Transfer Learning for Domain and Environment Adaptation in Serbian ASR

Branislav Z. Popović, *Member, IEEE,* Edvin T. Pakoci, and Darko J. Pekar

***Abstract* — In automatic speech recognition systems, the training data used for system development and the data actually obtained from the users of the system sometimes significantly differ in practice. However, other, more similar data may be available. Transfer learning can help to exploit such similar data for training in order to boost the automatic speech recognizer's performance for a certain domain. This paper presents a few applications of transfer learning in the context of speech recognition, specifically for the Serbian language. Several methods are proposed, with the goal of optimizing system performance on a specific part of the existing speech database for Serbian, or in a noisy environment. The experimental results evaluated on a test set from the desired domain show significant improvement in both word error rate and character error rate.**

***Keywords* — Automatic speech recognition, Kaldi speech recognition toolkit, transfer learning, noise adaptation, Serbian.**

I. Introduction

AUTOMATIC speech recognition (ASR) is a technology that allows computers to convert spoken words into text, i.e., to transcribe what has been said. It has many contemporary applications in areas that involve communication between humans and machines. These applications include dictation (automatic transcription) systems, voice assistant applications for smartphones, various smart home uses, automated call centers, as well as an array of tools for aiding people with certain disabilities.

Paper received June 25, 2020; revised September 28, 2020; accepted September 30, 2020. Date of publication December 25, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dragana Šumarac Pavlović.

This paper is revised and expanded version of the paper presented at the 27th Telecommunications Forum TELFOR 2019 [16].

This research was supported by the Science Fund of the Republic of Serbia, #6524560, AI – S-ADAPT, and by the Serbian Ministry of Education, Science and Technological Development through the project no. 451 03-68/2020-14/200156: "Innovative scientific and artistic research from the Faculty of Technical Sciences activity domain".

Branislav Z. Popović is with the Department for Power, Electronic and Telecommunication Engineering, Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia, Department for Music Production and Sound Design, Academy of Arts, Alfa BK University, Nemanjina 28, 11000 Belgrade, Serbia, and Computer Programming Agency Code85 Odžaci, Železnička 51, 25250 Odžaci, Serbia (phone: +381613207989; e-mail: bpopovic@uns.ac.rs).

Corresponding author Edvin T. Pakoci is with AlfaNum Speech Technologies, Bulevar vojvode Stepe 40, 21000 Novi Sad, Serbia (e-mail: edvin.pakoci@alfanum.co.rs).

Darko J. Pekar is with AlfaNum Speech Technologies, Bulevar vojvode Stepe 40, 21000 Novi Sad, Serbia (e:mail: darko.pekar@alfanum.co.rs).

In the past several years, many advances were reported in ASR research for the Serbian language. Most recent research is directed towards language modeling, as the inflectivity of Serbian posed many problems for large vocabulary ASR systems. Different recurrent neural network based language models were trained and tested, as well as variants that use embedding vectors as word representations and incorporate additional lexical and morphological features [1] - [2]. On the other hand, the latest acoustic models involved purely sequence-trained deep neural networks with subsampling, specifically designed to better model longer temporal contexts [3] - [4]. These models included accent-specific vowel models, Mel-frequency cepstral coefficients (MFCCs), pitch features and i-vectors [5] for the purpose of adaptation to different speakers and channels. There were also experiments using end-to-end architectures, but they did not provide improvements in error rates [6].

Speech database expansion is the basic way to upgrade acoustic models. This can be done by preparing and adding speech data for more speakers, more utterance types or different environments. However, database expansion can also be performed artificially in several ways – by modifying existing spoken data using speech speed and pitch manipulation, utilizing waveform scaling (volume manipulation), or by adding some amount of noise (artificial or real-life recorded noise) to the existing database. All of these methods have already been examined for Serbian (see Sect. 3 for details on what kind of expansion was used for experiments in this paper) [7].

Another way to create more robust acoustic models, especially for more specific domains, is transfer learning. Transfer learning is a machine learning method where a model developed for one task (e.g. large vocabulary continuous speech recognition in a general setting) is reused as the starting point for a model on another task (e.g. ASR for a given domain or speech style, different emotions, different channel or environment, or even a completely different language) [8]. It is also referred to as domain adaptation. The resulting models should show improved performance on the second task. The condition that the models trained for the first task are general enough is of most importance. Transfer learning is particularly useful when there is a limited amount of data for the second task [8] - [9]. For all the various types of transfer learning techniques, deep neural networks that have many hidden layers and are trained using contemporary methods have been shown to suit the best [10]. An illustrative diagram of the notion of transfer learning in general is displayed in Fig. 1.

TABLE 1: TRAINING PARAMETERS FOR THE BASELINE AND TRANSFER LEARNING (TL) MODELS (BOTH EXPERIMENTS).

Experiment	Training	#layers	#neurons	Layer splicing	#epochs (iterations)	Primary LR factor	Output LR factor
Exp. 1	Baseline	8	625	3+4	4 (247)	-	-
	TL +3L	8+3		3+7	+4 (80)	0.25	1.00
	TL WT	8		3+4	+4 (80)	0.25	1.00
				3+4	+4 (80)	0.50	0.50
Exp. 2	Baseline	10	1024	4+5	5 (2235)	-	-
	TL WT N				+4 (7)	0.25	1.00
	TL WT SN				+4 (164)		

In this paper, a transfer learning method is examined where the additional acoustic model training is performed on a smaller part of the larger database, which includes utterances typical for ASR systems that are based on command-response human-machine interaction (HMI). As a result, the final system should have significantly higher word recognition rates on utterances of that type and be more robust in interactions from the desired domain. This is especially important when the desired use on smartphones and similar platforms is taken into account [11] – here, the small amount of available in-domain data is paired with the goal of having somewhat simpler neural network models for the generally less powerful hardware of mobile phones (therefore simply increasing the number of hidden layers in the neural network or the number of neurons per layer is not the preferred method for improving the recognition rates).

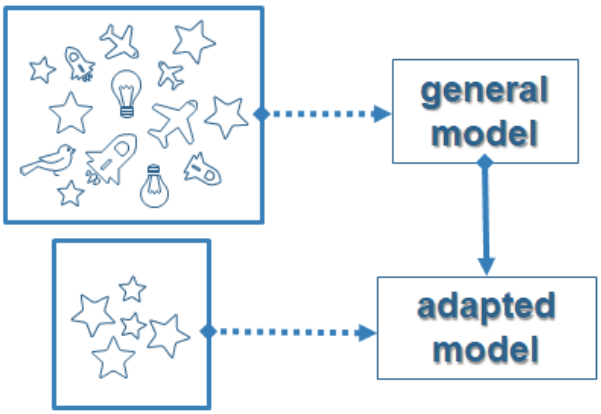

Fig. 1. Transfer learning diagram.

Another transfer learning application is also examined in the paper – model adaptation to a noisy environment. Here, the general ASR model is tuned using recordings of pure noise, or a combination of clean speech and pure noise for various environmental settings.

The remainder of this paper is organized as follows. Section II discusses the experimental setup and the applied training methods. In Section III, all of the used speech databases are briefly introduced – both the whole database and the HMI part for tuning, as well as databases with noise. The experimental results are then presented in Section IV. Finally, conclusions are drawn and future work is pointed out in Section V.

II. EXPERIMENTAL SETUP

The baseline model in experiment 1 is a so-called "chain" sub-sampled time-delay deep neural network (TDNN), trained using cross-entropy training and a sequence-level objective function [3], [4], [7]. The training procedure consists of the pre-DNN and the DNN phase. The pre-DNN phase involves the extraction of features from short frames of audio signals (static features including 14 MFCCs, energy and 3 pitch-related features – probability of voicing, log-pitch and delta-pitch, as well as dynamic features representing the first and second derivatives of the static features – the final feature vector is 54-dimensional), initial flat-start monophone HMM-GMM training, triphone HMM-GMM training (targeting 3500 HMM states and 35000 Gaussians), as well as speaker adaptive training (SAT, targeting the same model complexity), where the possibility of model adaptation based on maximum likelihood linear regression (MLLR) for individual speakers is introduced [12]. The final pre-DNN HMM-GMM model (the SAT model) is then used to provide input data alignments for the deep neural network (DNN) training. The DNN phase uses 40 high-resolution MFCCs as features, calculated on frames with duration of 30 ms and time shift of 10 ms between the adjacent frames, as well as the three previously described pitch-based features, and finally a 100 dimensional speaker identity vector, or i-vector, producing a 143 dimensional feature vector as input.

The neural network in experiment 1 consists of eight hidden layers, each containing 625 neurons. The lower TDNN layers are trained using temporal context windows that include the preceding, the current and the following frame (it is said that they are spliced in a {-1, 0, 1} manner), while the training of higher layers is performed using windows of three frames as well, but with 3-frame-long gaps between them (spliced in a {-3, 0, 3} manner). For example, if the three initial layers were spliced in a {-1, 0, 1} manner, and the other four layers (4-7) in a {-3, 0, 3} manner, that is denoted as "3+4" in Table 1. The acoustic model was trained using the widely used Kaldi speech recognition toolkit. The network is trained for 4 epochs, and the total number of training iterations is, as always, determined by the quantity of available training data [3].

The DNN in the second experiment differs in complexity – it has 10 hidden layers, 1024 neurons each. Other parameters are the same. It was trained for 5 epochs (2235 iterations). The database it was trained on also differs (see Sect. 3 for details).

The language model is a 3-gram ARPA language model, trained using the SRILM toolkit [13], the Kneser-Ney smoothing method and an *n*-gram pruning parameter of 10^{-7}, which have provided optimal results in previous

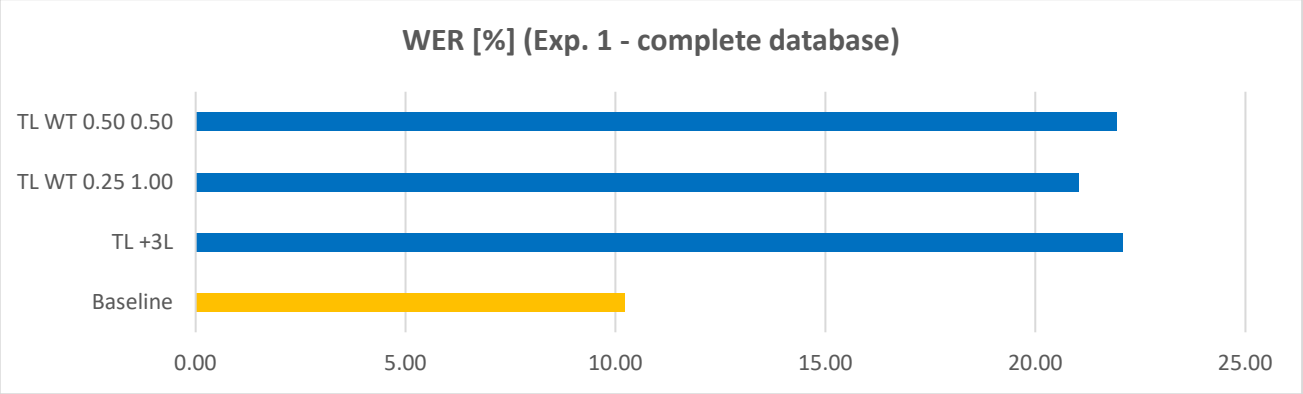

Fig. 2. Word error rate for baseline and transfer learning models ("complete" database).

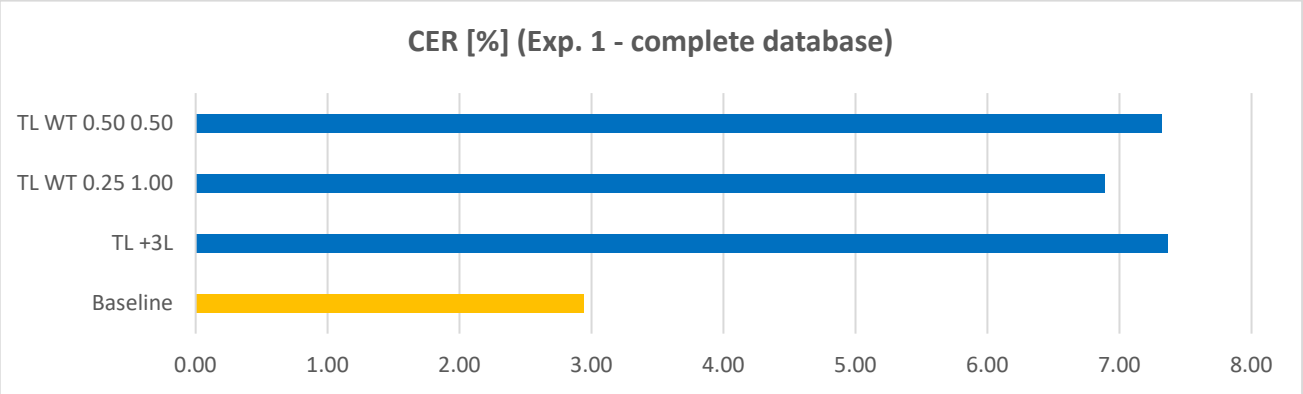

Fig. 3. Character error rate for baseline and transfer learning models ("complete" database).

research [14]. The procedure resulted in 249809 unigrams, about 1.87 million bigrams and 551 thousand trigrams.

For experiment 1, in order to retain the acoustic variability and knowledge provided by the baseline model trained on the "complete" database (see Sect. 3), and at the same time adapt the model for the specific HMI domain, three different transfer learning configurations have been examined. In the first setup, 3 additional layers are added to the original neural network configuration ("TL +3L"). The transferred layers are trained with a smaller learning rate (LR) factor of 0.25 (denoted as primary LR factor in Table 1; the values are given in the range of 0 to 1, 0 meaning that the transferred parameters are immutable). The learning rate factor for the additional layers is set to 1. The network is trained for 4 additional epochs (80 iterations based on provided data).

In the second setup, weight transfer approach was applied ("TL WT"), i.e., the network is pre-trained using the "complete" database, and then fine-tuned using parts of the database from the specific domain [15]. The LR factor for transferred layers was set to 0.25, while the last layer (output layer) was trained with the LR factor of 1. For the third setup, a similar configuration has been used, the difference being in the primary learning rate factor, which is set to 0.5 (faster tuning), as well as in the LR factor for the output layer, which is set to the same value (0.5).

For experiment 2, only the weight transfer approach was used, with LR factors of 0.25 (primary) and 1 (output). The tuning was performed using only pure noise recordings for the first setup (experiment 2A – "TL WT N"), or a combination of regular spoken data and noise recordings for the second setup (experiment 2B – "TL WT SN"). In both cases, the tuning lasted for 4 additional epochs, but of course, there were many more iterations in the second case (a lot more training data). All of the parameters of the above-mentioned setups can be observed in Table 1.

III. Database Description

A. Exp. 1: Complete Database

The "complete" database in experiment 1 is the recently expanded speech database for the Serbian language [7]. The database consists of audio book recordings (recorded in a studio environment, spoken by professional speakers, 32 male and 64 original female speakers, 168 hours of data in total), radio talk show recordings (179 hours of data, 21 male and 14 female speakers), and mobile phone HMI recordings (61 hours in total, 169 male and 181 female speakers). Audio data is sampled at 16 kHz, 16 bits per sample, mono PCM. For the first and second part of the database (books and talk shows), the original data for each speaker (for speakers with enough data) is separated into chunks of 30-35 minutes at most. Various combinations of speech speed and pitch modifications have been applied to each of these chunks in order to equalize the amount of audio data per speaker, resulting in a certain number of mutually distinct sub-speakers (398 and 420 distinct sub-speakers for audio books and radio shows, respectively). A different amount of noise is also added for the baseline model training – the resulting database included the "pure" database (recordings without artificially added noise), SNR 11 database (with added noise and signal-to-noise ratio of 11 dB), SNR 13 database (signal-to-noise ratio of 13 dB)

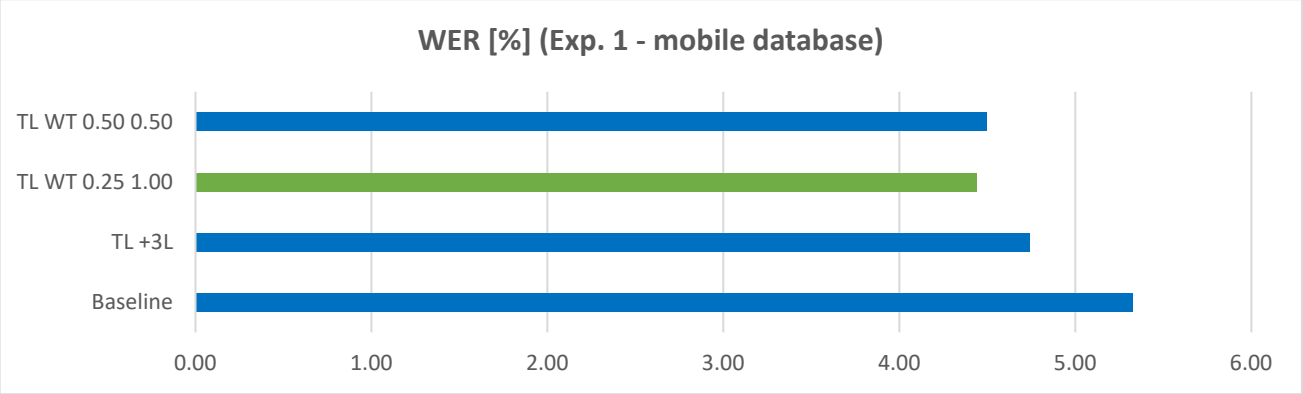

Fig. 4. Word error rate for baseline and transfer learning models (mobile HMI database).

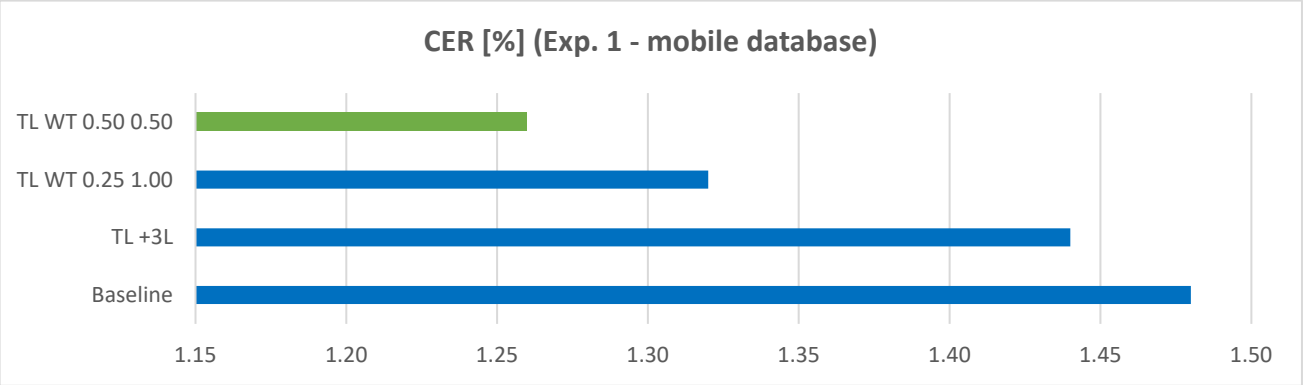

Fig. 5. Character error rate for baseline and transfer learning models (mobile HMI database).

and SNR 15 database (signal-to-noise ratio of 15 dB), as well as recordings with pure noise. These noise recordings varied in type – traffic noises, "cocktail party" noises, construction noises, wind noises, etc. For testing purposes, 29 hours of audio data was extracted from 81 sub-speakers (which were excluded from the training set in their entirety).

B. Exp. 1: Transfer Learning Database

The Serbian "mobile" speech database (the third part of the "complete" database) consists of mobile phone recordings – these include various commands, questions, numbers, dates, locations, and some other inquiry-based utterances, which can be expected in an interaction with a voice assistant type application on a smartphone. The utterances are freely spoken, and are a lot shorter in comparison to the other two database parts. The vocabulary is highly domain-oriented and much smaller (consisting of less than 4000 different words). This time, the "pure" database and the SNR 11 database (this was determined to be the usual amount of noise to be expected in this type of communication), as well as the pure noise recordings, were used for DNN tuning.

C. Exp. 2: Complete Database

The "complete" database in the second experiment included the "pure" part (with no added noise) of the "complete" database from the first experiment, as well as additional Croatian audio data (due to the similarity of the two languages this is possible) [7]. The Croatian data also includes audio books, radio talk shows and "mobile" HMI recordings, which added up to around 535 hours (therefore doubling the original amount of data), separated into 897 and 845 distinct male and female sub-speakers, respectively. On top of that, noised recordings (from both the Serbian and Croatian parts) with a SNR of 17 dB were added. The training data was finally tripled using speed perturbation, i.e., by using added speed-up and slowed-down speech (by 15% each). The test database is the same as in experiment 1, but in some tests noise was added with a SNR of 9 dB or 13 dB.

D. Exp. 2: Transfer Learning Databases

For the first setup (exp. 2A), only recordings with pure noise were used for model tuning (the same recordings mentioned in Sec. 3A). In total, there is around 7.5 hours of them. This database was multiplied 64 times, to be comparable to the "complete" database in length.

For the second setup (exp. 2B), a part of the "complete" database was used – the whole Serbian "pure" unmodified part (408 hours), plus additional 174 hours of Serbian audio books and radio show recordings, totaling to about 582 hours. This database was then randomly sampled, so about a fifth of the audio files were actually used for transfer learning (about 116 hours). Finally, the pure noise database multiplied 8 times was added to this set, achieving a 2:1 ratio between spoken and noise data.

IV. Experimental Results

The results of experiment 1 are given in Figs. 2 to 5. In Fig. 2 and 3, the results are given for the "complete" database – the baseline and transfer learning configurations in terms of word error rate (WER) and character error rate (CER). All transfer learning approaches led to a significant increase in error rates in comparison to the baseline model, indicating that the models have been fine-tuned for a

specific domain.

In Fig. 4 and 5, the results are presented for the Serbian "mobile" HMI database. A significant decrease in both WER and CER (16.7% and 10.8% relative decrease, respectively) has been obtained. The second setup provided the best WER (4.44%). The best CER was obtained for the third setup (1.26%). Better results in terms of error rates have been obtained for all three configurations in comparison to the baseline model. The results are comparable with improvements from weight transfer approaches described in [15] given for English, or even better (probably because of the highly specific target domain in our case).

In Fig. 6 to 9, the results are given for experiment 2. For the first setup (exp. 2A, Fig. 6 and 7), an attempt has been made to adapt the whole network using only pure noise recordings. As expected, the approach led to significant disruption in terms of WER, since the whole network was adapted only to noise. E0 represents the original (baseline models), while E1-E4 correspond to epochs 1 to 4. After the initial "shock", the network started to regain accuracy for epochs 3 and 4. The point of this experiment was to try to address the problem of insertions, i.e., words recognized in places where there was only background noise. As the probability of noise has been increased, the number of insertions dropped – from the initial number of 10000, to around 6000 after one epoch, and to between 4000 and 5000 after the following epochs. Unfortunately, the number of deletions (words missed) and substitutions has been significantly raised as well, so the error rates rose too.

For the second approach (Fig. 8 and 9), a mixture of clean speech and pure noise recordings was applied (see Sect. 3D). The idea was to retain model variability and provide robustness in noisy conditions. Even though this experiment did not provide the desired results, the number of insertions has dropped slightly after the first epoch. Furthermore, in some cases in practice, for limited domains and smaller vocabularies, human subjects did report an increased accuracy and system stability.

V. Conclusion

Various transfer learning methods are employed in this paper in order to adapt more general acoustic models for domain-specific recognition, e.g. interactions with a voice assistant application on a smartphone. All of the methods provided improvements on the given domain in relation to the baseline model. However, this came with the cost of obtaining inferior results on other types of test utterances, which is expected. An additional approach presuming environmental (noise) adaptation is also examined in the paper. Although experimental results show an overall increase in error rates, a significant drop in the number of inserted words was achieved in some of the tests, showing that even this approach may have some practical uses. Future work includes experiments with other types of architectures and procedures for domain adaptation, as well as further optimization of the existing methods and specifically the environmental adaptation approach.

References

[1] E. Pakoci, B. Popović and D. Pekar, "Using morphological data in language modeling for Serbian large vocabulary speech recognition," in *Computational Intelligence and Neuroscience, Special Issue on Advanced Signal Processing and Adaptive Learning Methods*, vol. 2019, 8 pages, 2019.

[2] B. Popović, E. Pakoci and D. Pekar, "A comparison of language model training techniques in a continuous speech recognition system for Serbian," in *Proceedings of the 20th International Conference on Speech and Computer (SPECOM) – Lecture Notes in Artificial Intelligence*, vol. 11096, pp. 522-531, Leipzig, Germany, September 2018.

[3] E. Pakoci, B. Popović and D. Pekar, "Improvements in Serbian speech recognition using sequence-trained deep neural networks," in *SPIIRAS Proceedings*, vol. 3, no. 58, pp. 53-76, 2018.

[4] V. Peddinti, D. Povey and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. of the 16th Annual Conf. of the International Speech Communication Association (INTERSPEECH)*, pp. 3214-3218, Dresden, Germany, September 2015.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-end factor analysis for speaker verification," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788-798, 2011.

[6] B. Popović, E. Pakoci and D. Pekar, "End-to-end large vocabulary speech recognition for the Serbian language," in *Proceedings of the 19th International Conference on Speech and Computer (SPECOM) – Lecture Notes in Artificial Intelligence*, vol. 10458, pp. 343-352, Hatfield, UK, September 2017.

[7] E. Pakoci, "Influence of morphological features on language modeling with neural networks in speech recognition systems," Ph.D. thesis, Dept. Power, Electronic and Telecommunication Engineering, University of Novi Sad, Serbia, 2019.

[8] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, vol. 1, pp. 242, 2009.

[9] S. Pan and Q. Yang, "A survey on transfer learning," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[10] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proc. of the 28th International Conference on Machine Learning (ICML)*, Workshop on Unsupervised and Transfer Learning, pp. 1-20, Bellevue, WA, USA, July 2011.

[11] B. Popović, E. Pakoci, N. Jakovljević, G. Kočiš and D. Pekar, "Voice assistant application for the Serbian language," in *Proceedings of the 23rd Telecommunications Forum (TELFOR)*, pp. 858-861, Belgrade, Serbia, November 2015.

[12] D. Povey, H-K.J. Kuo and H. Soltau, "Fast speaker adaptive training for speech recognition," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1245-1248, Brisbane, Australia, September 2008.

[13] A. Stolcke, J. Zheng, W. Wang and V. Abrash, "SRILM at sixteen: update and outlook," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 5-9, Waikoloa, HI, USA, December 2011.

[14] E. Pakoci, B. Popović and D. Pekar, "Language model optimization for a deep neural network based speech recognition system for Serbian," in *Proc. of the 19th International Conf. on Speech and Computer (SPECOM) – Lecture Notes in Artificial Intelligence*, vol. 10458, pp. 483-492, Hatfield, UK, September 2017.

[15] P. Ghahremani, V. Manohar, H. Hadian, D. Povey and S. Khudanpur, "Investigation of transfer learning for ASR using LF-MMI trained neural networks," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 279-286, Okinawa, Japan, December 2017.

[16] B. Popović, E. Pakoci and D. Pekar, "Transfer learning in automatic speech recognition for Serbian," in *Proceedings of the 27th Telecommunications Forum (TELFOR)*, pp. 309-312, Belgrade, Serbia, November 2019.

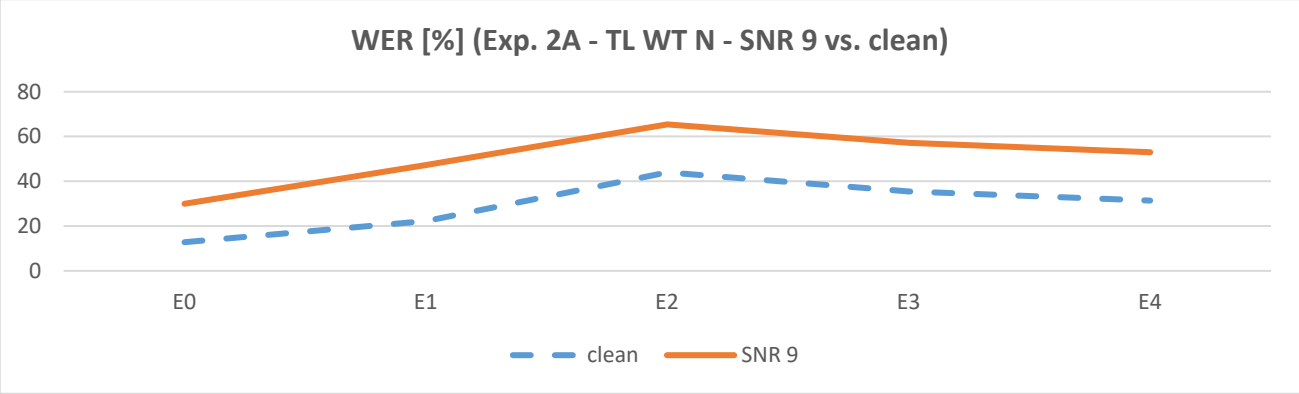

Fig. 6. Word error rate on clean speech and SNR 9 test set for the baseline model (EO) and epochs E1 to E4.

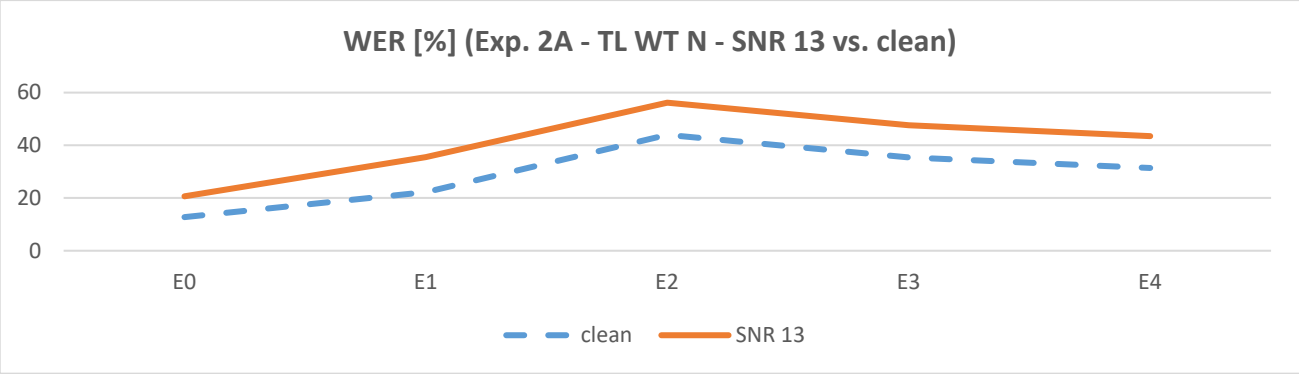

Fig. 7. Word error rate on clean speech and SNR 13 test set for the baseline model (EO) and epochs E1 to E4.

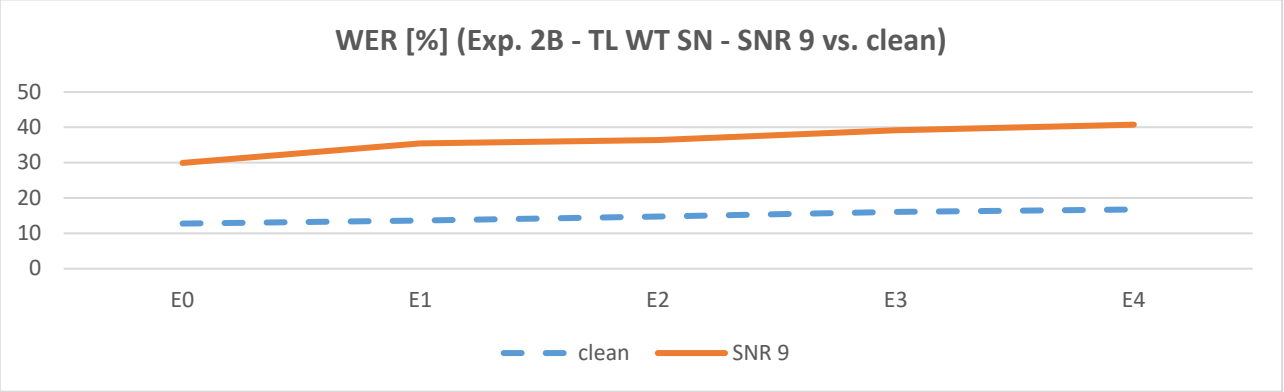

Fig. 8. Word error rate on clean speech and SNR 9 test set for the baseline model (EO) and epochs E1 to E4.

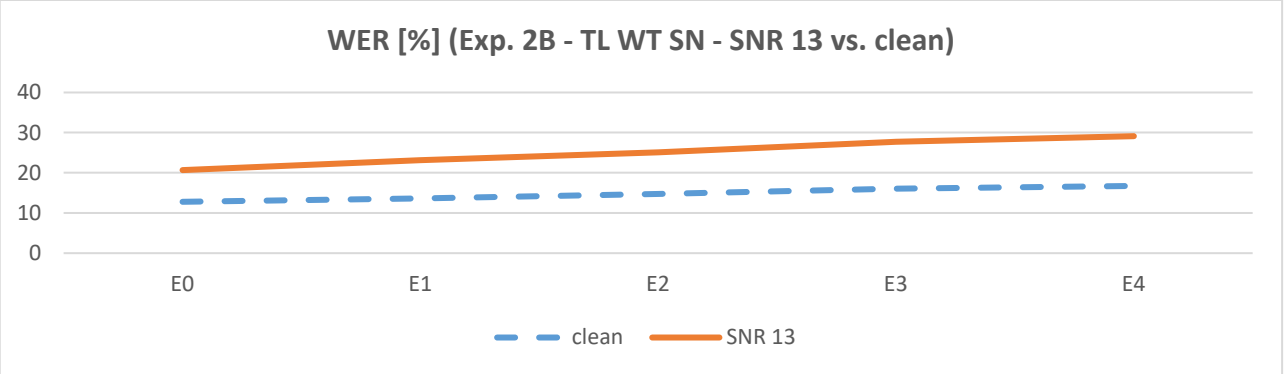

Fig. 9. Word error rate on clean speech and SNR 13 test set for the baseline model (EO) and epochs E1 to E4.