THE NDSC TRANSCRIPTION SYSTEM FOR THE 2016 MULTI-GENRE BROADCAST CHALLENGE

Xu-Kui Yang[1], *Dan Qu*[1*], *Wen-Lin Zhang*[1], *Wei-Qiang Zhang*[2]

[1] National Digital Switching System Engineering and Technological R&D Center, Zhengzhou, China
[2]Department of Electronic Engineering, Tsinghua University, Beijing, China
gzyangxk@gmail.com, qudanqudan@sina.com, zwlin_2004@163.com

ABSTRACT

The NDSC speech-to-text transcription system for 2016 multi-genre broadcast challenge is described. Various acoustic models based on deep neural network (DNN), such as hybrid DNN, long short term memory recurrent neural network (LSTM RNN), and time delay neural network (TDNN), are trained. The system also makes use of recurrent neural network language models (RNNLMs) for re-scoring and minimum Bayes risk (MBR) combination. The final WER of the speech-to-text task on the testing dataset is 18.2%. Furthermore, to simulate real applications where manual segmentations were not available an automatic segmentation system based on long-term information is proposed. WERs based on the automatically generated segments were slightly worse than that based on the manual segmentations.

Index Terms **—** speech recognition, broadcast transcription, deep neural networks, long-term information, Kaldi

1. INTRODUCTION

This paper describes the development of our speech-to-text transcription systems for the 2016 Multi-Genre Broadcast (MGB) challenge [1], a challenge task for state-of-the-art transcription systems of Arabic TV programs. The Arabic data used was taken from Aljazeera TV and consists of audio from 19 programs collected during 10 year (from 2005 to 2015). Comparing with the 2015 MGB challenge [1], there are fewer genres which mainly concentrate in conversation, interview, and report. But contents of the provided data cover 12 domains, namely politics, economy, society, culture, media, law, science, religion, education, sport, medicine, and military. Moreover, approaching 30% of the speech is spoken in different dialectal Arabic (e.g. Egyptian, Gulf, Iraqi, Levantine, North African etc.). Also, there are a few speeches spoken in other languages like English and French [2]. Thus, it is a challenging task with a much higher word error rate (WER) for the variability across different genres, dialects, and contents.

Firstly, auditory filter bank-based long-term information was used for training data selection. Then various deep neural networks (DNN) acoustic models were developed, including Hybrid DNN [4][5], long short term memory recurrent neural network (LSTM RNN) [6][7], and time delay neural network (TDNN) [8]. All systems were trained by utilizing the Kaldi speech recognition toolkit [3]. The n-gram language model were trained for first pass decoding and the recurrent neural network language models (RNNLMs) [9] were used for rescoring in second pass decoding stage. Lattice-based minimum Bayes risk (MBR) combination method [10] was utilized to combine different systems and reduce WER further. Finally, to simulate real applications where manual segmentations were not available, an automatic segmentation system was developed based on long-term filter bank information [11]. The new segmentation method not only makes use of the long-term information about speech signals, but also benefits from the logarithmic decomposition. WERs based on the automatically generated segments were slightly worse than that based on the manual segmentations.

The outline of this paper is as follows. Section 2 describes the related techniques used in our systems. A detailed description of the NDSC transcription system is given in Section 3. In Section 4, we introduce the automatic segmentation system based on long-term information. Section 5 presents the performance of our systems on evaluation dataset, and conclusions are drawn in Section 6.

2. RELATED WORK

In this section, we present a brief introduction to acoustic models, language models, system combination, and long-term information used in our transcription systems.

2.1. Acoustic models

For their excellent modeling capabilities and superior feature representations, deep neural networks have been widely used in speech recognition, especially for constructing high-performance acoustic models. Among them, the most classical one is hybrid DNN acoustic model,

[1] www.mgb-challenge.org
* corresponding author: Dan Qu (Email: qudanqudan@sina.com)

in which DNN is used as a scaled likelihood estimator replacing the GMM in a traditional GMM-HMM system.

Due to the nonstationarity of speech signals, the contextual information is critical for a speech recognition task. Compared with DNNs, RNNs have a powerful advantage for long contextual information representations because there are cyclic connections in the hidden layers of RNNs to model temporal correlations. However, the traditional RNNs suffer the gradient exploding or vanishing problems when being trained by stochastic gradient descent (SGD) algorithm [12]. Thus, the LSTM RNN is proposed to alleviate this problem. The primary difference between LSTM RNN and the traditional RNN is that in LSTM RNN linear recurrent connections are used instead of non-linear ones in the conventional RNN, leading to more smooth back propagation of gradients.

Another alternative neural network architecture is TDNN [8] which is seen as a precursor to convolutional neural networks (CNNs) [13][14]. Without the affine transform in the initial layer as a standard DNN, the TDNN still has the ability to modeling long-term temporal dependencies from short-term input speech features because the temporal resolutions which TDNN operates at increases from layer to layer. Moreover, under the assumption that neighboring activations are correlated, the sub-sampling processing is done to speed up the TDNN training and reduce the model size [15].

2.2. Language Models

N-gram LMs, predicting current word according to the previous *N*-1 words, are the most commonly used language models. Larger parameter *N* yields better contextual modeling ability, but requires more training data to avoid the data sparse problem. Although some smoothing methods can mitigate the impact of data sparseness, N can't be set to a large value, mostly 3 or 4. Namely, the N-gram LM has a limited ability to modeling context.

Comparing with N-gram LMs, RNNLMs can model longer context, resulting with significant WER reduction of speech recognition. However, due to the huge computational complexity, RNNLMs are usually used for re-scoring in a two-pass decoding strategy.

2.3. System Combination

By utilizing the complementarities of different systems, system combination has been used in various speech processing tasks to obtain better results. The commonly used combination methods are one-best-based combinations like recognizer out voting error reduction (ROVER) and lattice-based combinations such as MBR combination. MBR combination is a lattice-based system combination under the MBR decoding framework. In this method, the lattices from different systems are merged into one topology, and then the merged lattice is decoded to get a better result.

2.4. Long-term Information

Because of simplicity, robustness, and superior performance, long-term information [17][18][19] of speech signals has been studied deeply and widely used in voice activity detection. Among these methods, long-term information based on auditory filter banks is more discriminative and robust. Auditory filter banks, such as Mel filter banks and Pitch filter banks are a set of filter designed according to the function of human cochlea. The output of auditory filter bank can be regarded as a type of spectral decomposition in logarithmic forms. And the non-linear decomposition can represent some important acoustic cues like formant more explicitly.

The long-term spectrum divergence based on auditory filter banks between speeches and noises defined as the deviation of long-term spectral envelops respect to the average noise spectrums is given by:

$$D_*(l) = 10\log_{10}\left(\frac{1}{K}\sum_k \frac{E_*^2(k,l)}{\bar{N}_*^2(k,l)}\right) \quad (1)$$

where, K is the number of filter banks and R_d -order long-term spectral envelop is defined as follow:

$$E_*(k,l) = \max\left\{X_*(k,l-R_d+j)\middle| j=0,1,\cdots,2R_d\right\} \quad (2)$$

where, $\boldsymbol{X}_*$ represents pitch features or Mel spectrum, and $X_*(k,l)$ is the k^{th} band amplitude of $\boldsymbol{X}_*$ at frame l .

The noise spectrum $\boldsymbol{N}_*$ is estimated from $\boldsymbol{X}_*$ by the MMSE-based estimator [20]. And the average noise spectrum $\bar{N}_*(k,l)$ for the k^{th} band at frame l is defined as:

$$\bar{N}_*(k,l) = \frac{1}{l}\left((1-\alpha)(l-1)\bar{N}_*(k,l-1) + \alpha N_*(k,l)\right) \quad (3)$$

where, $N_*(k,l)$ is the noise feature value of the k^{th} band at frame l and $\bar{N}_*(k,1) = N_*(k,1)$.

The long-term information based on Mel filter is named long-term Mel spectral divergence (LTMD) and the other one based on pitch filter is called long-term pitch divergence (LTPD).

3. THE DEVELOPED SYSTEM

3.1. Data Used

The 2016 MGB challenge used audios from more than 3000 episodes spanning over 19 programs with a total duration of 1200 hours. The audios were recognized using QCRI Arabic LVCSR system, and then the automatic transcription was aligned with the human transcription to generate small speech segments with time information. Phone matched error rate (PMER) and word matched error rate (WMER) are calculated for data selection. Moreover, average word duration (AWD) in second was computed to reject unreliable segments.

The training set was selected with a zero WMER and an AWD ranging from 0.3s to 0.7s, leading to about 720 hours of data. Then, the selected data was filtered by using LTMD

(see Subsection 3.2.1 for detail). Finally, the developed system was trained with a corpus containing approximately 680 hours of data. It seems that there are some mistakes for AWD value calculation for each segment in the transcripts. Thus, we re-computed the AWD measure while selecting the training data.

Acoustic transcripts corresponding to the training set described above include 5 million words which would be used for 3-gram LMs training. For RNNLMs training, a large text corpus containing 12 years archive of articles from Aljazeera.net was available, yielding a total of 110 million words.

Both development and evaluation datasets include about 13 hour audios from 17 different program episodes with manual segmentations and transcriptions.

A grapheme-based lexicon supplied by QCRI was used.

3.2. ASR Systems

3.2.1. LMTD filtering

To begin with, we did some experiments following the MGB Kaldi recipe, namely a corpus containing 250 hours of data was used for training, selected from the top 500 episodes with a WMER of less than 80%. A LDA+MLLT+SAT GMM-HMM model was trained with 39 dimensional features (13 dimensional MFCCs, together with the relevant differential coefficients). All the feature dimensions were normalized by speaker-cluster level cepstral mean and variance normalization (CMVN).

As noticed in [2], we also found that there were some bad segments which embedded quite a long non-speech, even after we selected the data with a combination of WMER and AWD. Thus, a further filtering was carried out. Firstly, the speech/non-speech detection was done by comparing LTMDs of each frame with a pre-defined threshold. Then, the number of non-speech frames in each segment was counted. If this number accounts for 15% of the total frame number, the segment was considered to be a bad segment, and should be discarded.

After data filtering as the aforementioned method, a training set with 225 hour audio data is obtained. The same configuration is used to training a GMM-HMM speaker adapted model. To evaluate the two systems, the development dataset was decoded to calculate the WER. The results are listed in Table 1.

As a result, it can be seen that the filtered training dataset can help improve the system's performance. Hence, all the experiments below were based on the filtered dataset.

Filter	Duration	%WER
-	250h	39.2
LMTD	225h	**38.7**

Table 1. The %WER on dev dataset of GMM-HMM models trained on different datasets

3.2.2. Hybrid DNN system

To develop the hybrid DNN system, the 40 dimensional Mel filter bank (FBK) features and 3 pitch features were used along with both the first and second-order derivatives. CMVN was performed on these features. The final DNN input feature vectors are LDA-MLLT-fMLLR on top of FBK and pitch feature and cover 11 context frames, splicing current feature frame with 5 preceding and 5 succeeding frames. The DNN models have an input layer of 1419 units ($43\times3\times11$), 6 hidden layers with 1024 or 2048 sigmoid units in each layer, and the soft-max output layer with 10k tri-phone state targets. The networks were initialized using restricted Boltzmann machine (RBM) pre-training [21] and fine-turned using the cross-entropy (CE) criterion [22]. Table 2 compares the results of hybrid DNN system with different structures. The CE model with a more complex structure performs ($1419\times2048^6\times10000$) better. The discriminative sequence training experiments were done on this CE model. The 3-gram LM described above was used.

Structure	%WER
$1419\times1024^6\times10000$	30.3
$1419\times2048^6\times10000$	**29.4**

Table 2. The %WER on dev dataset of hybrid DNN system trained with different structures.

State-level minimum Bayes risk (sMBR) training [23] was performed 5 iterations to improve performance. The WERs of each iteration are listed in Table 3 and leads to a total reduction in WER of 3.7% absolute over the CE model.

Iteration	0	1	2	3	4	5
%WER	29.4	26.6	26.1	25.9	25.8	**25.7**

Table 3. %WER for MPE sequence training hybrid DNN system on dev dataset (the structure is $1419\times2048^6\times10000$).

3.2.3. LSTM system

Three copies of the training data corresponding to speed perturbations [24] of 0.9, 1.0, and 1.1 were created. The features used for LSTM system training were 40 dimensional MFCC with the appended 100 dimensional i-Vector. A LDA-MLLT-SAT GMM-HMM model was built to generate the basic clustered state alignment for neural network training. The LSTM RNN architecture had 3 LSTM hidden layers, where each LSTM layer had 1024 cells, and 3 projection layers with 256 units to reduce the number of parameters. The output state label was delayed by 5 frames. The output layer has 10k target units.

3.2.4. TDNN system

The TDNN system [15] was implemented using the 'chain' model available in the latest Kaldi recipe (egs/swbd/s5c). The 'chain' models are a new type of DNN-HMM model using the log-probability of the correct phone sequence as the objective function, instead of a frame-level objective. Moreover, the frame rate at the output of the neural network is 3 times smaller than that at the input. Thus, a modified

HMM topology was used, which allowed the HMM to be traversable in one transition.

The input features and speaker adapted models are the same as the LSTM system.

The performance of these three types of DNN system is showed in Table 4. It can be seen that both LSTM system and TDNN system are superior to standard hybrid DNN system, and TDNN system is a little better than LSTM system.

System	NN type	Feature	%WER
K01	DNN	FBK+Pitch	25.7
K02	LSTM	MFCC+i-vector	23.1
K03	TDNN	MFCC+i-vector	**23.0**

Table 4. %WER for different types of DNN system on dev dataset.

3.3. RNNLM Re-score

The rnnlm toolkit [25] was used for RNNLMs training. About 200k most frequent words from the additional text data were selected and added to enrich the vocabulary size. The pronunciations of these words were generated by the grapheme-to-phoneme (G2P) mapping. The RNNLMs parameters were set as the toolkit recommended. The number of neurons in hidden layer was 400, 5 steps of back propagation through time (BPTT) [26] training was used, and the amount of classes was 400.

System	NN type	%WER
K04	DNN	24.6
K05	LSTM	21.5
K06	TDNN	**21.3**

Table 5. %WER of RNNLM re-scoring for different types of DNN system and vocabulary sizes on dev dataset.

The results of RNNLM rescoring for different types of DNN system are shown in Table 5. With RNNLMs re-scoring, the system performances are improved up to 1.7% absolute. The best performance from TDNN system achieves 21.3% on development dataset.

3.4. System Combination

We evaluated MBR combination on development dataset, as shown in Table 6. The results show that MBR combination reduces the error rate and has a 0.9% absolute reduction in WER.

System	NN type	%WER
K07	K04:K06	20.8
K08	K04:K05:K06	20.4

Table 6. WER (%) on dev dataset for MBR combination; ':'represents MBR combination.

4. AUTOMATIC SEGMENTATION

In this section, we will describe our proposed automatic segmentation system based on long-term information in detail. The framework of the developed automatic segmentation system is shown in Fig. 1. The final segments are obtained by using a 3-step method, namely initial segmentation based on LTMD, fine-turning based on LTPD, and the final post-processing.

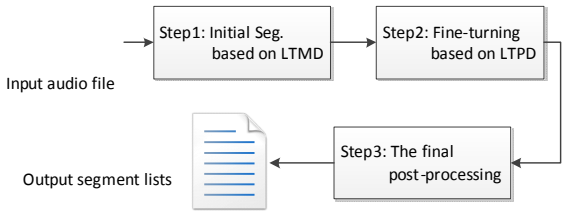

Fig. 1. The automatic segmentation system framework.

4.1. Initial Segmentation Based on LTMD

To raise segmentation efficiency, the initial segmentation is implemented by utilizing LTMD, computed with long frame length and shift, both of which were 50ms. And the order of long-term spectral envelop R_d was set to be 12. The distributions of LTMD for different types of audio (speech, music, and silence) on development dataset are estimated and shown in Fig. 2. It can be seen that LTMD can still be used for speech/non-speech detection though the distributions for music and silence almost overlap. To reduce the speech missing rate, the threshold is set to a small value of 10.5. Some simple strategies similar to hang-over scheme [27] are used to smooth the segmentation.

However, quite a part of non-speech is misclassified because of the small threshold. A further processing with more sophisticated classification criteria is required.

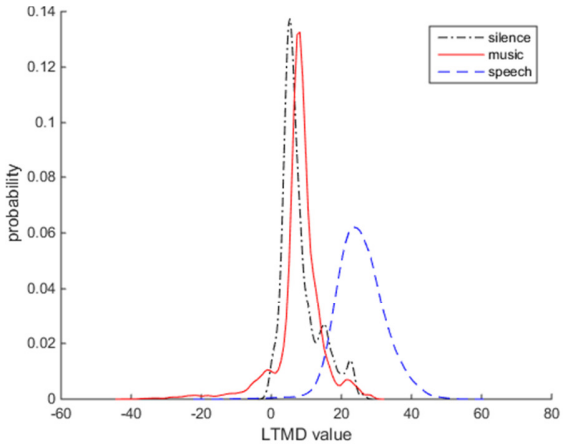

Fig. 2. The LTMD distribution of different audio types.

4.2. Fine-turning Based on LTPD

After the initial segmentation, more precise speech/non-speech detection was only done on speech segments to remove the embedded non-speech frames. When computing LTPD, the frame length and shift were 30ms and 10ms respectively, and $R_d = 6$.

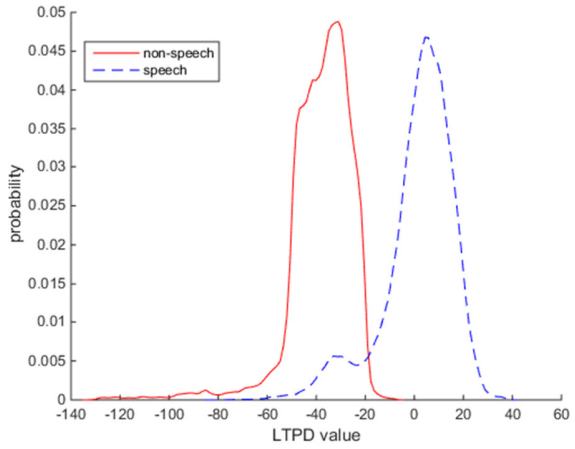

Fig. 3. The LTPD distribution of speech and non-speech audio.

The distributions of LTPD for speech and non-speech on development dataset were estimated. As the distributions shown in Fig. 3, the threshold for speech/non-speech detection is -20. It can be seen that there is a small peak in the speech distribution when the LTPD value is about -30. This is because there are some silence segments with short duration embedded in speech segments which aren't marked by the manual transcripts.

4.3. Final Post-processing

In the final post-processing stage, if the duration of a non-speech segment were less than 0.25s, it was merged with its adjacent speech segments. If the duration of a non-speech segment was less than 2.5s, we divided this segment in the middle. The first part was merged with the preceding speech segment and the second part was merged with the succeeding speech segment.

Finally, the audios in development dataset were divided into 2447 speech segments and 167 non-speech segments by the automatic segmentation system. The WER on these speech segments is shown in Table 7. The best performance of automatic segmentation system achieves 21.6%. The degradation of this segmentation system from manual segmentation is 1.2% absolute. And RNNLMs re-scoring still helps to reduce the error rate significantly while the effect of MBR combination seems to be limited. The reason for this appeared to be that the speech segments divided by the automatic system are quite long.

System	NN type	Re-scoring	%WER
K01	DNN	-	26.9
K02	LSTM	-	23.2
K03	TDNN	-	23.2
K04	DNN	✓	25.4
K05	LSTM	✓	22.0
K06	TDNN	✓	21.9
K07	K04:K06	-	21.6
K08	K04:K05:K06	-	21.6

Table 7. The %WER on dev dataset of automatic segmentation system

5. PERFORMANCE ON EVALUATION DATASET

The systems performances on evaluation dataset are shown in Table 8. It can be seen that the system K08 gives 18.2% WER with manual segmentation as well as 19.4% WER with automatic segmentation. This system was submitted as the primary system in the transcription evaluation task.

System	Manual				Automatic			
	orig	ndnp	glm	norm	orig	ndnp	glm	norm
K01	30.3	25.0	24.8	24.2	30.6	25.3	25.1	24.7
K02	26.7	21.0	20.7	20.2	27.0	21.4	21.1	20.6
K03	27.5	21.6	21.4	20.9	27.5	21.7	21.5	21.0
K04	28.1	22.6	22.5	22.0	28.7	23.3	23.2	22.7
K05	24.9	19.0	18.8	18.3	25.3	19.6	19.3	18.9
K06	25.4	19.4	19.2	18.7	25.7	19.8	19.6	19.1
K07	24.7	18.9	18.6	18.2	25.3	19.6	19.4	19.0
K08	24.4	18.5	**18.2**	17.8	25.3	19.6	**19.4**	19.0

Note: 'orig' is scoring the results with the original text, which may have punctuation/diacritization. 'ndnp' is scoring results after removing any punctuation or diacritization. 'glm' is using the Global Mapping File (GLM), which is the official number in the competition. 'norm' is after normalizing the Alef, yaa, and taa marbouta in the Arabic text.

Table 8. The %WER on evaluation dataset of all systems; 'Manual' means the manual segments and 'Automatic' represents automatic segments.

6. CONCLUSION

This paper describes the structure and development of NDSC Arabic transcription system for the 2016 MGB challenge. Different architectures of DNN-based acoustic model as well as RNNLMs re-scoring and MBR combination have been evaluated. The automatic segmentation based on long-term information is built to obtain suitable segments for speech transcription.

7. ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (No. 61175017, No. 61370034, and No. 61403224).

8. REFERENCES

[1] P. Bell, M.J.F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester and P. C. Woodland, "The MGB challenge: Evaluating multi-genre broadcast media transcription", In: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, 2015.

[2] A. Ali, P. Bell, J. Glass, Y. Messaoui, H. Mubarak, S. Renals, Y. Zhang "The MGB-2 Challenge: Arabic Multi-dialect Broadcast Media Recognition," In: *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT)*, 2016 .

[3] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit", In: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, Hawaii, USA, 2011.

[4] G.E. Dahl, D. Yu, L. Deng, et al., "Context-Dependent pre-trained deep neural networks for large-vocabulary speech

recognition," *IEEE Transactions on Audio, Speech and Language Processing*, 2012, 20(1):30-42.

[5] L. Deng , J.Y. Li, J.T. Huang, et al., "Recent advances in deep learning for speech research at Microsoft," *In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[6] A. Graves, A. Mohamed, G.E. Hinton, "Speech recognition with deep recurrent neural networks," *In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 6645–6649.

[7] A. Graves, N. Jaitly, A. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," *In: Proceedings of Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, Czech Republic, 2013, pp.273–278.

[8] V. Peddinti, G. Chen, D. Povey, S. Khudanpur, "Reverberation Robust Acoustic Modeling Using i-Vectors with Time Delay Neural Networks," In: *Proceedings of Interspeech*, Dresden, Germany, 2015, pp. 2440 – 2444

[9] T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, and S. Khudanpur, "Extensions of recurrent neural network language model," In: *Proceedings of ICASSP*, 2011, pp. 5528–5531.

[10]H. Xu, D. Povey, L. Mangu, et al., "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech and Language*, 2011, 25(4): 802-828.

[11]X.K. Yang, L. He, D. Qu, W.Q. Zhang, "Voice activity detection algorithm based on long-term pitch information," *EURASIP Journal on Audio, Speech, and Music Processing* 2016, 2016:14.

[12] X. Glorot, Y. Bengio, "Understanding the difficulty of training deep feed-forward neural networks," *J. Mach. Learn. Res.*, vol. 9, 2010, pp. 249–256.

[13]O. Abdel-Hamid, A. Mohamed, H. Jiang, G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," *In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Kyoto, Japan, 2012, pp.4277–4280.

[14]T.N. Sainath, B. Kingsbury, A. Mohamed, B. Ramabhadran, "Learning filter banks within a deep neural network framework," *In: Proceedings of Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, Czech Republic, 2013, pp.297–302.

[15]V. Peddinti, D. Povey, S Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," *In: Proceedings of Interspeech*, 2015.

[16]A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech and Signal* Processing, vol. 37, no. 3, pp. 328–339, 1989.

[17]J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3-4, pp. 271-287, April 2004.

[18]P.K. Ghosh, A. Tsiartas, S. Narayanan, "Robust Voice Activity Detection Using Long-Term Signal Variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600-613, 2011.

[19]Yanna Ma, Akinori Nishihara, "Efficient voice activity detection algorithm using long-term spectral flatness measure," *EURASIP Journal on Audio, Speech and Music Processing*, 2013:21.

[20]Timo Gerkmann, and Richard C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383-1393, 2012.

[21]G.E. Hinton, "A Practical Guide to Training Restricted Boltzmann Machines", *Technical Report, UTML TR 2010-003,* Department of Computer Science, University of Toronto, 2010.

[22]K Vesely, A Ghoshal, L Burget, and D Povey, "Sequence discriminative training of deep neural networks," *In: Proceedings of Interspeech*, 2013.

[23]H. Xu, D. Povey, L. Mangu, & J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance", *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

[24]T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," *In: Proceedings of Interspeech*, 2015.

[25]T. Mikolov, S. Kombrink, A. Deoras, et al. "RNNLM - Recurrent neural network language modeling toolkit," In: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, Hawaii, USA, 2011, pp. 196-201.

[26]M. Bodén, "A Guide to Recurrent Neural Networks and Backpropagation," *In the Dallas project*, 2002.

[27]A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J. P. Petit, "ITU-T Recommendation G.729 Annex B: A Silence Compression Scheme for Use with G.729 Optimized for V.70 Digital Simultaneous Voice and Data Applications," *Communications Magazine, IEEE*, 35(9): 64-73, September 1997.