ACHIEVING HUMAN PARITY IN CONVERSATIONAL SPEECH RECOGNITION

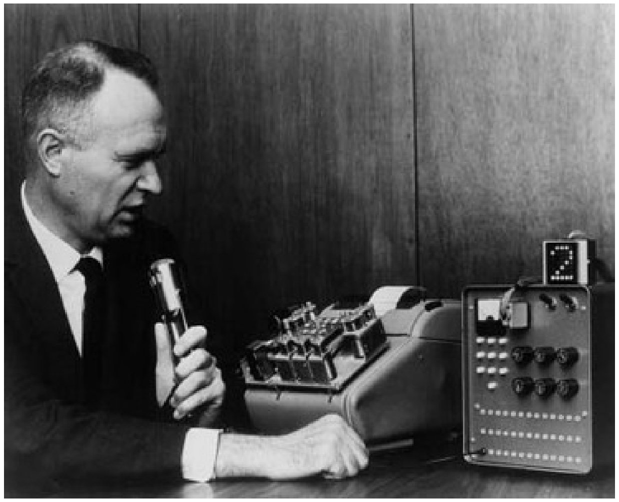
Paper by: W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu and G. Zweig
Presented by: Samin Yasar

WHAT DOES IT MEAN?

- Human Parity is the condition of being equal to humans.
- Speech Recognition (is also known as Automatic Speech Recognition (ASR), or computer speech recognition) is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program.
- Hence, achieving human parity in conversational speech recognition is the state a computer reaches when it can identify everyday human speech as well as a regular human being can.
- This topic falls under the field of Artificial Intelligence, more specifically Machine Learning.

WHAT IS ITS BACKGROUND?

- Digit Recogniser(1952) –recognizing spoken numerical digits.
- Shoebox by IBM(1960’s) – recognize digits and arithmetic commands
- Whither Speech Recognition(1969)
- Speech Understanding research by DARPA(1970’s)
- Hidden Markov Model(1980’s)
- Dragon Dictate (1990’s) – recognize 30-40 words a minute
- Whither Speech Recognition: The Next 25 Years(1993)

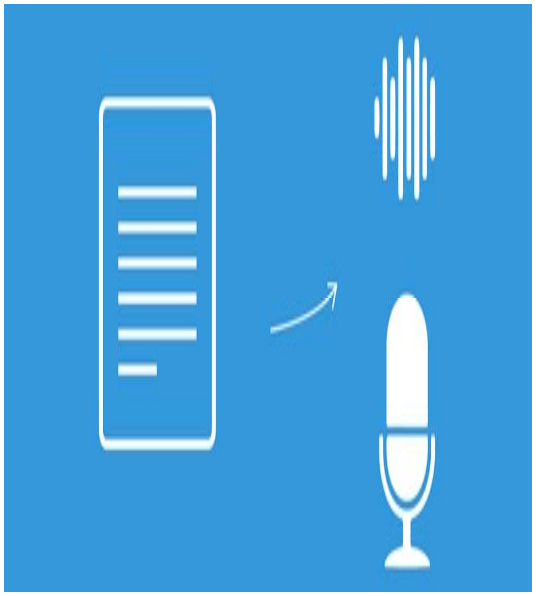


WHAT ARE ITS APPLICATIONS AND WHAT DRIVES ITS DEVELOPMENTS?

- In today's world, speech recognition systems have a wide variety of applications in different sectors including education where physically handicapped students unable to type use it to enter text verbally, medical sector where medical transcriptions are are done verbally and in the tech sector where it is used in robots and in digital assistants in smartphones.
- Other sectors using speech recognition includes the banking sector, tourism, military, communication sector, etc.
- Although it is this widely used, speech recognition isn't 100% efficient. The desire to make it as perfect as possible as well as its growing demands drives the development of this topic.

WHAT DISTINGUISHES THIS TOPIC FROM ASSOCIATED TOPICS?

- Speech generation is different from associated topics like speech generation(TTS) and natural language processing(NLP).
- Speech recognition is used for dictation purposes while NLP is used for tasks like automatic summarization and information retrieval and TTS is used for tasks such as voice enabled email and radio broadcasting .
- Popular examples of speech recognition include Windows speech recognition and dragon while examples of NLP include digital assistants such as Siri while examples of TTS include Ivona and Natural reader,



OBJECTIVE ADDRESSED BY THIS PAPER AND PRIOR RESEARCH DONE TOWARDS THIS GOAL

- This paper talks about how the latest automated systems today have reached human parity and describes how it has done so by explaining about the different algorithms, tools and techniques used in the system.
- This paper expands upon the the research paper done by Microsoft called "The Microsoft 2016 Conversational speech recognition system".
- Other papers about research done towards this goal include "Transcription methods for consistency, volume and efficiency", "Very deep convolutional networks for large-scale image recognition", Front-end factor analysis for speaker verification", etc.

HUMAN PERFORMANCE

- Two pass transcription used where a transcriber transcribed data from scratch on the first pass and on the second pass a second transcriber does error correction.
- NIST 2000 Test set used
- Same Audio segment given to the speech recognition system.
- 5.9% for error rate for switchboard portion and 11.3% for the CallHome portion.
- It was observed that the the performance of the artificial system aligns almost exactly with the performance of people on both sets.

CONVOLUTIONAL AND LSTM NEURAL NETWORKS

- 3 types of CNNs used:

1.VGG architecture – uses smaller 3x3 filter, deeper and applies up to 5 convolutional layer

2. ResNet architecture – adds a a linear transformation each layer's input to the layer's output

3. LACE(layer-wise context expansion) model which is a Time Delay neural network.

- Bidirectional LSTMs(Long short-term memory) used
- Spatial smoothing - data points are averaged with their neighbours

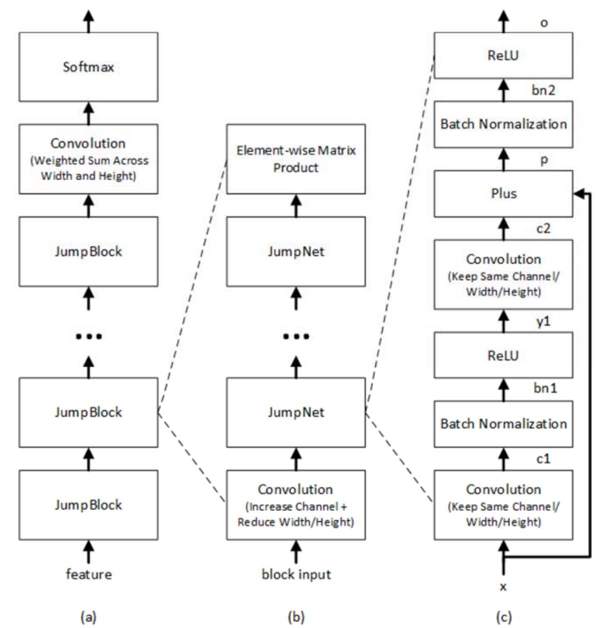


Fig. 1. LACE network architecture

Table 1. Comparison of CNN architectures

VGG Net (85M Parameters)	**Residual-Net (38M Parameters)**	**LACE (65M Parameters)**
14 weight layers	49 weight layers	22 weight layers
40x41 input	40x41 input	40x61 input
3 – conv 3x3, 96	3 – [conv 1x1, 64 conv 3x3, 64 conv 1x1, 256]	5 – conv 3x3, 128
Max pool	4 – [conv 1x1, 128 conv 3x3, 128 conv 1x1, 512]	5 – conv 3x3, 256
4 – conv 3x3, 192	6 – [conv 1x1, 256 conv 3x3, 256 conv 1x1, 1024]	5 – conv 3x3, 512
Max pool	3 – [conv 1x1, 512 conv 3x3, 512 conv 1x1, 2048]	5 – conv 3x3, 1024
4 – conv 3x3, 384	Average pool	1 – conv 3x4, 1
Max pool	Softmax (9000)	Softmax (9000)
2 – FC – 4096		
Softmax (9000)		

SPEAKER ADAPTIVE MODELING

- i-vector used
- Learnable weight matrix added
- Log-filterbank features

Table 3. Performance improvements from i-vector and LFMMI training on the NIST 2000 CTS test set

Configuration	WER (%)							
	ReLU-DNN		ResNet-CNN		BLSTM		LACE	
	CH	SWB	CH	SWB	CH	SWB	CH	SWB
Baseline	21.9	13.4	17.5	11.1	17.3	10.3	16.9	10.4
i-vector	20.1	11.5	16.6	10.0	17.6	9.9	16.4	9.3
i-vector+LFMMI	17.9	10.2	15.2	8.6	16.3	8.9	15.2	8.5

LATTICE FREE SEQUENCE TRAINING

- Cross entropy training
- optimizing the model parameters using the maximum mutual information (MMI) objective function.
- Perform a forced alignment of the training data to select lexical variants and determine frame-aligned senone sequences.
- Compress consecutive framewise occurrences of a single senone into a single occurrence.
- Estimate an unsmoothed, variable-length N-gram language model from this data, where the history state consists of the previous phone and previous senones within the current phone.

LM RESCORING AND SYSTEM COMBINATION

- RNN-LM setup
- LSTM-LM setup
- Training data
- RNN-LM and LSTM-LM performance
- System Combination

Table 4. LSTM perplexities (PPL) as a function of hidden layers, trained on in-domain data only, computed on 1997 CTS eval transcripts.

Language model	PPL
letter trigram input with one layer (baseline)	63.2
+ two hidden layers	61.8
+ three hidden layers	59.1
+ four hidden layers	59.6
+ five hidden layers	60.2
+ six hidden layers	63.7

Table 5. Perplexities (PPL) of the four LSTM LMs used in the final combination. PPL is computed on 1997 CTS eval transcripts. All the LSTM LMs are with three hidden layers.

Language model	PPL
RNN: 2 layers + word input (baseline)	59.8
LSTM: word input in forward direction	54.4
LSTM: word input in backward direction	53.4
LSTM: letter trigram input in forward direction	52.1
LSTM: letter trigram input in backward direction	52.0

Table 6. Performance of various versions of neural-net-based LM rescoring. Perplexities (PPL) are computed on 1997 CTS eval transcripts; word error rates (WER) on the NIST 2000 Switchboard test set.

Language model	PPL	WER
4-gram LM (baseline)	69.4	8.6
+ RNNLM, CTS data only	62.6	7.6
+ Web data training	60.9	7.4
+ 2nd hidden layer	59.0	7.4
+ 2-RNNLM interpolation	57.2	7.3
+ backward RNNLMs	-	6.9
+ LSTM-LM, CTS + Web data	51.4	6.9
+ 2-LSTM-LM interpolation	50.5	6.8
+ backward LSTM-LM	-	6.6

MICROSOFT COGNITIVE TOOLKIT (CNTK)

- Flexible, Terse Model Definition
- Multi-Server Training using 1-bit SGD
- Computational performance

Table 7. Runtimes as factor of speech duration for various aspects of acoustic modeling and decoding, for different types of acoustic model

Processing step	Hardware	DNN	ResNet-CNN	BLSTM	LACE
AM training	GPU	0.012	0.60	0.022	0.23
AM evaluation	GPU	0.0064	0.15	0.0081	0.081
AM evaluation	CPU	0.052	11.7	n/a	8.47
Decoding	GPU	1.04	1.19	1.40	1.38

EXPERIMENTS AND RESULTS

- Speech corpora
- Acoustic Model Details
- Overall Results and Discussion

Table 9. Comparative error rates from the literature and human error as measured in this work

Model	N-gram LM		Neural net LM	
	CH	SWB	CH	SWB
Povey et al. [54] LSTM	15.3	8.5	-	-
Saon et al. [51] LSTM	15.1	9.0	-	-
Saon et al. [51] system	13.7	7.6	12.2	6.6
2016 Microsoft system	13.3	7.4	11.0	5.8
Human transcription			11.3	5.9

Table 8. Word error rates (%) on the NIST 2000 CTS test set with different acoustic models. Unless otherwise noted, models are trained on the full 2000 hours of data and have 9k senones.

Model	N-gram LM		RNN-LM		LSTM-LM	
	CH	SWB	CH	SWB	CH	SWB
ResNet, 300h training	19.2	10.0	17.7	8.2	17.0	7.7
ResNet	14.8	8.6	13.2	6.9	12.5	6.6
ResNet, GMM alignments	15.3	8.8	13.7	7.3	12.8	6.9
VGG	15.7	9.1	14.1	7.6	13.2	7.1
VGG + ResNet	14.5	8.4	13.0	6.9	12.2	6.4
LACE	15.0	8.4	13.5	7.2	13.0	6.7
BLSTM	16.5	9.0	15.2	7.5	14.4	7.0
BLSTM, spatial smoothing	15.4	8.6	13.7	7.4	13.0	7.0
BLSTM, spatial smoothing, 27k senones	15.3	8.3	13.8	7.0	13.2	6.8
BLSTM, spatial smoothing, 27k senones, alternate dictionary	14.9	8.3	13.7	7.0	13.0	6.7
BLSTM system combination	13.2	7.3	12.1	6.4	11.6	6.0
Full system combination	13.0	7.3	11.7	6.1	**11.0**	**5.8**

ERROR ANALYSIS

- compare the errors made by the artificial recognizer with those made by human transcribers
- machine errors are substantially the same as human ones, with one large exception: confusions between backchannel words and hesitations.
- It is speculated that this is due to the nature of the Fisher training corpus, where the "quick transcription" guidelines were predominately used
- We see that the human transcribers have a somewhat lower substitution rate, and a higher deletion rate.

Table 13. Overall substitution, deletion and insertion rates.

	CH		SWB	
	ASR	**Human**	**ASR**	**Human**
sub	6.5	4.1	3.3	2.6
del	3.3	6.5	1.8	2.7
ins	1.4	0.7	0.7	0.7
all	11.1	11.3	5.9	5.9

Table 10. Most common substitutions for ASR system and humans. The number of times each error occurs is followed by the word in the reference, and what appears in the hypothesis instead.

CH		SWB	
ASR	**Human**	**ASR**	**Human**
45: (%hesitation) / %bcack	12: a / the	29: (%hesitation) / %bcack	12: (%hesitation) / hmm
12: was / is	10: (%hesitation) / a	9: (%hesitation) / oh	10: (%hesitation) / oh
9: (%hesitation) / a	10: was / is	9: was / is	9: was / is
8: (%hesitation) / oh	7: (%hesitation) / hmm	8: and / in	8: (%hesitation) / a
8: a / the	7: bentsy / bensi	6: (%hesitation) / i	5: in / and
7: and / in	7: is / was	6: in / and	4: (%hesitation) / %bcack
7: it / that	6: could / can	5: (%hesitation) / a	4: and / in
6: in / and	6: well / oh	5: (%hesitation) / yeah	4: is / was
5: a / to	5: (%hesitation) / %bcack	5: a / the	4: that / it
5: aw / oh	5: (%hesitation) / oh	5: jeez / jeeze	4: the / a

Table 11. Most common deletions for ASR system and humans.

CH		SWB	
ASR	**Human**	**ASR**	**Human**
44: i	73: i	31: it	34: i
33: it	59: and	26: i	30: and
29: a	48: it	19: a	29: it
29: and	47: is	17: that	22: a
25: is	45: the	15: you	22: that
19: he	41: %bcack	13: and	22: you
18: are	37: a	12: have	17: the
17: oh	33: you	12: oh	17: to
17: that	31: oh	11: are	15: oh
17: the	30: that	11: is	15: yeah

Table 12. Most common insertions for ASR system and humans.

CH		SWB	
ASR	**Human**	**ASR**	**Human**
15: a	10: i	19: i	12: i
15: is	9: and	9: and	11: and
11: i	8: a	7: of	9: you
11: the	8: that	6: do	8: is
11: you	8: the	6: is	6: they
9: it	7: have	5: but	5: do
7: oh	5: you	5: yeah	5: have
6: and	4: are	4: air	5: it
6: in	4: is	4: in	5: yeah
6: know	4: they	4: you	4: a

RESULT

- The automatic speech recognition had a variable rate of 5.8% and11.0% for Switchboard and CallHome subsets respectively compared to the 5.9% and 11.3% for the professional transcribers.
- This is means that for the first time ASR performance is on par with actual human performance meaning human parity has been achieved.

THE PROS AND CONS

- **Pros**

I. Benefits people with visual and physical inabilities

II. Hands free

III. More time efficient as people are faster at speaking than typing

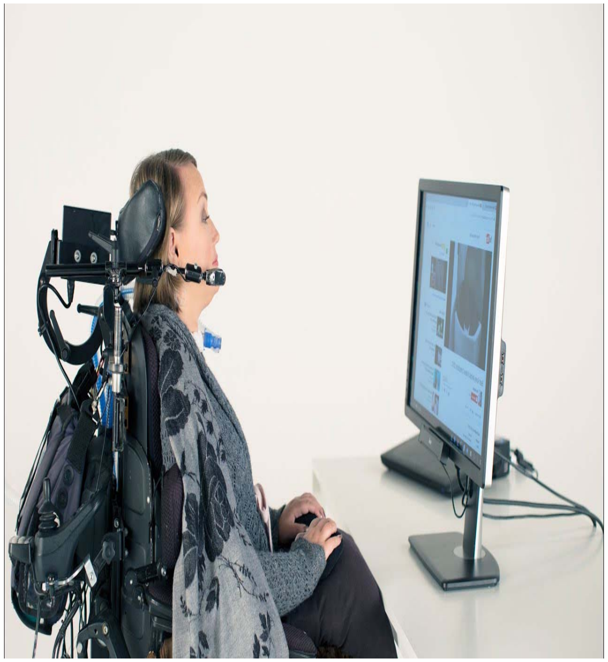
IV. Fewer spelling errors

- **Cons**

I. Errors can be a huge problem

II. Loss of jobs

III. People have to speak very clearly for the ASR to understand



FUTURE STEPS TO FULLY MEET ORIGINAL OBJECTIVE

- More training data
- Better algorithms
- Stacking more algorithms together
- Algorithm tuning
- Reframing the problems

CONCLUSION

This paper has shown that speech recognition has reached a level where it now on par, if not better than, with humans. It has shown that by combining different algorithms, VGG, ResNet and LACE, and techniques ,like Lattice free MMI, and using the latest tools and technology has enabled us to reach this level. Although human parity has been reached, this is not the end of this research as Speech Recognition is still a long way from being perfect.

- M. L. Glenn, S. Strassel, H. Lee, K. Maeda, R. Zakhary, and X. Li, “Transcription methods for consistency, volume and efficiency”, in LREC, 2010.
- K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, arXiv preprint arXiv:1409.1556, 2014.
- R. K. Srivastava, K. Greff, and J. Schmidhuber, “Highway networks”, CoRR, vol. abs/1505.00387, 2015.
- P. Ghahremani, J. Droppo, and M. L. Seltzer, “Linearly augmented deep neural network”, in Proc. IEEE ICASSP, pp. 5085–5089. IEEE, 2016.
- A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures”, Neural Networks, vol. 18, pp. 602–610, 2005.
- H. Sak, A. W. Senior, and F. Beaufays, “Long shortterm memory recurrent neural network architectures for large scale acoustic modeling”, in Proc. Interspeech, pp. 338–342, 2014.
- D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y.Wang, and S. Khudanpur, “Purely sequence-trained neural networks for ASR based on lattice-free MMI”, in Proc. Interspeech, pp. 2751–2755, 2016.
- X. Chen, X. Liu, Y. Qian, M. J. F. Gales, and P. C. Woodland, “CUED-RNNLM: An open-source toolkit for efficient training and evaluation of recurrent neural network language models”, in Proc. IEEE ICASSP, pp. 6000–6004. IEEE, 2016.
- A. Stolcke et al., “The SRI March 2000 Hub-5 conversational speech transcription system”, in Proceedings NIST Speech Transcription Workshop, College Park, MD, May 2000.
- Microsoft Research, “The Microsoft Cognition Toolkit (CNTK)”, https://cntk.ai.