International Journal of Scientific Advances
ISSN: 2708-7972
Volume: 3 | Issue: 5 | Sep - Oct 2022 Available Online: www.ijscia.com
DOI: 10.51542/ijscia.v3i5.7

Armenian Speech Recognition System: Acoustic and Language Models

Varuzhan H. Baghdasaryan
National Polytechnic University of Armenia

*Corresponding author details: Varuzhan H. Baghdasaryan; varuzh2014@gmail.com

ABSTRACT
Nowadays automatic speech recognition (ASR) is an important task for machines. Several applications such as speech translation, virtual assistants and voice bot systems use ASR to understand human speech. Most of the research and available models are for widely used languages, such as English, German, French, Chinese and Spanish. This paper presents the Armenian speech recognition system. As a result of this research developed acoustic and language models for the Armenian language (modern ASR systems combine acoustic and language models to achieve higher accuracy). RNN-based Baidu's Deep Speech deep neural network was used to train the acoustic model, and the KenLM toolkit was used to train the probabilistic language model. The acoustic model was trained and validated on ArmSpeech Armenian native speech corpus using transfer-learning and data augmentation techniques and tested on the Common Voice Armenian database. The language model was built based on the texts scraped from Armenian news websites. Final models are small in size and can be run and do real-time speech-to-text tasks on IoT devices. Testing on the Common Voice Armenian database the model gave 0.902565 WER and 0.305321 CER without the language model, and 0.552975 WER and 0.285904 CER with the language model. The paper aims to describe environment setup, data collection, acoustic and language models training processes, as well as final results and benchmarks.

Keywords: Armenian ASR; speech recognition system; speech-to-text; acoustic model; language model

INTRODUCTION
According to Wikipedia [1], automatic speech recognition (ASR) is a technology which converts spoken language into text. For popular languages, there are many publicly available models to do tasks like automatic speech recognition. This paper presents the Armenian speech recognition system.

Speech-to-text systems are mainly implemented through machine learning techniques to generate text from a human speech in an end-to-end approach. Some of these techniques are not based on neural networks, such as Hidden Markov Model (HMM) [2, 3] which is a stochastic, statistical model. The best implementation and use of the Hidden Markov Model is CMUSphinx [4] open-source speech recognition system.

In the recent period, models based on not-neural-network solutions are replacing by neural networks. There are many well-known neural networks for this task. Some of them have high-speed performance and low accuracy and some of them have high accuracy but a long time of processing. One of these neural networks is Baidu's Deep Speech [5]. Mozilla's DeepSpeech [6] is an open-source implementation of Baidu's Deep Speech deep neural network. The engine is based on a recurrent neural network (RNN) [7, 8] and consists of 5 layers of hidden units.

Recurrent neural networks are modern and have a tendency to develop and it is preferable to train the language model using DeepSpeech or its modified version (no need to invent a new bicycle).

In the frame of research 2 models will be developed: the acoustic model [9] and the language model [10]. The acoustic model and language model work together to produce better accuracy of prediction. The acoustic model uses a sequence-to-sequence algorithm, to learn which acoustic signals correspond to which letters in the language alphabet (outputs probabilities for each class of character, not at the word level). To distinguish homonyms (words that sound the same but are spelled differently), a language model comes to the rescue.

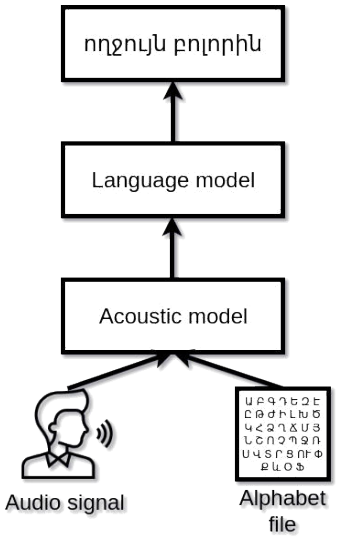

FIGURE 1: The principle of the Armenian ASR system.

The language model predicts which words will follow each other in a sequence (n-gram modelling).

The acoustic model defines the relationship between a speech audio signal and the phonemes, while the language model defines the relationship between language words and their sequences in the language. For acoustic model training and validating used ArmSpeech Armenian spoken language corpus [11, 12] totally of 15.7 hours. For acoustic model testing used Common Voice [13] Armenian database.

Language model training is based on the KenLM library [14]. KenLM is a Language Model Toolkit that enables build an n-gram [15, 16] language model. Necessary data for language model training was scraped from Armenian news websites articles about medicine, sport, culture, lifestyle and politics, and normalized by main normalization rules [17].

RELATED WORKS

Before proceeding to the actual material, it is necessary to carry out research on previous works. This will help to understand their achievements, disadvantages and advantages and try to use that experience in this work. There are many pieces of research for Russian, English, German, French and Mandarin Chinese languages.

Aashish Agarwal and Torsten Zesch [18] introduced the German ASR system based on DeepSpeech. For acoustic model training, they used 3 publicly available datasets: Voxforge, Tuda-De and Mozilla Common Voice. All datasets are multispeaker. Combined data were split into 70% of training, 15 % of validation, and 15% of the test set. For training a 3-gram language model Radeck-Arneth et al corpus was used, which consists of eight million filtered sentences (63.0% Wikipedia, 22.0% Europarl, and 14.6% crawled sentences). Table 1 shows the final results across three datasets.

TABLE 1: result [18].

Train	Test	WER (%)
Voxforge	Voxforge	72.1
Tuda-De	Voxforge	96.8
Mozilla	Voxforge	73.1
Tuda-De, Mozilla	Voxforge	66.2
Tuda-De	Tuda-De	26.8
Voxforge	Tuda-De	98.5
Mozilla	Tuda-De	84.9
Voxforge, Mozilla	Tuda-De	83.8
Mozilla	Mozilla	79.7
Tuda-De	Mozilla	94.8
Voxforge	Mozilla	87.1
Tuda-De, Voxforge	Mozilla	80.5

This paper [19] presents the Russian speech recognition system again based on DeepSpeech. Used datasets are:

- 'yt-vad-1k': 1000 hours of audio recordings extracted from videos on YouTube (by over 1000 people in a variety of recording conditions and with varying degrees of background noise),
- 'voxforge-ru-clean': 11.5 hours of audio-transcripts pairs,
- 'yt-vad-650-clean', 650 hours labelled audio.

The best result is a 22% word error rate (WER) in the case of training on yt-vad-1k-train and yt-vad-650-clean-train datasets and testing on voxforge-ruclean-test set (with use of the LM language model).

SPEECH CORPUS

As mentioned above an acoustic model represents the relationship between an audio signal and the phonemes or other linguistic units. The acoustic model learns language features from the corpora which contain the set of audio recordings and corresponding transcripts. So, the acoustic model is the final output of the neural network that must perform the speech-to-text operations.

Training speech-to-text engines with multi-speaker corpora give the ability to extract all the features of language and increase the efficiency of the system when it is used by users of different genders, ages, timbres and accents.

The dataset used for acoustic model training and validation is ArmSpeech [11, 12], which is an Armenian Spoken Language Corpus totally 15.7 hours.

ArmSpeech is a free speech corpus, in which audio clips are extracted from free-to-use and publicly available Armenian audiobooks or recorded by voice volunteers. Data were split into 80% of training and 20% of validation sets. Figure 2 shows the ratio of male and female speeches in the ArmSpeech corpus.

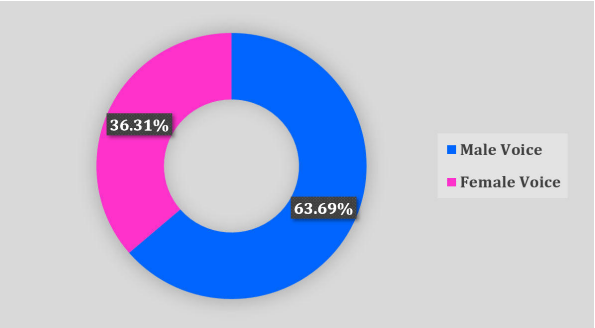

FIGURE 2: ratio of male and female speeches in the ArmSpeech corpus [11, 12].

Statistics of the two releases of ArmSpeech and other related information are presented in Table 2.

TABLE 2: statistics of ArmSpeech corpus [11,12].

Specifications	ArmSpeech first release	ArmSpeech second release	Two releases together
Total duration	11:46:26	04:00:52	15:47:19
Minimum sample duration	0.72 seconds	0.62	0.62
Maximum sample duration	10.00 seconds	13.96	13.96
Mean sample duration	6.8 seconds	2.7	4.9
Total number of samples	6206	5378	11584
Total number of unique sentences (words or phrases)	6205	4838	11026
Total symbols	414685	160729	575414

Specifications	ArmSpeech first release	ArmSpeech second release	Two releases together
Minimum number of symbols in samples	2	1	1
Maximum number of symbols in samples	144	135	144
Mean number of symbols in each sample	66.82	29.89	49.67
Total words	80632	26039	106671
Unique words	16847	9391	23062
Minimum number of words in samples	1	1	1
Maximum number of words in samples	31	19	31
Mean number of words in each sample	12.99	4.84	9.21

For acoustic model testing used the Common Voice [13] Armenian database. The Common Voice corpus is a multilingual speech corpus consisting of audio-transcript pairs.

Both ArmSpeech and Common Voice contain mono-channel, 16-bit audio clips with a 16000 Hz sampling rate and 256 kbps bit rate. Audio samples are in WAV (lossless compression) format.

ACOUSTIC MODEL PREPARATION

As the acoustic model results from the neural network training process, it is essential to describe the structure and fundamental components of NN used for making an acoustic model. In the frames of research used Mozilla's DeepSpeech [6], which implements Baidu's Deep Speech ASR [5] and uses TensorFlow. It is a deep recurrent neural network (RNN) [7, 8] composed of 5 hidden layers and performs supervised learning tasks.

Neural network hidden layers are:

- the first layer (fully connected layer), which input parameters are MFCC [20, 21, 22, 23] features (coefficients) extracted from audio. Uses ReLU [24, 25] activation function (ReLU(x) = max (0, x)),
- the second and third layers are fully connected layers, which also use the ReLU [24, 25] activation function,
- bidirectional RNN layer (LSTM [26, 27] cells with tanh activation),
- ReLU.

Fully connected layer (output layer) that uses a softmax activation for normalization outputs probabilities for each character in the language's alphabet.

Figure 3 shows the architecture of Mozilla's implementation of Baidu's Deep Speech.

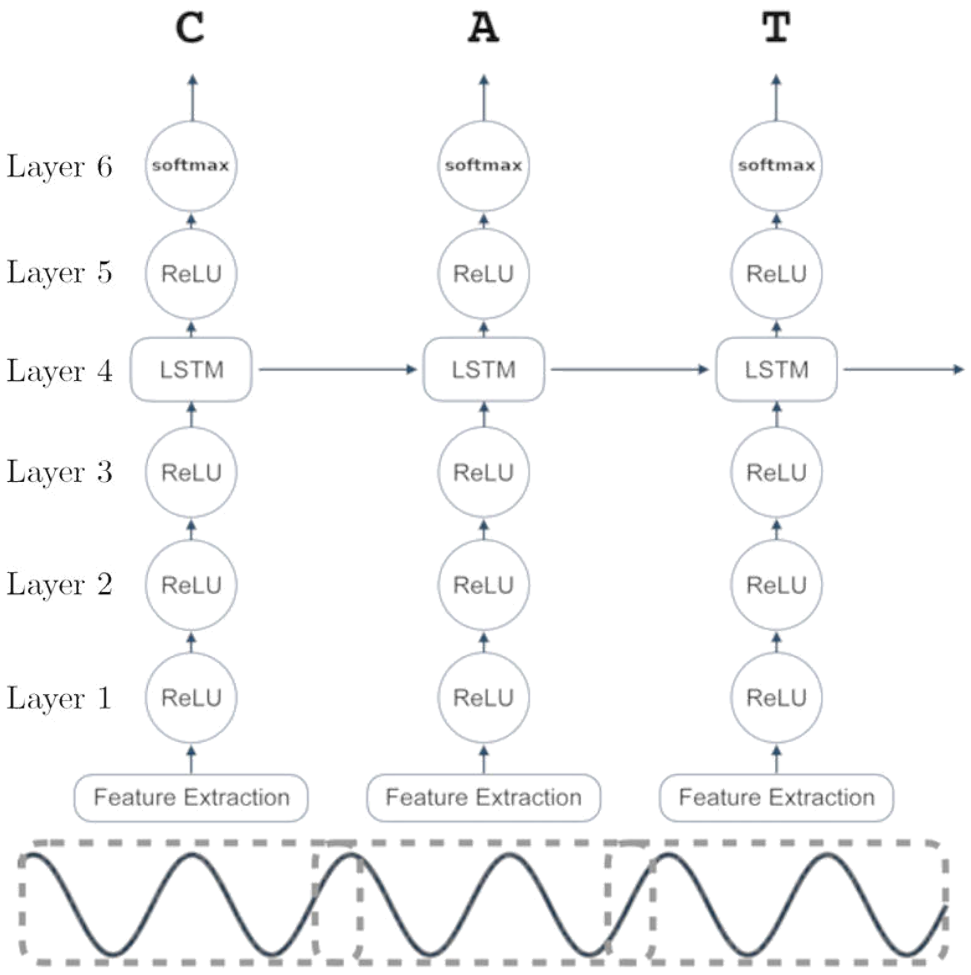

FIGURE 3: the architecture of Mozilla's implementation of Baidu's Deep Speech [6].

During audio processing, the system (Deepspeech) gives a probability for each character from the language alphabet. Then the system uses Connectionist Temporal Classification (CTC) [28] to improve the accuracy of the prediction.

LANGUAGE MODEL PREPARATION

It is a common approach to improve the accuracy of predictions of language acoustic models by using a statistical language model [10]. A statistical language model is a probability distribution over sequences of words and is used to give probabilities to words and phrases based on statistics from training data. The language model provides context to distinguish between words and phrases that sound phonetically similar.

The language model is an n-gram model [15, 16]. For the language model training, the data, which is a text file (sentences) should be very large and include all areas of language use to ensure higher accuracy and transcription of the text of any aspect of life. That's why it is preferable to collect language model resources from books, news articles, etc.

After collecting, this text data must pass the normalization stage. During this stage all punctuation marks must be removed, digits and dates must be replaced with their alphabetical representation and the text must be converted to lowercase. This can be done either manually or by an automation application, which must be developed by taking into account language specifications, grammar and rules.

Mozilla DeepSpeech uses the KenLM toolkit [14] to make queries to the language model [10, 29]. KenLM is a library which implements PROBING and TRIE data structures. According to the KenLM paper PROBING data structure uses linear probing hash tables (ensures high speed) and the TRIE data structure is a trie with bit-level packing.

For most common English speech-to-text models the language model trained with KenLM uses the LibriSpeech normalized LM training text [30], which contains millions of lines of English sentences.

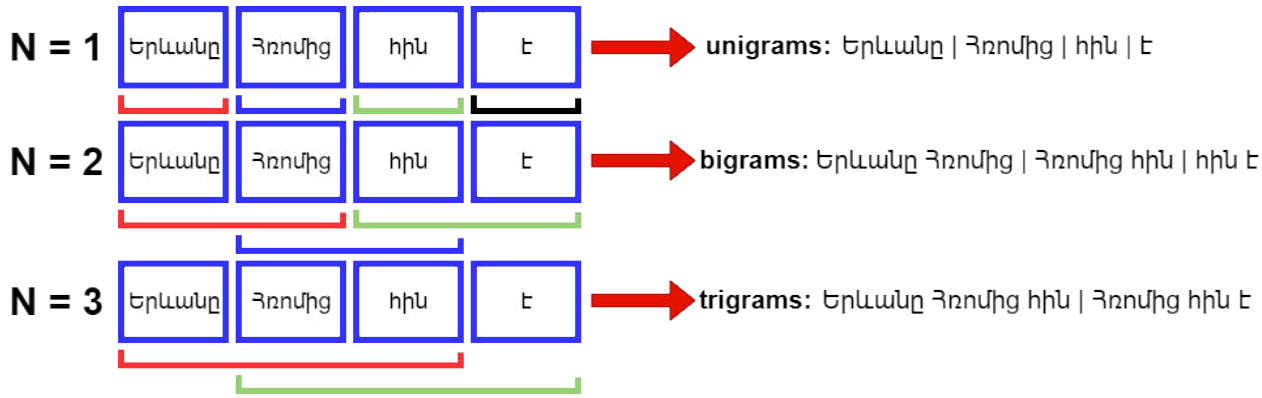

FIGURE 4: n-gram model.

The text for language model training was scraped from Armenian news websites articles about lifestyle, music, culture, politics and sport. Scraping was done using a small console application written in Python programming language. This application also normalized all text by the rules mentioned above. The total number of processed articles is 1884833. The final result is a text containing normalize sentences in 9539350 lines.

TRAINING HYPERPARAMETERS & SERVER

To improve the accuracy of an acoustic model can be used either fine-tuning or transfer learning techniques.

Fine-tuning uses the same alphabet.txt (characters) and a set of checkpoints from another model (another acoustic model of the same language) to make a new model.

Transfer learning uses the new language alphabet to train a model from another language model (neural network drops the old alphabet). This technique is the best to train a new language acoustic model from another pre-trained and similar language.

Armenian language acoustic model was trained using the transfer-learning technique [31, 32], which gives the ability to drop certain layers from a pre-trained model (pre-trained acoustic model for English) and initialize those layers for a new language. Mozilla's DeepSpeech system allows reinitialising five layers (from end to beginning: the output layer, penultimate layer, LSTM layer and the next two fully connected layers) when using transfer learning. Onno Eberhard and Torsten Zesch [33] in their research have shown that models give the best accuracy when dropping 2 or 3 layers.

Also applied data augmentation technique to increase dataset size and speech variances. In the frame of experiment augmentation types that are used are:

- sample domain augmentation (overlay, reverb, resample, codec, volume),
- spectrogram domain augmentation (pitch, tempo, frequency mask),
- multi-domain augmentation (time mask, dropout, add, multiply).

Training, validation and test batch sizes are 1 (it gives the higher effectiveness of training). The total number of training epochs is 150 with a learning rate of 0.001. Also used Tensorflow's CuDNN RNN backend for training on GPU. The width of hidden layers is 2048. The dropout rate for feedforward layers is 0.05. The training was done on a machine running 64-bit Linux OS (Ubuntu 20.04 LTS) with Intel Core i7-9700 CPU @ 3 GHz. GPUs are one NVIDIA GeForce GTX 1660 TUF and one NVIDIA GeForce GTX 1650 GPU with a totally of 10GB of memory. One epoch (with training and validation stages) of acoustic model training took approximately 0.9 hours.

RESULTS

The two most important units of measurement of acoustic and language efficiency and accuracy are CER (character error rate) [34] and WER (word error rate) [34, 35].

The WER shows the accuracy of the language model (how accurately it recognises a word) [10, 34, 35] and The CER shows the accuracy of the acoustic model (how accurately it recognises a character) [9, 34].

Both WER and CER can be computed as [34, 35]:

$$WER/CER = (S + D + I) / N = (S + D + I) / (S + D + C),$$

where:

- 'S' is the number of substitutions,
- 'D' is the number of deletions,
- 'I' is the number of insertions,
- 'C' is the number of correct words/characters,
- 'N' is the number of words/characters in the reference (N=S+D+C).

Both metric takes references (a list of references for each speech input) and predictions (a list of transcriptions to score) as input and outputs a float representing the word/character error rate.

After training the acoustic model first test was without the KenLM language model. The test dataset is a multi-speaker Common Voice Armenian dataset consisting of 1264 samples. The acoustic model gave 0.902565 WER and 0.305321 CER. The WER and CER are poor because wasn't used the language n-gram model.

Figure 5 shows training and validation loss curves (loss shows the difference between prediction and the expected output), which are going down over time.

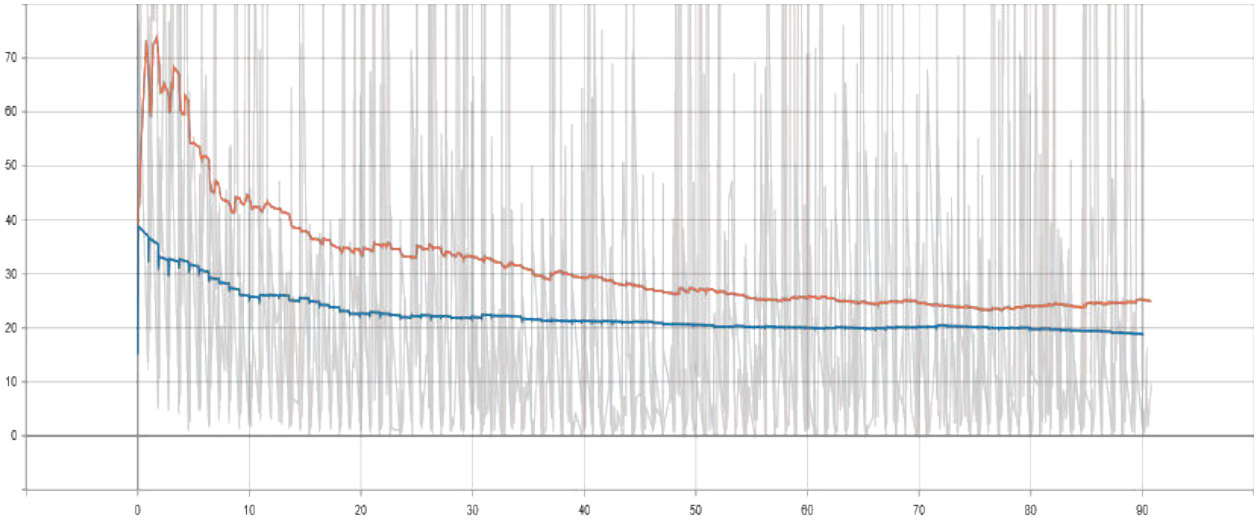

FIGURE 5: loss diagram of training and validation epochs.

After building the KenLM language model and testing on the same common voice dataset 2 models together (acoustic and language model) gave 0.552975 WER and 0.285904 CER.

CONCLUSION

In the frames of research developed an Armenian-language speech recognition system that delivers the WER of 0.552975 and CER of 0.285904. Both acoustic and language models are small in size and can be used on IoT devices to do simple speech-to-text tasks. It is planned to fine-tune the models with additional Armenian datasets and improve the accuracy and performance of the models.

REFERENCES

[1] Speech recognition. In Wikipedia. https://en.wikipedia.org/wiki/Speech_recognition.

[2] Mariette Awad, Rahul Khanna. Hidden Markov Model. Efficient Learning Machines, pages 81-104, January 2015.

[3] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi. Hidden markov model speech recognition arrangement. US Patent 4,587,670, May 6 1986.

[4] CMU Sphinx. In Wikipedia. https://en.wikipedia.org/wiki/CMU_Sphinx.

[5] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng. Deep Speech: Scaling up end-to-end speech recognition, arXiv preprint arXiv:1412.5567, 19 December 2014.

[6] Mozilla, 'Project DeepSpeech', 2021. https://github.com/mozilla/ DeepSpeech.

[7] Alex Sherstinsky. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. arXiv:1808.03314v9, 31 January 2021.

[8] Recurrent neural network. In Wikipedia. https://en.wikipedia.org/wiki/Recurrent_neural_network.

[9] Acoustic model. In Wikipedia. https://en.wikipedia.org/wiki/Acoustic_model.

[10] Language model. In Wikipedia. https://en.wikipedia.org/wiki/Language_model.

[11] Varuzhan H. Baghdasaryan. ArmSpeech: Armenian Spoken Language Corpus. International Journal of Scientific Advances (IJSCIA), Volume 3| Issue 3: May-Jun 2022, pages 454-459.

[12] Varuzhan H. Baghdasaryan. Extended ArmSpeech: Armenian Spoken Language Corpus. International Journal of Scientific Advances (IJSCIA), Volume 3| Issue 4: Jul-Aug 2022, pages 573-576.

[13] Ardila, Rosana, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers and Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus. Proceedings of the 12th Language Resources and Evaluation Conference, May 2020, pages 4218-4222.

[14] Kenneth Heafield. KenLM: Faster and Smaller Language Model Queries. WMT at EMNLP, Edinburgh, Scotland, United Kingdom, July 2011, pages 30-31.

[15] M. Siu and M. Ostendorf. Variable n-grams and extensions for conversational speech language modeling. IEEE Transactions on Speech and Audio Processing. Volume 8, no. 1, 2000, pages 63–75.

[16] n-gram. In Wikipedia. https://en.wikipedia.org/wiki/N-gram.

[17] Text normalization. In Wikipedia. https://en.wikipedia.org/wiki/Text_normalization.

[18] Aashish Agarwal and Torsten Zesch. German End-to-end Speech Recognition based on DeepSpeech. Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers, 2019, pages 111-119.

[19] Iakushkin, Oleg & Fedoseev, George & Shaleva, Anna & Degtyarev, Alexander & Sedova, Olga. Russian-Language Speech Recognition System Based on Deepspeech. Proceedings of the VIII International Conference 'Distributed Computing and Grid-technologies in Science and Education' (GRID 2018), Dubna, Moscow region, Russia, September 10 - 14, 2018.

[20] Mel-frequency cepstrum. In Wikipedia. https://en.wikipedia.org/wiki/Mel-frequency_cepstrum#:~:text=Mel%2Dfrequency%20cepstral%20coefficients%20(MFCCs,%2Da%2Dspectrum%22).

[21] J. Martinez, H. Perez, E. Escamilla and M. M. Suzuki, 'Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques', CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers, 2012, pp. 248-251, doi: 10.1109/CONIELECOMP.2012.6189918.

[22] Vibha Tiwari. MFCC and its applications in speaker recognition. International Journal on Emerging Technologies 1(1): 19-22(2010). ISSN: 0975-8364.

[23] Shikha Gupta, Jafreezal Jaafar, Wan Fatimah wan Ahmad and Arpit Bansal. FEATURE EXTRACTION USING MFCC. Signal & Image Processing: An International Journal (SIPIJ) Vol.4, No.4, August 2013. DOI: 10.5121/sipij.2013.4408 101.

[24] Abien Fred M. Agarap. Deep Learning using Rectified Linear Units (ReLU). arXiv:1803.08375v2, 7 February 2019.

[25] Arnold M. Pretorius, Etienne Barnard, Marelie H. Davel. ReLU and sigmoidal activation functions. FAIR 2019, pages 37-48.

[26] Alex Graves and Jurgen Schmidhuber. Framewise Phoneme Classification with Bidirectional LSTM Networks. Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31 - August 4, 2005.

[27] Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. Neural Computation, 9(8):1735-1780, 1997.

[28] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. ICML 2006, Pittsburgh, USA, pp. 369-376.

[29] Heafield K., et al. Scalable modified Kneser-Ney language model estimation. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Short Papers, Volume 2, 2013, pages 690-696.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

[31] Kunze, Julius et al. (2017). Transfer Learning for Speech Recognition on a Budget. Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL, Vancouver, Canada, August 3 2017. Association for Computational Linguistics, pages 168–177.

[32] Li, Bryan, Xinyue Wang, and Homayoon S. M. Beigi. Cantonese Automatic Speech Recognition Using Transfer Learning from Mandarin. CoRR, 2019.

[33] Onno Eberhard and Torsten Zesch. Effects of Layer Freezing when Transferring DeepSpeech to New Languages. arXiv:2102.04097v1, 8 February 2021.

[34] Morris, Andrew & Maier, Viktoria & Green, Phil. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. 10.21437/Interspeech.2004-668, 2004.

[35] Woodard, J.P. and Nelson, J.T. An information theoretic measure of speech recognition performance. Workshop on standardisation for speech I/O technology, Naval Air Development Center, Warminster, PA, 1982.