, many scientists from around the world have worked on this topic [5, 7, 8]. Areas in which deep learning was included range from data processing to artificial intelligence. There is the hierarchy of functions, so that high-level functions are defined in terms of low-level functions, and therefore this is called deep architecture. Typically, of the models that are included in this class are based on uncontrolled learning ideas [9].

In our opinion, in-depth training is most applicable to our task of creating a test signal and generate a stray interference. It was proven itself very well when working with acoustic information and in tasks where speech recognition is required, especially for large-scale tasks [8]. In addition to the improved training procedures, the main factors were an increase in the computing power of the technology that contributed to the recent successes studies of deep neural networks (DNNs), the availability of more training data, and better software development.
Into the interior of approx second neural network (DNN), which first demonstrated significant improvements compared to the Gaussian mixture model (GMM) for acoustic modeling, researchers Y. Gong, R. Haeb – Umbach, of L. Deng, and the J . Li in [5] describes the main advantages of DNN:

- it works better for tasks of recognition of noisy speech, in comparison with the other methods of machine learning;
- methods of machine learning , depending on the speaker, not much better than independent of the DNN’s speaker;
- DNNs work much better for acoustic modeling if we use one or more convolutional stratum, giving some is invariant to the differences in the voice path between the speakers;
- can be used not only for acoustic modeling;
- DNN architecture can be used for multi-tasking training in several different ways, and is much more effective as well as the GMM, to use the data of a problem for performance related problems.

The term “deep learning” existing today characterizes a wide range of methods, one way or another related to machine learning, as well as architectures based on the use of many levels of non-linear information processing [10]. There are many different deep learning algorithms, but recurrent neural networks and convolutional neural networks have gained much more popularity.

The convolutional neural network (CNN) are a kind of deep discrimination architecture, in which each model comprises the most accurate level and combining stratum which situated above each other. Many weights are used in the convolutional stratum, and the combining stratum, on the other hand, it selects the inference resulting from the convolutional stratum and reduces the data rate of the lower stratum. The distribution of weight, together with correctly selected union schemes, leads to CNN invariance properties. Some argue that the small invariance observed in CNN does not allow using it for solving complex problems of pattern recognition. Nevertheless, CNN have shown their efficacy when it is used in tasks of computer vision or image recognition. In addition, CNN can be tried to be used in speech recognition tasks, if you apply some changes in the method designed for image analysis.

Recurrent Neural Network (RNN) is such a deep networks wich is used in unsupervised learning, in cases when the number of input data stratum in the network can be comparable, or even equal to its length [6]. RNNs are developed by using the same set of weights recursively along the branching structure, and the branches are traversed in topological order from start to end. RNN is used primarily to predict the future sequence of data through the usage of previous data samples. Recursive neural networks are very effective when it comes to simulation of sequence data, such as text or speech. However, until nowadays, these networks were not widely used, since it is believed that it is extremely difficult to train to cover long-term dependencies. In recent years, through the work of J. Martens managed to overcome this limitation, including through the use of approximate second-order information or stochastic estimates of curvature. In some newly published researches of RNN [11], trained in the technique of free Hessian optimization, they demonstrated that they are able to generate consecutive text characters.

A good example of the study of recurrent neural networks, in terms of noise removal, is the work of L.-R. Dai, C.-H. Lee, L. Sun, and J. Du [12]. The authors propose the recurrent neural networks with short-term memory (LSTM - RNN) for noise removal and speech enhancement. The method proposed in [22] is intended to extract basic pure speech from the observed noise speech signal. In conventional algorithms, the same as in the spectral subtraction and based on the MMSE estimate crate spectral amplitude, the settings are not controlled and based on mathematical assumptions about the speech and noise , is often necessary to observe artifacts (eg, background noise) and limited performance in enhanced speech.

In [13] proposed LSTM – RNN, wich was compared with the DNN approach, in the problems of reducing the effect of noise on speech recognition at a low signal-to-noise ratio (SNR).

Authors, L.-R. Dai, C.-H. Lee, L. Sun, and J. Du offers small revived ensemble structure with a multi-purpose co-ed, in order to fully use the capabilities of the set of learning objectives and consistently improve objective measures as the voice quality and intelligibility for invisible noise scenarios. A single jointly studied LSTM model can provide comparable performance with several LSTM - RNN ensembles with less computational complexity and significantly less model size.

The algorithm’s feature and model LSTM, the network able to dynamically evaluate information that you want to update, save, and print. Thus, we can effectively use temporary information. In addition, LSTM responds well to some features of speech and noise, which is good for the subsequent separation operation, especially for non-stationary noise.
As a rule, experiments are performed on the TIMIT database [14]. In [12], authors used 1,115 types of noise during the training phase. Statements of the training set TIMIT were distorted with each type of noise on six different levels the SNR, at 20dB to -5dB increments 5 dB. The authors aimed to improve voice recognition by reducing the effect of noise using the trained LSTM method.

Three sets of test data were used to assess the capabilities of acoustic models.

1) “Pure”: it belongs to the field of pure training data (~ 41 k statements).
2) “Noisy”: it is obtained from a “clean” data set by the same modeling method, but different sources of multimedia noise are selected, and the reverberation time is taken from another interval (520, 920) ms.
3) “realistic”: it is collected in a real room (~ 2 k statements). Pure speech and multimedia noise are reproduced by speakers and recorded by several microphones located at different points in the room.

Accuracy of recognition is improved with increasing amount of input data that is approximately 800 hours. However, the authors noticed that the minimum error value (WER) is achieved by using approximately 4800 hours of aces (800 hours of input, repeated 6 times), and then begins to increase it again. Most likely, this is due to, perhaps the first overloaded viscous system repetitive noise (fig. 2). According to the graphs in fig. 2, it can be observed that the algorithm works very well in an environment with static noise, since the graph is as even as possible, which suggests that the use of speech-like interference (as an analogue of realistic test data) based on spurious noise and it should be more effective than the use of white noise.

![Figure 2. Graphs of the dependence of the percentage of errors on the number of repetitions in LSTM.](image)

The article [12] compares the quality scores of continuous speech PESQ (speech quality) and short-term objective intelligibility - STOI (speech intelligibility) for different systems in the test set. Based on their research, it follows that the basic DNN-DM algorithm based on DNN is not very effective for speech recognition in a "noisy" environment, an ensemble of various LSTM (Multiple Models) models outperforms it in all indicators at all SNR levels, which indicates its effectiveness and is the most advantageous in terms of results, especially with low SNR. However, this ensemble requires significant computing power from the technology on which the research is used.

Alternatively, L.-R. Dai, C.-H. Lee, L. Sun, and J. Du uses the results of the LSTM-RN approach to multipurpose learning, which is able, while saving about 40% of the computing power of the technique, to give results very close to the ensemble, which is the most acceptable recognition method for research.

Machine learning methods for speech recognition do not work well enough at low signal-to-noise ratios yet, but there are good prospects for methods using deep neural networks, especially in recurrent neural networks (taking into account the relative speed with good recognition rates).

Thus, at best with both standard teaching and the learner model, the end learner can get a relative error reduction (WER) of about 10% with pure learning data, 29% when adding various media noise systems to the raw data, and 20% when recognizing under conditions in which test cases are close to realistic.
4. Conclusions
This feature of deep neural network methods can be applied in information protection problems using
not arbitrary speech as a test signal, but a test signal is known in advance from our recognition
algorithm, formed from the speech of several speakers covering differences in spectral characteristics,
formant distribution and amplitude composition of speech.
It is also necessary to develop requirements for the generation of interference based on the “speech
choir”, masking them most productively from recognizing intercepted information both by a person
and using the methods described above, while at the same time introducing as little discomfort as
possible to people working indoors.
Thus, the methodology for assessing security requires improvement. It should be based on modern
speech recognition algorithms, take into account the individual characteristics of the premises, the
speech of the speakers, as well as the possibility of applying interference based on the “speech choir”.

5. References
[1] Zhelezniak I K, Makarov Yu K and Horev A A 2000 Some methodical approaches to the
assessment of the effectiveness of the protection of the speech information (M. Special
equipment. No4) pp 39–45
[2] Pokrovsky N B 1962 Calculation and measurement of speech intelligibility (M.: Svyazizdat)
[3] Ivanov A V, Reva I L, Trushin V A and Tudevdagva U 2014 Correction of the methodology for
assessing the protection of speech information from leakage through technical channels in
conditions of speech forcing (Scientific Herald of the Novosibirsk State Technical University.
No2 (55)) pp 183–189
[4] Nassif A B, Shahin I, Attili I, Azzeh M and Shaalan Kh 2019 Speech Recognition Using Deep
Neural Networks: A Systematic Review (IEEE Access-February)
[5] Li J, Deng L, Haeb-Umbach R and Gong Y 2016 Robust automatic speech recognition. A
bridge to practical applications (Academic Press is an imprint of Elsevier)
[6] Deng L 2014 A tutorial survey of architectures, algorithms, and applications for deep learning
(APSIPA Trans. Signal Inf. Process., vol. 3, no. e2) pp 1–29
[7] Dahl G, Mohamed A and Hinton G 2009 Deep belief networks for phone recognition (in Proc.
NIPS Workshop Deep Learning for Speech Recognition and Related Applications)
[8] Hinton G and Salakhutdinov R 2006 Reducing the dimensionality of data with neural networks
(Science, Vol. 313. №5786) pp 504 - 507
[9] Cho Y and Saul L K 2009 Kernel methods for deep learning (in Proc. Adv. Neural Inf. Process.
Syst. (NIPS), vol. 22) pp 342–350
[10] LeCun Y, Bengio Y and Hinton G 2015 Deep learning (Nature vol. 521, no. 7553) p 436
[11] Martens J and Sutskever I 2011 Learning recurrent neural networks with Hessian-free
optimization (in Proc. 28th Int. Conf. Mach. Learn. (ICML)) pp 1033–1040
[12] Dai L-R, Lee C-H, Du J and Sun L 2017 Multipurpose deep training for improvement of speech
on the basis of LSTM – RNN (in Proc. Hands-Free Speech Commun. Microphone grating
(HSCMA)) pp 136-140
[13] Weninger F, Hershey J R, Roux J L, and Schuller B 2014 Discriminatively trained recurrent
neural networks for single-channel speech separation (IEEE Global Conference on Signal and
Information Processing (GlobalSIP) pp 577–58
[14] Garofolo J 1988 Getting Started with the DARPA TIMIT (CD-ROM: An Acoustic Phonetic
Continuous Speech Database)