http://dx.doi.org/10.14236/ewic/HCI2018.150

Broadcast Language Identification & Subtitling System (BLISS)

Jinling Wang[1] Karla Muñoz Esquivel[2] James Connolly[3] Kevin Curran[2] Paul Mc Kevitt[4]
Faculty of Computing, Engineering & Built Environment, [1]Ulster University (Jordanstown), Newtownabbey, BT37 0QB, [2]Ulster University (Magee), Derry/Londonderry, BT48 7JL, Northern Ireland
{j.wang, kc.munoz-esquivel, kj.curran}@ulster.ac.uk

[3]Department of Computing, Letterkenny Institute of Technology (LYIT), Port Road, Letterkenny, IRL- F92 FC93, Co. Donegal, Ireland
James.Connolly@lyit.ie

[4]Faculty of Arts, Humanities & Social Sciences, Ulster University (Magee), Derry/Londonderry, BT48 7JL, Northern Ireland
p.mckevitt@ulster.ac.uk

Accessibility is an important area of Human Computer Interaction (HCI) and regulations within many countries mandate that broadcast media content be accessible to all. Currently, most subtitles for offline and live broadcasts are produced by people. However, subtitling methods employing re-speaking with Automatic Speech Recognition (ASR) technology are increasingly replacing manual methods. We discuss here the subtitling component of BLISS (Broadcast Language Identification & Subtitling System), an ASR system for automated subtitling and broadcast monitoring built using the Kaldi ASR Toolkit. The BLISS Gaussian Mixture Model (GMM)/Hidden Markov Model (HMM) acoustic model has been trained with ~960 hours of read speech, and language model with ~900k words combined with a pronunciation dictionary of 200k words from the LibriSpeech corpus. In tests with ~5 hours of unseen clean speech test data with little background noise and seen accents BLISS gives recognition accuracy of 91.87% based on the WER (Word Error Rate) metric. For ~5 hours of unseen challenge speech test data, with higher-WER speakers, BLISS's accuracy reduces to 75.91%. A BLISS Deep Learning Neural Network (DNN) acoustic model has also been trained with ~100 hours of read speech data. It's accuracy for ~5 hours of unseen clean and unseen challenge speech test data is 92.88% and 77.27% respectively based on WER. Future work includes training the DNN model with ~960 hours of read speech data using CUDA GPUs and also incorporating algorithms for background noise reduction. The BLISS core engine is also intended as a Language Identification system for broadcast monitoring (BLIS). This paper focuses on its Subtitling application (BLSS).

Automatic Speech Recognition (ASR), Accent, Automated Subtitling, Background Noise, BLISS, Human-Computer Interaction, Kaldi, LibriSpeech

1. INTRODUCTION

An important aspect of Human Computer Interaction (HCI) is accessibility, involving production of text from speech (Subtitles/Captions) (Romero-Fresco, 2014) for those who cannot hear and audio from video (Audio Description) (Fryer, 2016) for those who cannot see. Manual production of time-aligned transcriptions of audio-visual content requires considerable effort. It is prone to manual typing errors, slow for real-time delivery and human subtitlers can be costly (Alvarez et al., 2016). For live subtitles, re-speaking techniques are combined with off-the-shelf Automatic Speech Recognition (ASR) engines to produce subtitles. With re-speaking, the audio content is re-spoken by a professional speaker. This results in speech with reduced accents and noise which can be processed by ASR engines with an acceptable accuracy for live subtitling. However, re-speaking causes delays in real-time subtitling tasks and requires the re-speaker to dictate the audio content in a speaker independent manner.

Improving the quality of subtitles for people with audio and visual impairments is an important focus for the UK's communications regulator, Ofcom (Ofcom, 2013). Broadcasters are required to measure and improve the quality of live broadcasts so that subtitles are synchronised with video and speech in addition to achieving high accuracy rates. In terms of broadcast monitoring, broadcasters must also ensure that correct language playout occurs in different geographic regions.

Advanced ASR technology can help solve the problems of subtitle delay during live broadcasts in addition to improving the accuracy obtained. We discuss here a platform called BLISS (Broadcast Language Identification & Subtitling System) for performing language identification and subtitling. The core BLISS technology is based on advanced ASR with the use of Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), and Deep

Neural Networks (DNNs) implemented within the Kaldi Toolkit platform (Kaldi, 2018). BLISS is tailored to the traditional and online broadcast media and video production industries. The BLISS core engine is also intended as a Language Identification system for broadcast monitoring (BLIS) (Connolly et al., 2014), a software application which identifies the language of the broadcast and alerts and even corrects if it is the wrong language for a particular region. In this paper we focus on its Subtitling application (BLSS).

Key problems BLISS addresses include:

a) Latency and hence speed of transcription from spoken word to text output.
b) Accuracy and compliance performance to mitigate reputation damage and financial penalties.
c) Reduction of costs through automation of human-based manual transcription.

In this paper section 2 gives a brief review of recently developed ASR technology. In Section 3, the design, architecture and implementation of BLISS is discussed. Section 4 discusses experimental results from testing BLISS and the impact of new unseen accents and noise. Section 5 concludes and discusses future work.

2. BACKGROUND & RELATED WORK

ASR is concerned with developing technologies that enable the recognition and translation of spoken language into text by computerised systems. ASR is also referred to as automatic Speech-To-Text (STT). In recent years, ASR technology has made remarkable progress, but the design of ASR systems still needs to pay careful attention to problems such as accents and noise.

Most ASR systems rely on phoneme recognition and word decoding. Classification algorithms (e.g., GMMs) are used on highly specialised features such as Mel Frequency Cepstral Coefficients (MFCCs) or Perceptual Linear Predictive coefficients (PLPs) so that a distribution of possible phonemes for each frame can be obtained (Kaur et al., 2016; Karpagavalli & Chandra, 2016). A HMM with a pre-trained language model is used to find the most likely sequence of phonemes that can be mapped to the output words during the decoding phase. HMMs are often utilised to handle the temporal variability of speech, and have been popular because they are flexible, versatile, and have a consistent statistical framework (Mohamed et al., 2009, 2012; Stevenson, 2016).

An alternative to GMMs evaluating performance is a feed-forward neural network, which takes several frames of coefficients as input and produces posterior probabilities over HMM states as output. DNNs have proven successful for acoustic modelling in speech recognition especially for large-scale tasks with examples being DNN-HMM hybrid systems (Hinton et al., 2006, 2012; Hinton & Salakhutdinov, 2006; Woodland et al., 2015), CNN-HMM hybrid systems (Sercu & Goel, 2016) and end-to-end ASR systems (Song & Cai, 2015; Collobert et al., 2016; Liptchinsky et al., 2017).

2.1 DNN-HMM hybrid systems

By taking advantage of DNN discriminative power, several successful hybrid DNN-HMM ASR systems have been developed for phoneme recognition (Hinton et al., 2012, Pan et al., 2012). DNNs have been used for acoustic model likelihood computation. Here DNNs outperformed traditional GMMs in predicting emission probabilities of HMM states representing phonemes in a hybrid model setup. DNNs have given promising results for large vocabulary continuous speech recognition (LVCSR) tasks, showing significant gains over GMM/HMM systems on a wide variety of small and large vocabulary tasks (Seide et al., 2011; Dahl et al., 2011, 2012, 2013; Li et al., 2013; Jaitly et al., 2012; Sainath et al., 2013; Zhang & Woodland, 2015).

A trained DNN output is not the end result of an ASR system, but instead supplies a HMM with the best acoustic modelling information to predict the target HMM states. The advantage of DNNs over GMMs in ASR is their ability to predict many thousands of tied triphone HMM states. This creates a large number of HMM classes and also inherently adds to the amount of training data and time needed to initialise a DNN-HMM system. For the 2015 Multi-Genre Broadcast (MGB) challenge, Woodland et al. (2015) outline a speech to text model containing a segmentation system based on DNNs. The model uses HTK 3.5 for building the DNN-based hybrid and tandem acoustic model in a joint decoding framework. The final system had the lowest (23.7%) Word Error Rate (WER) metric error rate for speech-to-text transcription on the MGB evaluation data (Bell et al., 2015).

2.2 CNN-HMM hybrid systems

Convolutional Neural Networks (CNNs) can be used to model correlation between spatial and temporal signals, and reduce spectral variance in acoustic features for ASR. Hybrid ASR systems incorporating CNNs with HMMs/GMMs have achieved promising results with various benchmarks (Abdel-Hamid et al., 2013; Sainath et al., 2013). CNNs are a more effective model for speech compared to extensively used fully-connected acoustic DNN models. The number of convolutional layers, the optimal number of hidden units along with the best pooling strategy, and the best input feature type for CNNs should all be considered. Comparing CNNs to DNNs and GMMs show that CNNs can have a 13-30%

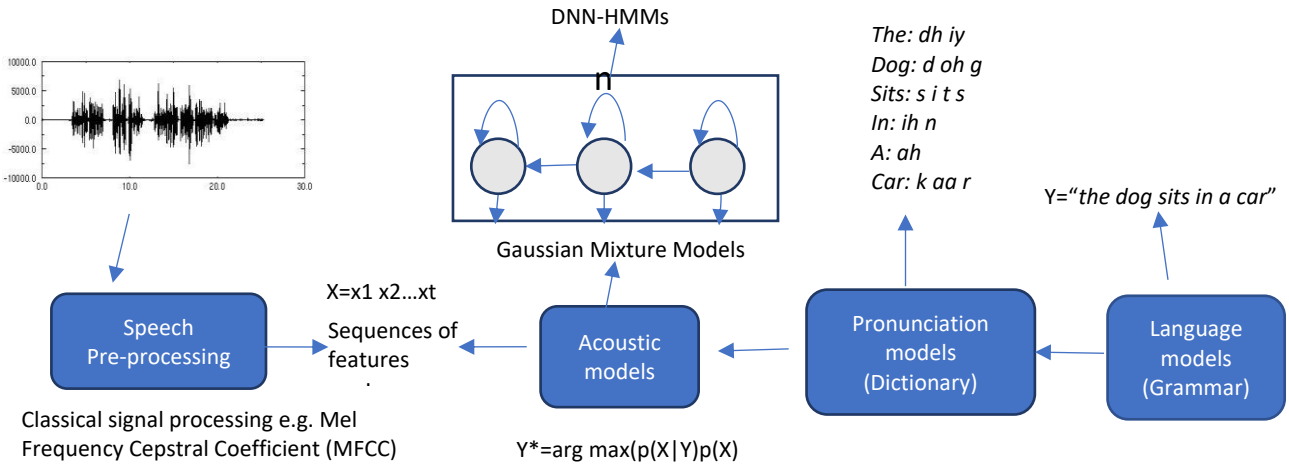

Figure 1: *Architecture of BLISS*

improvement over GMMs, and a 4-12% improvement over DNNs, on a variety of LVCSR tasks such as the 400 hours Broadcast News and 300 hours Switchboard tasks (Sercu & Goel, 2016).

2.3 End-to-End systems

Hybrid systems have been developed for phoneme recognition and decoding by HMMs. Recurrent neural networks (RNNs) can handle recognition and decoding simultaneously. Connectionist Temporal Classification (CTC) with RNNs (Zhang & Pezeshki, 2016) for labelling unsegmented sequences makes it feasible to train an 'end-to-end' ASR system instead of using hybrid settings. However, RNNs are computationally expensive and sometimes difficult to train. Inspired by the advantages of both CNNs and the CTC approach, an end-to-end ASR model was developed for sequence labelling, by combining hierarchical CNNs with CTC directly without recurrent connections. In evaluating the approach in the TIMIT phoneme recognition task, this model is not only computationally efficient, but also competitive with existing baseline systems. Moreover, CNNs have the capability to model temporal correlations with appropriate context information. LVCSR systems perform differently in terms of accuracy depending on the ASR task.

Clean read speech data gives better results than broadcast speech data. There are several subtitling tools on the market that enable the conversion of audio into text, e.g., the SAVAS project automatic live and batch subtitling application for several European languages such as Basque & Portugese (Alvarez et al., 2016). However, we still do not have human-level ASR systems for the automated subtitling task (Maxwell, 2018).

3. BLISS DESIGN & IMPLEMENTATION

Here we discuss the Design and Implementation of BLISS in terms of requirements analysis, architecture, and implementation with the Kaldi Toolkit and LibriSpeech dataset for model training and testing.

3.1 Customer requirements analysis

We have conducted requirements analysis and customer conversations with 25 Vendors and End Users within the 2017 ICURe NI LIC Lean Launch programme hosted by the SETsquared Partnership, Ulster University and Queen's University Belfast. Benchmarking of subtitle quality requires at least 95% accuracy in terms of WER or NER (Number Edition Recognition) metrics (Alvarez et al., 2016), with some customers requiring 100%. Vendors and End Users quote subtitler charging costs of e.g. ~£3.50/min., and sales pricing of e.g. ~£550/hr. with differences depending on nature of media content. Most Vendors access an external bank of outsourced subtitlers with internal staff quality checking and packaging. Vendors are currently investigating solutions that use cutting-edge ASR technology to automatically transcribe speech and format it for subtitling and captioning purposes, e.g. Red Bee Media (Maxwell, 2018). Any ASR system that reduces the cost, time and penalties for subtitling (Ofcom, 2013) would be of huge benefit to the subtitling industry.

3.2 BLISS architecture design

Figure 1 illustrates the modules in BLISS. Raw speech is converted to sequences of feature vectors using classical signal processing methods, e.g. MFCCs. The data X is a sequence of frames of audio features x1, x2…xt. *Y* represents text sequences. Using the language models, a sequence of words is produced. In the pronunciation dictionary for each word there is a pronunciation model for how this word is spoken. The pronunciation model with associated probabilities is written as a sequence of phonemes (or pronunciation tokens) which are basic unit of

sound, and then converted into sequences of text with corresponding pronunciation tokens. The models are fed into an acoustic model to test token sounds. Acoustic models are typically built using three state left-to-right GMMs which output frames of data. When the acoustic model is built, recognition can be performed by conducting inference on data received. For example, when some waveforms are received and their features (*X*) are extracted, the acoustic model deciphers the sequence of Y's that would cause this sequence of *X* with the highest probability. Traditionally, each of the components shown in Figure 1 were completed with traditional statistical methods (e.g., HMMs) but neural networks have recently proven superior.

3.3 Implementation of BLISS

Here we discuss the implementation of BLISS in terms of LibriSpeech datasets for training and testing, Language Model (LM) and Acoustic Model.

3.3.1. Training and testing datasets
BLISS was trained and tested on the LibriSpeech[1] dataset, and in addition a USA Alabama broadcast News dataset was used for testing. Table 1 gives details on the LibriSpeech datasets.

***Table 1**: LibriSpeech dataset (Panayotov et al., 2015)*

Data	Subset	Hours	Mins./ Spkr.	Gender	
				F	M
Train	train_100c	100.6	25	125	126
	train_360c	363.6	25	439	482
	train_500h	496.7	30	564	602
Test	test_c	5.4	8	20	20
	test_h	5.1	10	17	16

For example, the *train_100c* dataset includes ~100 hours of clean speech data, 28,539 utterances with 990,101 words that are spoken by 125 female and 126 male speakers. Each speaker spoke for 25 minutes. Speakers in the LibriSpeech corpus were ranked according to transcript WER and were divided roughly in the middle. The lower-WER speakers were designated as, *clean (c),* and the higher-WER speakers were designated as, *challenge (h)*. The *test_c* dataset includes ~5.4 hours of unseen clean test data, 2,620 utterances with 52,576 words that are spoken by 20 females and 20 males. Each speaker spoke for 8 minutes. The *test_h* dataset includes ~5.1 hours of unseen challenge test data, 2,939 utterances with 52,343 words that are spoken by 17 females and 16 males. Each speaker spoke for 10 minutes.

3.3.2. BLISS Language Model (LM)[2]
The full, non-pruned 3-gram LM was trained using the most frequent 200k word vocabulary from 14,500 public domain books in which about 803 million tokens and 900k unique words were selected (Panayotov et al., 2015). The pronunciation lexicon includes a 206,510 word pronunciation dictionary as some words have more than one pronunciation.

3.3.3. BLISS Acoustic Models
train_100c (~29k utterances) and data subsets (*2k, 5k* and *10k*) were used to train BLISS early-stage acoustic models. For the monophone stages of acoustic model development the shortest utterances were selected to facilitate data alignment from a flat start. The *5k, 10k, train_100c* and the entire about 960 hours training subset utterances were incrementally aligned using preceding built models. Model *tri5b* is a Speaker Adapted Training (SAT) model that was built on Feature space Maximum Likelihood Linear Regression (FMLLR)-adapted features on the ~960 hours mixed data.

4. EXPERIMENTAL RESULTS

The BLISS *tri5b* model was tested on both the unseen clean (*test_c*) and challenge (*test_h*) test datasets listed in Table 1 and it gave WER results of 8.13% and 24.09% respectively. The USA Alabama News data was used to test background noise and accent effects, e.g. *AN2088*: "there was that survey which said that the united states wasn't going to use nukes to help south korea", with all words in the pronunciation dictionary except "nukes".

Figure 2 shows the performance of *tri5b* in terms of WER results without and with unseen accent, noise and music on the unseen USA Alabama News audio utterance *AN2088* using the full, non-pruned 3-gram BLISS LM. *test_c*, *test_h* and *AN2088* are all unseen during training. However, although the training data has audio data with accents found in *test_c* and *test_h*, accents from the USA Alabama news English audio data are unseen during *tri5b* model training. BLISS performance on USA Alabama News data with new unseen accents, without or with white noise and music gives WER results of 52.63%, 57.89% and 94.74% respectively. Figure 2 shows background noise and music affect performance, with music having a greater effect than noise.

A BLISS *DNN* model trained with the BLISS *train_100c* acoustic model using the setup discussed in Povey et al. (2015) was compared to the BLISS *tri4b* model also built using *train_100c*. Results show the BLISS *DNN* model achieves better

[1] The LibriSpeech American accent English dataset can be downloaded from: *http://www.openslr.org/12/*.

[2] The LibriSpeech LM can be downloaded from: *http://www.openslr.org/11/*.

performance than the BLISS *tri4b* model using the same volume of acoustic training data (100 hours) with the full, non-pruned 3-gram LM. WER for the BLISS *DNN* model decreases to 7.12% from that using the BLISS *tri4b* model (9.74%) for *test_c* clean audio data. WER for the *DNN* model decreases to 22.73% from that using BLISS *tri4b* model (32.93%) for *test_h* challenge audio data.

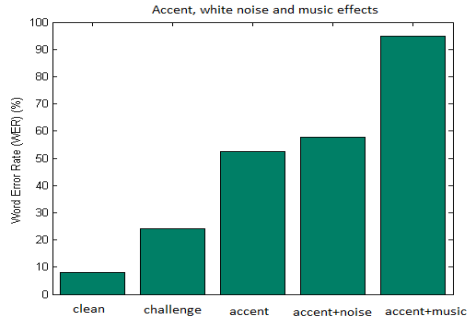

Figure 2: *Effect of unseen accents, white noise and background music on WER*

5. CONCLUSION & FUTURE WORK

In this paper we discussed the design and implementation of a software platform called BLISS (Broadcast Language Identification & Subtitling System) for performing subtitling (BLSS) and language identification (BLIS) within the broadcast entertainment industry, with a focus on subtitling. BLISS is based on customer requirements analysis conducted with more than 25 Vendors and End Users. The LibriSpeech USA English audio book read speech and USA Alabama broadcast News datasets were used to build and test the BLISS acoustic models using the Kaldi ASR Toolkit. BLISS gives promising WER metric results of 8.13% and 24.09% respectively on ~5 hours of unseen test clean (*test_c*) and unseen test challenge (*test_h*) audio data subsets from the LibriSpeech corpus. BLISS performance on USA Alabama broadcast News data with new unseen accents gives WER results of 52.63%. Our experiments demonstrate that speech with new unseen accents, background noise and music degrade BLISS model performance, with music degrading more than white noise. The BLISS *DNN* model performs better than the BLISS GMM/HMM *tri4b* model giving WER results of 7.12% (over 9.74%) with *test_c*, and 22.73% (over 32.93%) with *test_h*. Future work includes developing further BLISS DNNs models and methods for noise removal.

6. ACKNOWLEDGEMENTS

BLISS R&D has been funded by Invest NI Proof of Concept (PoC) awards PoC-607 (BLISS) & PoC-318 (Song Form Intelligence) & NI LIC ICURe travel funds. We would like to thank Paul Malcolmson of Invest NI; Fergus Begley & Dr. John Macrae of Ulster University Office of Innovation; Alan Scrase & Don Spalinger from SETsquared; BLISS Mentors Stephen Craig, Ben Dair, Andrew Lambourne, Dr. Pablo Romero-Fresco & Mark Caldwell; MC+P Consulting, Prof. Mike McTear & Dr. Arantza del Pozo for Market Assessment & Dr. Junxiu Liu & Mike McCool for technical support.

7. REFERENCES

Abdel-Hamid, O., Deng, L. & Yu, D. (2013). Exploring convolutional neural network structures and optimization for speech recognition. Interspeech, ISCA, Lyon, France, 3366-3370.

Alvarez, A., Mendes, C., Raffaelli, M., Luıs, T., Paulo, S., Piccinini, N., Arzelus, H., Neto, J., Aliprandi, C. & del Pozo, A. (2016). Automating live and batch subtitling of multimedia contents for several European languages. Multimed. Tools Appl., Vol. 75, 10823–10853.

Bell, P., Gales, MJF., Hain, T., Kilgour, J., Lanchantin, P., Liu, X., McParland, A., Renals, S., Saz, O., Wester, M. & Woodland, PC. (2015). The MGB challenge: Evaluating multi-genre broadcast media recognition. In Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU2015), Scottsdale, Arizona, USA, 687-693.

Collobert, R., Puhrsch, C. & Synnaeve, G. (2016). Wav2letter: an end-to-end convnet-based speech recognition system. CoRR, Vol. abs/1609.03193.

Connolly, J., Curran, K., McKevitt, P., Macrae, J. & Craig, S. (2014). Broadcast Language Identification System (BLIS). In: Proc. of the 16th Irish Machine Vision and Image Processing Conference (IMVIP-14), Ulster University, UK.

Dahl, G., Yu, D., Deng, L. & Acero, A. (2011). Large vocabulary continuous speech recognition with context-dependent DBN-HMMs. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 4688-4691.

Dahl, G., Yu, D., Deng, L. & Acero, A. (2012). Context Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition, Vol. 20, No.1, 30–42.

Dahl, G., Sainath, T. & Hinton, G. (2013). Improving DNNs for LVCSR using rectified linear units and dropout. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, ca8609-8613.

Fryer, L. (2018). An introduction to Audio Description: a practical guide. New York, USA: Routledge.

Hinton, G., Osindero, S. & The, Y. (2006). A fast learning algorithm for deep belief nets. Neural Computation, Vol. 18, 1527-1554.

Hinton, G. & Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks, Science. Vol. 313, No. 5786, 504-507.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N. & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modelling in Speech Recognition. IEEE Signal Processing Magazine, Vol. 29, No. 6, 82–97.

Jaitly, N., Nguyen, P., Senior, A. W. & Vanhoucke, V. (2012). Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition. The 13th International Speech Communication Association, in Proc. Interspeech, New York, USA, 2578-2581.

Jaitly, N., (2017). Lecture 12: End-to-End Models for Speech Processing, Stanford University School of Engineering, https://www.youtube.com/watch?v=3MjIkWxXigM&app=desktop.

Kaldi (2018). http://kaldi-asr.org/doc/index.html.

Karpagavalli, S. & Chandra, E. (2016). A Review on Automatic Speech Recognition Architecture and Approaches, International Journal of Signal Processing, Image Processing and Pattern Recognition. Vol. 9, No. 4, 393-404.

Kaur, I., Kaur, N., Ummat, A., Kaur, J., Navjot, K. (2016). Automatic Speech Recognition: A Review. InternatIonal Journal of Computer SclEnce and Technology (IJCST), Vol. 7, Issue 4, Oct.-Dec.

Li, D., Hinton, G. & Kingsbury, B. (2013). New Types of Deep Neural Network Learning for Speech Recognition and Related Applications: An Overview. IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 8599-8603.

Li, D., Li, J., Huang, J., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., Gong, Y. & Acero, A. (2013). Recent Advances in Deep Learning for Speech Research at Microsoft. IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, B.C., Canada, 8604-8608.

Liptchinsky, V., Synnaeve, G. & Collobert, R. (2017). Letter-Based Speech Recognition with Gated ConvNets. CoRR, Vol. abs/1712.09444.

Maxwell, H. (2018). Can we talk about the other 7%? https://www.redbeemedia.com/blog/can-talk-7/, Red Bee Media Blog.

Mohamed, A., Dahl, G. & Hinton, G. (2009). Deep belief networks for phone recognition. In Proc. NIPS Workshop on Deep Learning for Speech Recognition and Related Applications. Vancouver, B. C., Canada, 1-9.

Mohamed, A., Dahl, G. & Hinton, G. (2012). Acoustic modeling using deep belief networks. IEEE Trans. on Audio, Speech, and Language Processing, Vol. 20, No. 1, 14–22.

Ofcom (2013). Measuring the quality of live subtitling, https://www.ofcom.org.uk/__data/assets/pdf_file/0017/51731/qos-statement.pdf.

Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In International Conference on Acoustics, Speech and Signal Processing (ICASSP), Queensland, Australia, 5206–5210.

Pan, J., Liu, C., Wang, Z., Hu, Y. & Jiang, H. (2012). Investigation of Deep Neural Networks (DNN) for Large Vocabulary Continuous Speech Recognition: Why DNN Surpasses GMMS in Acoustic Modelling. In Proc. of 8th International Symposium on Chinese Spoken Language Processing (ISCSLP'2012), Hong Kong, 301-305.

Povey, D., Zhang, X. & Khudanpur, S. (2015). Parallel Training of Deep Neural Networks with Natural Gradient and Parameter Averaging. In Proc. of 3rd International Conference on Learning Representations (ICLR2015), San Diego, USA.

Romero-Fresco, P. (2014). Subtitling through speech recognition: respeaking. Manchester, UK: St. Jerome Publishing.

Sainath, T., Mohamed, A., Kingsbury, B. & Ramabhadran, B. (2013) Deep convolutional neural networks for lvcsr. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 8614-8618.

Seide, F., Li, G. & Yu, D. (2011). Conversational Speech Transcription Using Context-Dependent Deep Neural Networks. In Proc. Interspeech, Florence, Italy, 444-447.

Sercu, T. & Goel, V. (2016). Advances in Very Deep Convolutional Neural Networks for LVCSR. Multimodal Algorithms and Engines Group, IBM, T.J. Watson Research Center, USA.

Song, W. & Cai, J. (2015). End-to-End Deep Neural Network for Automatic Speech Recognition. Technical Report, Department of Computer Science, Stanford University.

Stevenson, G. A. (2016). Aalysis of Pre-Trained Deep Neural Networks for Large-Vocabulary Automatic Speech Recognition. LLNL-TH-698797 July 28, Lawrence Livermore National Laboratory.

Woodland, P. C., Liu, X., Qian, Y., Zhang, C., Gales, M., Karanasou, P., Lanchantin, P., Wang, L. (2015). Cambridge University Transcription Systems for the Multi-Genre Broadcast Challenge, Automatic Speech Recognition and Understanding (ASRU), IEEE Automatic Speech Recognition and Understanding Workshop, Scottsdale, Arizona, USA, 639-646.

Zhang, C. & Woodland, P. C. (2015). A general artificial neural network extension for HTK. In Proc. Interspeech, Dresden, Germany, 3581-3585.

Zhang, Y. & Pezeshki, M. (2016). Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks. In Proc. Interspeech, San Francisco, USA, 410-414.