WER we are and WER we think we are?

EMNLP 2020

Rethinking Evaluation in ASR: Are our models robust Enough?

Interspeech 2021

The History of Speech Recognition to the Year 2030

By Awni Hannun

Erfan Loweimi

WER we are and WER we think we are

Piotr Szymański[1,2,*], Piotr Żelasko[3,*], Mikołaj Morzy[1,4], Adrian Szymczak[1],
Marzena Żyła-Hoppe[1], Joanna Banaszczak[1], Łukasz Augustyniak[1,3]
Jan Mizgajski[1,4], Yishay Carmiel[1]
[1] Avaya Inc., USA
[2] Wrocław University of Science and Technology, Wrocław, Poland
[3] Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
[4] Poznań University of Technology, Poznań, Poland
[*] Equal contribution
piotr.szymanski@pwr.edu.pl, pzelasko@jhu.edu

EMNLP 2020

Quantifying the ASR-NLP Gap

- Is ASR a solved problem? Depends ...
- *Quality* of SOTA ASR systems is over-estimated ...
  - $\text{WER}_{\text{Real-world}}$ vs $\text{WER}_{\text{Research Benchmarks}}$
    - H2H* Spontaneous conversation
    - (semi-)scripted, read, artificial conversation
- Benchmarks … demographically homogeneous … reliable real-world diversity representative?

* H2H: human to human

Human-Haman vs Human-Machine

- Human-machine interaction ... artificial & static
  - Simplified/short utterances, well-structured phrases, correct, grammar, interrogative or imperative (request/response)
- Human-human interaction ... natural & dynamic
  - Disfluencies, lack clear borders, incorrect termination, richer vocabulary, communicate via non-verbal channels, etc.
- Acoustically distinguishable → 81% accuracy (Alexa)

Experimental Setup

- Compare $WER_{\text{Real-world}}$ vs $WER_{\text{Research Benchmarks}}$
  - using 3 commercial SOTA ASR [telephone speech]
- Real-world proxy
  - Data from 50 **c**all **c**entre **c**onversations (CCC)
  - 8 kHz, 2.2h speech, #utter: 1595+1361, avg #wrds/utt: 10
- Research Benchmarks proxy
  - Hub’05 [SWBD + CallHome]

Real-world vs Research Benchmarks
Performance gap

ASR	CCC	SWBD	CallHome
ASR 1	17.9	11.62	17.69
ASR 2	19.2	11.45	18.6
ASR 3	16.5	10.2	15.85

Kaldi (Hybrid):	8.8%	13.5%
SAHR* (E2E):	6.7%	13.7%
SOTA**:	5.0%	9.0%

* Commercial ASR WER is 2X SOTA. Why?
  - – General acoustic and language model
  - – [5-min chunks + SAD] vs oracle segmentation

	ASR 1	ASR 2	ASR 3
Booking	21.19	22.16	20.95
Finance	16.82	18.46	15.83
Insurance 1	18.01	20.20	17.84
Insurance 2	15.25	17.11	13.73
Telecomm.	19.75	23.31	17.62
Agent	16.97	17.83	16.49
Customer	17.87	20.99	16.48

Domain ↔ Performance

Why WER for booking is high?

SAHR*: Stochastic Attention Head Removal

Zhang et al., Interspeech 2021

**: Super-specific ASR system for SWBD

Conclusion

- ASR for spontaneous human-human conversation is challenging!
  - $\text{WER}_{\text{Real-world}}$ >[>] $\text{WER}_{\text{Research Benchmarks}}$
- Call to action
  - Crowd-sourcing → Mozilla Common Voice → phone calls + transcription donation
  - Construct new ASR quality measures
  - Designing joint ASR+NLP tasks
  - …

INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Rethinking Evaluation in ASR: Are Our Models Robust Enough?

Tatiana Likhomanenko[1*], *Qiantong Xu*[1*], *Vineel Pratap*[1], *Paden Tomasello*[1], *Jacob Kahn*[1], *Gilad Avidov*[1], *Ronan Collobert*[1], *Gabriel Synnaeve*[2]

[1]Facebook AI Research, USA
[2]Facebook AI Research, France
{antares,qiantong}@fb.com

Motivation – Research Question ...

- *" … Are our models robust enough?"*
  - Is pushing numbers on a single benchmark practically valuable?
  - Is WER on a single benchmark a good proxy for performance on real-world data?
  - Does ASR progress on research benchmarks mean progress in ASR over real-world applications?

Motivation – Robustness Means ...

- **Q**: Are our models *robust* enough?
- Robustness ↔ Handling Mismatch
  - Acoustic mismatch → noise (Additive, Channel, Reverberation)
  - Domain/Genre mismatch
  - Research/Real-world mismatch (this paper)
- Robustness (AM*) ↔ Generalisation (ML**)

* AM: Acoustic Modelling
** ML: Machine learning

This paper ...

- **Goal**: Study ...
  - Generalisation from research to real-life
  - Practical usefulness of low $\text{WER}_{\text{Research Benchmark}}$
- **How**:
  - Build SOTA AM/LM using single/joint research dataset(s)
  - Evaluate on various research/real-world datasets
  - … investigate ASR knowledge transfer …

Experimental Setup

- Acoustic model:
  - Architecture: Transformer (36T blocks with 4 heads, d_{model}=762)
  - Training dataset: Single & Joint (+ Fine-tuning: 1h, 10h, 100h)
  - Loss: CTC; Decoding: greedy & beam-search
- Optimiser: Adagrad + LR decay factor 2 (WER plateau)
- Dropout (SA and FFN) + layer drop (FFN)
- Token set: 26 Eng. letters + aposhtrophe + word boundary
- Data augmentation: SpecAug (freq + time masking)
- Toolkit: Kaldi, Flashlight & wav2letter++

Datasets (1)

Data	kHz	Train (h)	Valid (h)	Test (h)	Speech
WSJ	16	81.5	1.1	0.7	read
TL	16	452	1.6	2.6	oratory
CV	48	693	27.1	25.8	read
LS	16	960	5.1+5.4	5.4+5.4	read
SB+FSH	8	300+2k	6.3	1.7+2.1	convers.
RV	16	5k	14.4	18.8+19.5+37.2	diverse

- **Research**
  - **WSJ** (Read), **T**ED **L**IUM (oratory), Mozilla **C**ommon **V**oice (Read), **L**ibri**S**peech (Read), **S**witch**B**oard (telephone conversation)
- **Real-world**
  - Facebook's in-house **R**obust **V**ideo [**RV**] (social media)

Datasets (2)

	Sec			Words			
Data	Train $\mu \pm \sigma$ (s)	Valid $\mu \pm \sigma$ (s)	Test $\mu \pm \sigma$ (s)	Train $\mu \pm \sigma$ (wrd)	Valid $\mu \pm \sigma$ (wrd)	Test $\mu \pm \sigma$ (wrd)	#wrds/sec
WSJ	7.8 ± 2.9	7.8 ± 2.9	7.6 ± 2.5	17 ± 7	16 ± 7	17 ± 6	2.1
TL	6 ± 3	11.3 ± 5.7	8.1 ± 4.3	17 ± 10	35 ± 20	24 ± 15	3.0
CV	5.7 ± 1.6	6.1 ± 1.8	5.8 ± 2.6	10 ± 3	10 ± 3	9 ± 3	1.6
LS	12.3 ± 3.8	6.8 ± 4.5	7 ± 4.8	33 ± 12	19 ± 13	19 ± 13	2.7
SB+FSH	3.7 ± 3.2	4 ± 3.1	2.1 ± 1.7	11 ± 12	12 ± 12	8 ± 8	3.0
RV	8.5 ± 1.9	11.6 ± 2.8	11.6 ± 2.7	21 ± 10	25 ± 13	29 ± 12	2.3

- **Research**
  - **WSJ** (Read), **T**ED **L**IUM (oratory), Mozilla **C**ommon **V**oice (Read), **L**ibri**S**peech (Read), **S**witch**B**oard (telephone conversation)
- **Real-world**
  - Facebook's in-house **R**obust **V**ideo (social media)

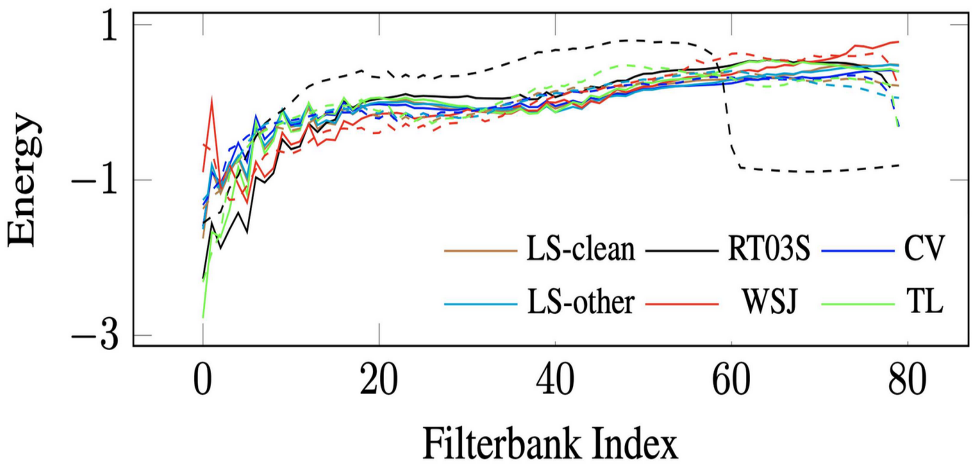
Language Model

- Architecture:
  - N-gram (1st pass), KN, 4-gram
  - Transformer (2nd pass)
    - Arch: Google Billion Words
- Data:
  - in-domain: Training corpus + Original LM
  - Generic: Common Crawl (CC)

Data/Vocab	in-dom. n-gram		in-dom. Transf.		CC 4-gram	
	Valid	Test	Valid	Test	Valid	Test
WSJ/162K	159	134	83	65	297	285
TL/200k	119	149	79	81	142	136
CV/168K	359	329	256	240	213	157
LS/200K	155/147	164/154	48/50	52/50	258/258	244/249
SB+FSH/64K	124	114/112	91	82/85	221	199/153
RV/200K	158	146	-	-	249	204

Unifying Audios

- Downsample all to 8kHz
- Similar FBank feature distribution (MVN per utterance)
- **Note**: Vanilla Up/Down sampling reduces the performance
  - SB: 8 → 16 kHz → ΔWER = + 1%abs
  - LS: 16 → 8 kHz → ΔWER = + 0.2%abs



Experimental Results (0)

Train	WSJ		TL		CV		LS				SB+FSH			average	
	nov93	nov92	valid	test	valid	test	dev-c	test-c	dev-o	test-o	RT03S	SB	CH	valid	test
SOTA		2.8	5.1	5.6				1.9		3.9	8.0	5.0	9.1		
WSJ	13.3	11.5	42.9	41.7	70.7	76.3	31.1	30.6	52.2	53.5	65.9	57.3	63.1	46.9	46.4
	8.1	6.4	28.4	28.9	54.5	61.7	16.4	16.7	36.8	38.7	52.3	44.2	49.7	34.0	34.3
	6.4	5.2	26.7	26.8	52.8	60.2	12.8	13.3	33.8	35.9	49.8	42.2	47.2	31.8	32.3
TL	12.9	10.7	7.4	7.5	30.8	34.7	9.7	9.8	20.0	20.4	28.3	20.0	28.4	18.9	18.4
	10.0	6.2	6.1	6.4	23.0	27.1	5.7	6.1	13.5	14.3	23.9	16.5	24.5	14.5	14.1
	6.9	5.4	5.8	6.0	22.0	26.1	4.0	4.5	10.1	11.7	23.3	16.6	24.8	13.0	13.3
CV	12.1	9.0	46.4	30.0	13.1	16.9	19.2	20.9	25.3	27.0	47.8	39.7	43.6	28.3	24.3
	6.7	4.1	38.2	23.4	10.8	13.8	14.3	16.1	18.3	20.1	37.1	29.9	34.2	21.8	18.3
	5.7	3.6	37.7	21.8	10.7	13.6	12.6	14.5	15.9	17.7	35.3	28.0	32.9	20.7	17.1
LS-960	13.6	11.0	12.7	13.4	30.0	34.1	2.8	2.8	7.1	7.1	36.4	27.1	33.8	19.5	18.8
	7.1	3.8	7.8	9.4	18.8	22.5	2.0	2.5	5.3	5.6	27.5	19.3	26.4	13.0	12.5
	4.9	3.6	7.3	8.6	18.1	22.0	1.5	2.1	4.3	4.7	25.9	18.3	25.3	11.8	11.9
SB+FSH	12.1	11.5	14.9	12.8	42.6	45.7	14.1	15.0	28.6	29.2	12.8	7.7	12.0	20.8	20.4
	6.4	5.2	8.5	8.8	31.7	36.0	7.1	7.9	19.1	20.4	10.4	6.5	10.3	14.0	14.5
	5.1	3.9	8.1	8.2	29.8	34.3	4.6	5.7	16.1	17.5	10.3	6.4	10.4	12.7	13.3
Joint	4.5	3.4	6.9	6.9	13.1	15.5	3.0	3.0	7.3	7.3	11.7	6.3	10.7	8.3	7.9
	3.1	2.0	5.4	5.7	10.5	12.6	2.0	2.5	5.2	5.6	9.8	5.9	9.5	6.5	6.4
	2.9	2.1	5.1	5.2	10.3	12.3	1.4	2.1	4.1	4.4	9.7	5.8	9.3	6.2	6.1
Joint CC	4.0	2.8	5.6	5.7	8.9	10.6	3.1	3.0	6.0	6.0	10.0	5.5	9.1	6.6	6.2

* Greedy decoding … No LM

* Beam-search decoding ... (first pass) in-domain n-gram LM

* Beam-search decoding … second pass rescoring by in-domain Transformer

* Joint CC → Joint, decoding with 4-gram

* Average of average (same weight for all datasets)

WSJ ...

Train	WSJ		TL		CV		LS				SB+FSH			average	
	nov93	nov92	valid	test	valid	test	dev-c	test-c	dev-o	test-o	RT03S	SB	CH	valid	test
SOTA		2.8	5.1	5.6				1.9		3.9	8.0	5.0	9.1		
WSJ	13.3	11.5	42.9	41.7	70.7	76.3	31.1	30.6	52.2	53.5	65.9	57.3	63.1	46.9	46.4
	8.1	6.4	28.4	28.9	54.5	61.7	16.4	16.7	36.8	38.7	52.3	44.2	49.7	34.0	34.3
	6.4	5.2	26.7	26.8	52.8	60.2	12.8	13.3	33.8	35.9	49.8	42.2	47.2	31.8	32.3
TL	12.9	10.7	7.4	7.5	30.8	34.7	9.7	9.8	20.0	20.4	28.3	20.0	28.4	18.9	18.4
	10.0	6.2	6.1	6.4	23.0	27.1	5.7	6.1	13.5	14.3	23.9	16.5	24.5	14.5	14.1
	6.9	5.4	5.8	6.0	22.0	26.1	4.0	4.5	10.1	11.7	23.3	16.6	24.8	13.0	13.3
CV	12.1	9.0	46.4	30.0	13.1	16.9	19.2	20.9	25.3	27.0	47.8	39.7	43.6	28.3	24.3
	6.7	4.1	38.2	23.4	10.8	13.8	14.3	16.1	18.3	20.1	37.1	29.9	34.2	21.8	18.3
	5.7	3.6	37.7	21.8	10.7	13.6	12.6	14.5	15.9	17.7	35.3	28.0	32.9	20.7	17.1
LS-960	13.6	11.0	12.7	13.4	30.0	34.1	2.8	2.8	7.1	7.1	36.4	27.1	33.8	19.5	18.8
	7.1	3.8	7.8	9.4	18.8	22.5	2.0	2.5	5.3	5.6	27.5	19.3	26.4	13.0	12.5
	4.9	3.6	7.3	8.6	18.1	22.0	1.5	2.1	4.3	4.7	25.9	18.3	25.3	11.8	11.9
SB+FSH	12.1	11.5	14.9	12.8	42.6	45.7	14.1	15.0	28.6	29.2	12.8	7.7	12.0	20.8	20.4
	6.4	5.2	8.5	8.8	31.7	36.0	7.1	7.9	19.1	20.4	10.4	6.5	10.3	14.0	14.5
	5.1	3.9	8.1	8.2	29.8	34.3	4.6	5.7	16.1	17.5	10.3	6.4	10.4	12.7	13.3
Joint	4.5	3.4	6.9	6.9	13.1	15.5	3.0	3.0	7.3	7.3	11.7	6.3	10.7	8.3	7.9
	3.1	2.0	5.4	5.7	10.5	12.6	2.0	2.5	5.2	5.6	9.8	5.9	9.5	6.5	6.4
	2.9	2.1	5.1	5.2	10.3	12.3	1.4	2.1	4.1	4.4	9.7	5.8	9.3	6.2	6.1
Joint CC	4.0	2.8	5.6	5.7	8.9	10.6	3.1	3.0	6.0	6.0	10.0	5.5	9.1	6.6	6.2

* Poor (the worst) ASR performance transfer from WSJ to others

→ Why? *Domain overfitting* …

… amount of data (81h), too clean, limited variability, etc.

TL, SB+FSH and LibriSpeech

Train	WSJ nov93	WSJ nov92	TL valid	TL test	CV valid	CV test	LS dev-c	LS test-c	LS dev-o	LS test-o	SB+FSH RT03S	SB+FSH SB	SB+FSH CH	average valid	average test
SOTA		2.8	5.1	5.6				1.9		3.9	8.0	5.0	9.1		
WSJ	13.3	11.5	42.9	41.7	70.7	76.3	31.1	30.6	52.2	53.5	65.9	57.3	63.1	46.9	46.4
	8.1	6.4	28.4	28.9	54.5	61.7	16.4	16.7	36.8	38.7	52.3	44.2	49.7	34.0	34.3
	6.4	5.2	26.7	26.8	52.8	60.2	12.8	13.3	33.8	35.9	49.8	42.2	47.2	31.8	32.3
TL	12.9	10.7	7.4	7.5	30.8	34.7	9.7	9.8	20.0	20.4	28.3	20.0	28.4	18.9	18.4
	10.0	6.2	6.1	6.4	23.0	27.1	5.7	6.1	13.5	14.3	23.9	16.5	24.5	14.5	14.1
	6.9	5.4	5.8	6.0	22.0	26.1	4.0	4.5	10.1	11.7	23.3	16.6	24.8	13.0	13.3
CV	12.1	9.0	46.4	30.0	13.1	16.9	19.2	20.9	25.3	27.0	47.8	39.7	43.6	28.3	24.3
	6.7	4.1	38.2	23.4	10.8	13.8	14.3	16.1	18.3	20.1	37.1	29.9	34.2	21.8	18.3
	5.7	3.6	37.7	21.8	10.7	13.6	12.6	14.5	15.9	17.7	35.3	28.0	32.9	20.7	17.1
LS-960	13.6	11.0	12.7	13.4	30.0	34.1	2.8	2.8	7.1	7.1	36.4	27.1	33.8	19.5	18.8
	7.1	3.8	7.8	9.4	18.8	22.5	2.0	2.5	5.3	5.6	27.5	19.3	26.4	13.0	12.5
	4.9	3.6	7.3	8.6	18.1	22.0	1.5	2.1	4.3	4.7	25.9	18.3	25.3	11.8	11.9
SB+FSH	12.1	11.5	14.9	12.8	42.6	45.7	14.1	15.0	28.6	29.2	12.8	7.7	12.0	20.8	20.4
	6.4	5.2	8.5	8.8	31.7	36.0	7.1	7.9	19.1	20.4	10.4	6.5	10.3	14.0	14.5
	5.1	3.9	8.1	8.2	29.8	34.3	4.6	5.7	16.1	17.5	10.3	6.4	10.4	12.7	13.3
Joint	4.5	3.4	6.9	6.9	13.1	15.5	3.0	3.0	7.3	7.3	11.7	6.3	10.7	8.3	7.9
	3.1	2.0	5.4	5.7	10.5	12.6	2.0	2.5	5.2	5.6	9.8	5.9	9.5	6.5	6.4
	2.9	2.1	5.1	5.2	10.3	12.3	1.4	2.1	4.1	4.4	9.7	5.8	9.3	6.2	6.1
Joint CC	4.0	2.8	5.6	5.7	8.9	10.6	3.1	3.0	6.0	6.0	10.0	5.5	9.1	6.6	6.2

* Average-wise (1): TL and SB+FSH … perform on par …
* Average-wise (2): LibriSpeech … single … best …
  → Why? Data amount (960h) + variability (clean + other)

Joint AM + CC (generic) LM

Train	WSJ		TL		CV		LS				SB+FSH			average	
	nov93	nov92	valid	test	valid	test	dev-c	test-c	dev-o	test-o	RT03S	SB	CH	valid	test
SOTA		2.8	5.1	5.6				1.9		3.9	8.0	5.0	9.1		
WSJ	13.3	11.5	42.9	41.7	70.7	76.3	31.1	30.6	52.2	53.5	65.9	57.3	63.1	46.9	46.4
	8.1	6.4	28.4	28.9	54.5	61.7	16.4	16.7	36.8	38.7	52.3	44.2	49.7	34.0	34.3
	6.4	5.2	26.7	26.8	52.8	60.2	12.8	13.3	33.8	35.9	49.8	42.2	47.2	31.8	32.3
TL	12.9	10.7	7.4	7.5	30.8	34.7	9.7	9.8	20.0	20.4	28.3	20.0	28.4	18.9	18.4
	10.0	6.2	6.1	6.4	23.0	27.1	5.7	6.1	13.5	14.3	23.9	16.5	24.5	14.5	14.1
	6.9	5.4	5.8	6.0	22.0	26.1	4.0	4.5	10.1	11.7	23.3	16.6	24.8	13.0	13.3
CV	12.1	9.0	46.4	30.0	13.1	16.9	19.2	20.9	25.3	27.0	47.8	39.7	43.6	28.3	24.3
	6.7	4.1	38.2	23.4	10.8	13.8	14.3	16.1	18.3	20.1	37.1	29.9	34.2	21.8	18.3
	5.7	3.6	37.7	21.8	10.7	13.6	12.6	14.5	15.9	17.7	35.3	28.0	32.9	20.7	17.1
LS-960	13.6	11.0	12.7	13.4	30.0	34.1	2.8	2.8	7.1	7.1	36.4	27.1	33.8	19.5	18.8
	7.1	3.8	7.8	9.4	18.8	22.5	2.0	2.5	5.3	5.6	27.5	19.3	26.4	13.0	12.5
	4.9	3.6	7.3	8.6	18.1	22.0	1.5	2.1	4.3	4.7	25.9	18.3	25.3	11.8	11.9
SB+FSH	12.1	11.5	14.9	12.8	42.6	45.7	14.1	15.0	28.6	29.2	12.8	7.7	12.0	20.8	20.4
	6.4	5.2	8.5	8.8	31.7	36.0	7.1	7.9	19.1	20.4	10.4	6.5	10.3	14.0	14.5
	5.1	3.9	8.1	8.2	29.8	34.3	4.6	5.7	16.1	17.5	10.3	6.4	10.4	12.7	13.3
Joint	4.5	3.4	6.9	6.9	13.1	15.5	3.0	3.0	7.3	7.3	11.7	6.3	10.7	8.3	7.9
	3.1	2.0	5.4	5.7	10.5	12.6	2.0	2.5	5.2	5.6	9.8	5.9	9.5	6.5	6.4
	2.9	2.1	5.1	5.2	10.3	12.3	1.4	2.1	4.1	4.4	9.7	5.8	9.3	6.2	6.1
Joint CC	4.0	2.8	5.6	5.7	8.9	10.6	3.1	3.0	6.0	6.0	10.0	5.5	9.1	6.6	6.2

* Joint acoustic model → better than single.mdl per dataset
* Joint AM + generic CC LM is ~ as good as Joint AM + in-domain LM

E. Loweimi

In-dom: 5-gram
CC: 4-gram

Research to Real-world Transfer

- **Baseline**:
  - Train/Dev/Test: RV [real-world data]
- **Single**
  - WSJ … poorest transfer
  - TL … best transfer
- **Joint**
  - Slightly worse than baseline
  - FT + 1h ~ on par w/ baseline
  - FT+ 10/100h → better than baseline

Train	LM	Valid	Test clean	Test noisy	Test extreme
RV	-	18.4	17.1	22.4	31.8
	in-dom.	12.8	15.7	20.9	29.8
WSJ	-	69.6	67.7	74.3	84.8
	in-dom.	56	54.9	62.4	71.8
TL	-	29.5	26	34.4	46.5
	in-dom.	22.1	21.4	29.4	40.6
CV	-	42.2	34.7	45.7	58
	in-dom.	31.6	27.3	37.7	49.4
LS-960	-	36.9	32.7	42.7	58.3
	in-dom.	24.4	24.6	33.5	45
SB+FSH	-	35.7	31.6	37.0	45.3
	in-dom.	28.6	26.6	32.5	41.0
Joint	-	23.6	19.2	25.5	35.0
	in-dom.	17.9	16.1	21.9	31.4
	CC	20.6	15.8	21.7	31.2
Joint + finetune RV-1h	-	22.5	18.4	23.6	34.3
	in-dom.	16.7	15.2	21.2	30.3
	CC	19.5	15.0	20.9	30.1
Joint + finetune RV-10h	-	20.8	17.1	23.4	33.0
	in-dom.	15.7	14.6	20.5	29.8
	CC	18.5	14.1	20.2	29.5
Joint + finetune RV-100h	-	18.9	15.5	21.2	31.4
	in-dom.	14.3	13.3	18.7	28.2
	CC	16.8	12.9	18.2	27.7

FT: Fine-tuning

Conclusion

- Are our models robust enough?
- Robustness … Generalisation … mismatch
- AM: Transformer + CTC
- LM: n-gram (1st pass) & Transformer (2nd pass); generic CC
- Generalisation from (single/joint) research to real-world
  - TD-LIUM, SwitchBoard transferable to real-world

The History of Speech Recognition to the Year 2030

Awni Hannun*

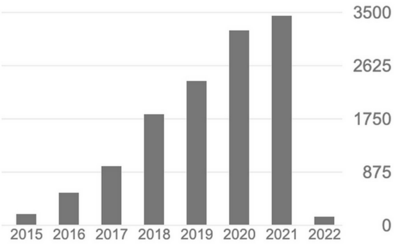
awni.hannun@gmail.com

* Distinguished Scientist at Zoom
* Facebook AI Research (FAIR)
* Baidu's Silicon Valley AI Lab
* PhD in Stanford University, advised by Andrew Ng

Take a look at his blog

Cited by

	All	Since 2017
Citations	12856	11990
h-index	26	24
i10-index	31	31



16, January, 2022

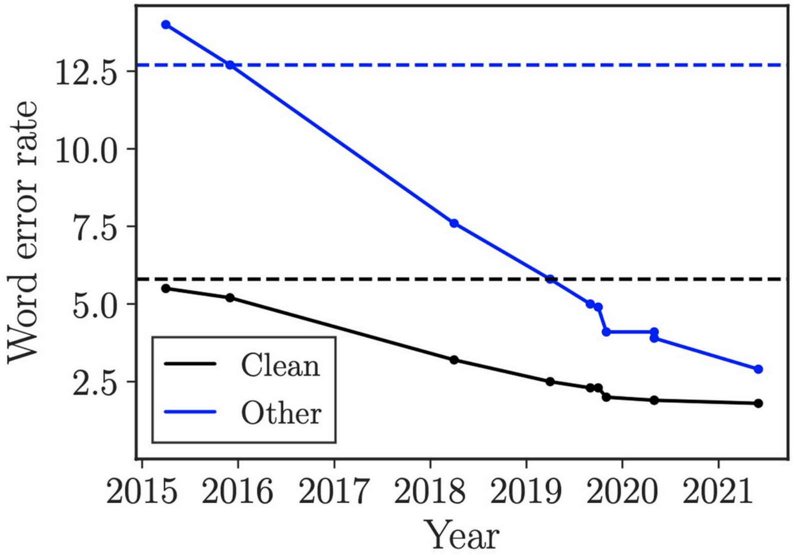
2010-2020
Remarkable Improvement in ASR

- 2010 – 2020 → dramatic ASR improvement

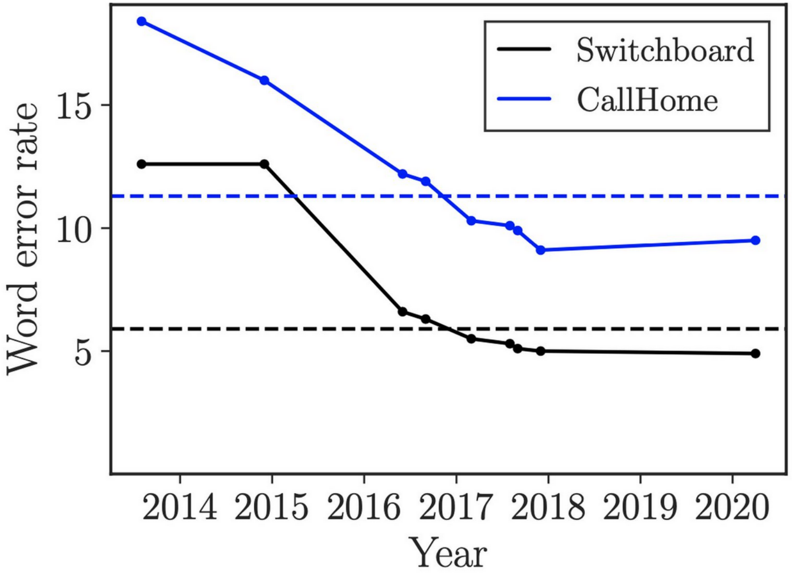
- What we can expect over the coming decade?
  - What is left?

YouTube

ASR vs HSR over time



(a) LibriSpeech

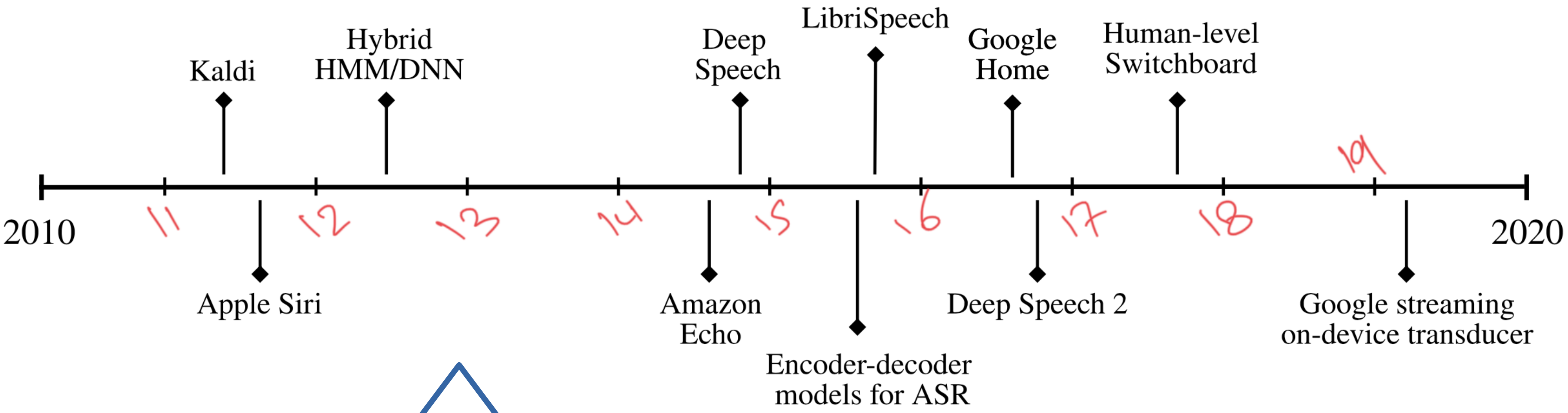


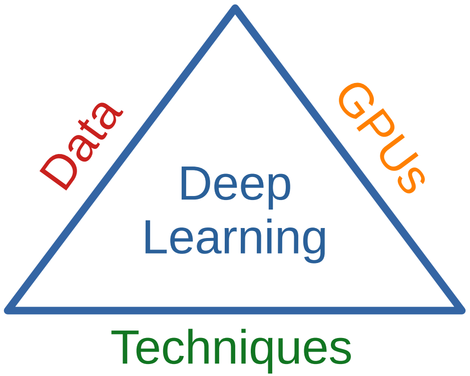
(b) Switchboard Hub5'00

* Dash lines: Human-level performance (professional transcriber)

* What is left if ASR is better than HSR?

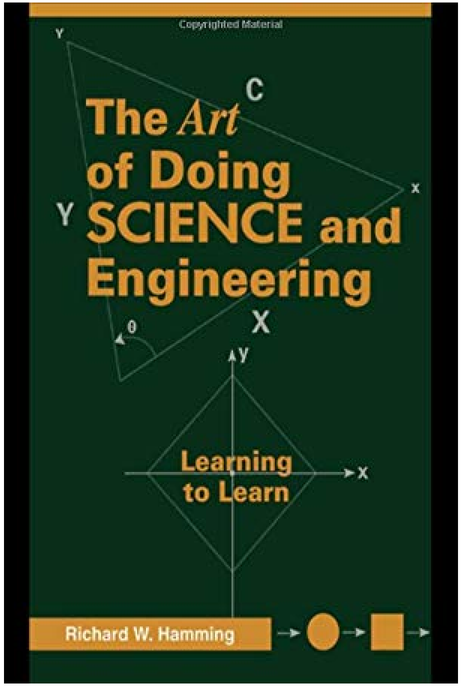
Timeline of major developments in ASR





Richard Hamming

Publish in 1997



Video Lectures in YouTube
(1995)

The history of Computing to the Year 2000

R. Hamming 1960

1968 ACM Turing Lecture

One Man's View of Computer Science

R. W. HAMMING

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey

The Unreasonable Effectiveness of Mathematics

R. W. Hamming

The American Mathematical Monthly, Vol. 87, No. 2. (Feb., 1980), pp. 81-90

Richard Hamming
(1915 – 1998)

Biography

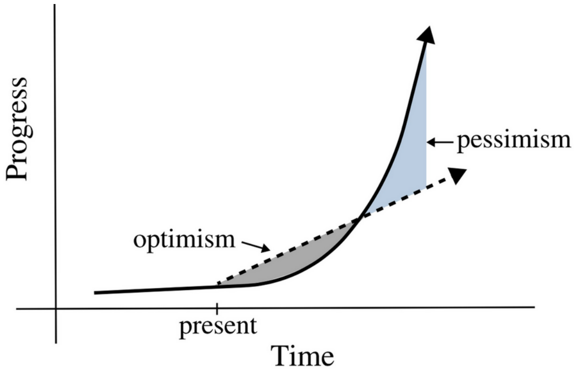
Quotes

Hamming’s Predictions

- *… by 2020 it would be fairly universal practice for the expert in the filed of application to do the actual program preparation rather than have experts in computers do the program preparation.*
- *NN … represent a solution to the programming problem … will probably play a large part in the future of computing*
- Pre-valence of
  - general-purpose rather than special-purpose hardware
  - digital over analog
  - high-level programming languages
  - fiber optic rather than copper wires
  - ...

Hamming was very good in predicting the future … How

- Technology forecasting is challenging …
- **Practice** → Friday afternoons … "great thoughts" … mused on the future
- **Mastering the fundamentals** → depth and breadth
- **Open-minded**

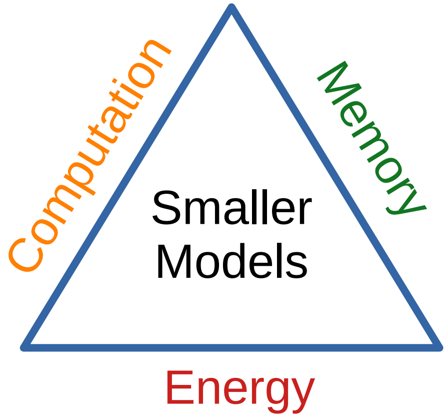


Research Prediction (1)

- Semi-supervised and self-supervised Learning are here to stay
- Goal: Leveraging unannotated data
- Approaches: Pseudo-labelling, CPC, etc.
- Challenges: scale (accessibility) + details
  - … Shift from research labs to engineering organizations
- Research implications:
  - Lighter-weight models, optimisation (faster training), incorporation of prior knowledge (for sample efficiency)

Research Prediction (2)

- ASR on/at the device/edge
- Why edge processing is important?
  - Data privacy … training+inference on device
  - Lower Latency + 100% availability [w/o internet]
- Research implications:
  - Sparsity [lottery ticket hypothesis, etc.]
  - Knowledge Distillation [directly]
  - Quantization



Research Prediction (3)

- ~~Improved WER on benchmark X with mdl/arch Y~~
  - Saturated on academic benchmark
    - Scale will solve new challenging tasks, too!
  - Practical value of low WER (*correlation*)
    - Low $\text{WER}_{\text{academic}} \overset{?}{\rightarrow}$ Low $\text{WER}_{\text{real-world}}$
    - Other quality metrics $\leftrightarrow$ human understanding
      - e.g., semantic error rate

Research Prediction (4) & (5)

- Transcription replaced with richer representations for downstream tasks, e.g. lattice/graph
- Personalisation to individual users
  - Leveraging context (topic, history, background, visual cues, facial expressions, etc.)
  - Narrow down the scope … underrepresented in training data
- On-device personalisation ... on-devices trainable/customisable … user/context

Application Prediction (1) & (2)

- 99% of transcription with ASR
- Voice assistants get better (incrementally, not fundamentally)
  - ASR is no longer a bottleneck
  - New bottlenecks: language understanding
    - How to maintain a conversation, etc.
- What is left?
  - A lot left to build ASR that works all the time, for everyone!

Summary

Table 1: Predictions for the progress in speech recognition research and applications by the year 2030.

Prediction
Self-supervised learning and pretrained models are here to stay.
Most speech recognition (inference) will happen at the edge.
On-device model training will be much more common.
Sparsity will be a key research direction to enable on-device inference and training.
Improving word error rate on common benchmarks will fizzle out as a research goal.
Speech recognizers will output richer representations (graphs) for use by downstream tasks.
Personalized models will be commonplace.
Most transcription services will be automated.
Voice assistants will continue to improve, but incrementally.

That’s It!

- Thanks for your attention!
- Q/A