Human Parity and Beyond

Jasha Droppo

Introduction

- The history of Automatic Speech Recognition (ASR) is one of solving progressively harder tasks over time, meeting or exceeding human performance.
- Collectively, we have recently solved the task of transcribing American English conversational telephone speech (CTS).
- This talk covers
  - Human parity in American English CTS.
  - An analysis of human and machine errors on this task.
  - What lies beyond human parity?

Acknowledgments

- Xuedong Huang
- Fil Alleva
- Zhehuai Chen
- Frank Seide
- Mike Seltzer
- Andreas Stolcke
- Lingfeng Wu
- Wayne Xiong
- Dong Yu
- Geoff Zweig

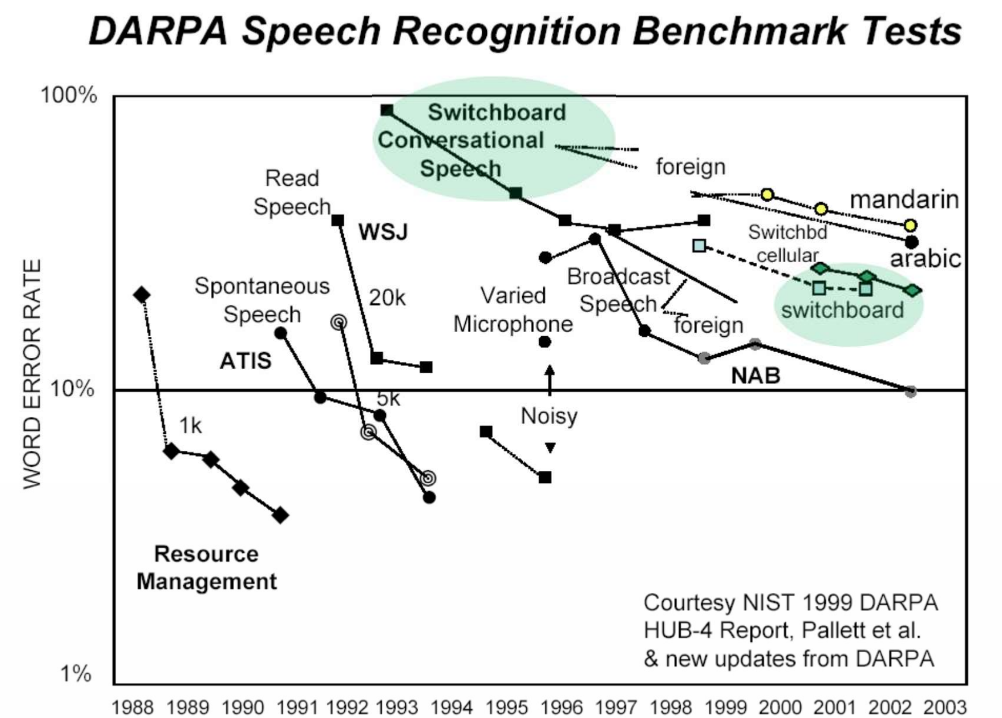
Introduction:
Task and History

The Human Parity Experiment

- Conversational telephone speech has been a benchmark in the research community for 20 years
  - Focus: strangers talking to each other via telephone, given a topic
  - Known as the “Switchboard” task in speech community
- Question: Can we achieve human-level performance?
- Top-level tasks:
  - Measure human performance
  - Build the best possible recognition system
  - Compare and analyze

30 Years of Speech Recognition Benchmarks

For many years, DARPA drove the field by defining public benchmark tasks



Read and planned speech:

RM

ATIS

WSJ

Conversational Telephone Speech (CTS):

Switchboard (SWB)
(strangers, on-topic)

CallHome (CH)
(friends & family, unconstrained)

History of Human Error Estimates for SWB

- Lippman (1997): 4%
  - based on "personal communication" with NIST, no experimental data cited
- LDC LREC paper (2010): 4.1-4.5%
  - Measured on a different dataset (but similar to our NIST evaluation set, SWB portion)
- Microsoft (2016): 5.9%
  - Transcribers were blind to experiment
  - 2-pass transcription, isolated utterances (no "transcriber adaptation")
- IBM (2017): 5.1%
  - Using multiple independent transcriptions, picked best transcriber
  - Vendor was involved in experiment and aware of NIST transcription conventions

Note: Human error will vary depending on

- Level of effort (e.g., multiple transcribers)
- Amount of context supplied (listening to short snippets vs. entire conversation)

Recent ASR Results on Switchboard

Group	2000 SWB WER	Notes	Reference
Microsoft	16.1%	DNN applied to LVCSR for the first time	Seide et al, 2011
Microsoft	9.9%	LSTM applied for the first time	A.-R. Mohammed et al, IEEE ASRU 2015
IBM	6.6%	Neural Networks and System Combination	Saon et al., Interspeech 2016
Microsoft	5.8%	First claim of "human parity"	Xiong et al., arXiv 2016, IEEE Trans. SALP 2017
IBM	5.5%	Revised view of "human parity"	Saon et al., Interspeech 2017
Capio	5.3%		Han et al., Interspeech 2017
Microsoft	5.1%	Current Microsoft research system	Xiong et al., MSR-TR-2017-39, ICASSP 2018

Microsoft System Overview and Results

System Overview

- Hybrid HMM/deep neural net architecture
- Multiple acoustic model types
  - Diverse architectures (convolutional and recurrent)
    - VGG, LACE, CNN, BLSTM, Resnet
  - Diverse senone sets
    - Different set size, different base phones
- Multiple language models
  - All based on LSTM recurrent networks
  - Different input encodings
  - Forward and backward running
- Advanced system combination
  - Model combination at multiple levels
  - Search for complementary acoustic model
  - Confusion-network based, weighted combination

Data used

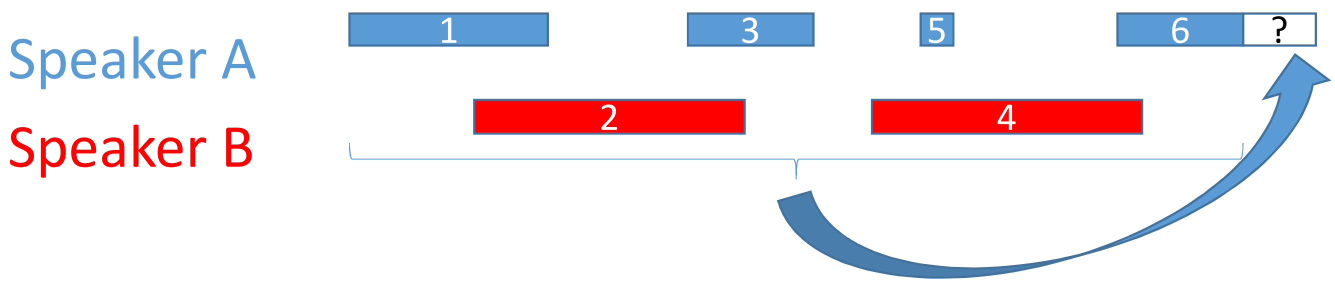
- Acoustic training: 2000 hours of conversational telephone data
- Language model training:
  - Conversational telephone transcripts
  - Web data collected to be conversational in style
  - Broadcast news transcripts
- Test on NIST 2000 SWB+CH evaluation set
- *Note:* data chosen to be compatible with past practice
  - NOT using proprietary sources

Language Modeling: Multiple LSTM variants

- Decoder uses a word 4-gram model
- N-best hypotheses are rescored with multiple LSTM recurrent network language models
- LSTMs differ by
  - Direction: forward/backward running
  - Encoding: word one-hot, word letter trigram, character one-hot
  - Scope: utterance-level / **session-level**

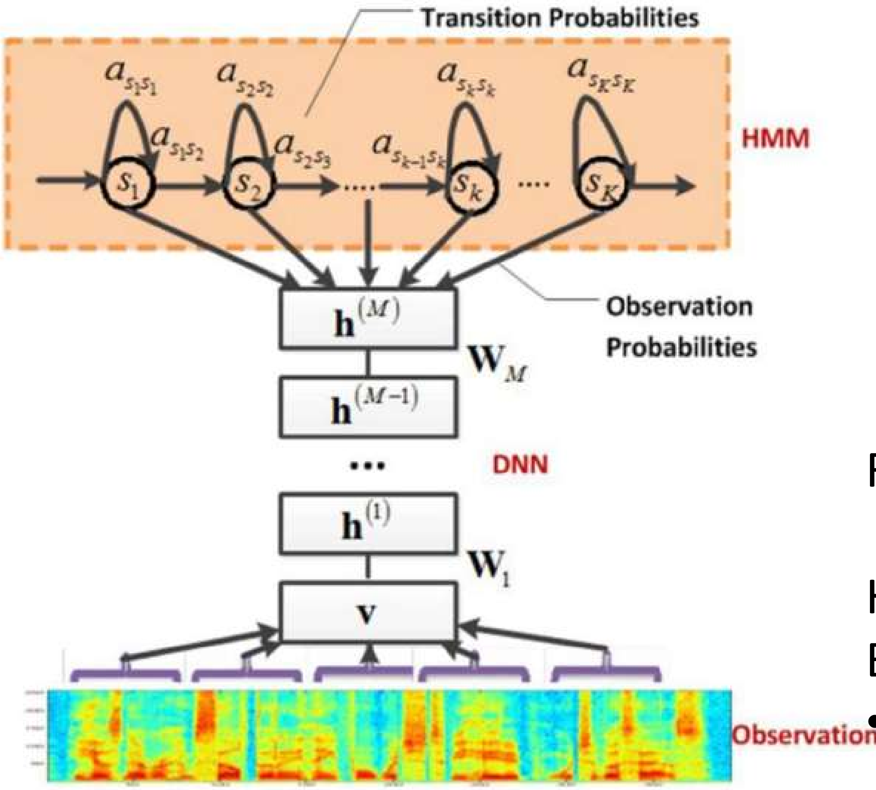
Session-level Language Modeling

- Predict next word from full conversation history, not just one utterance:



LSTM language model	Perplexity
Utterance-level LSTM (standard)	44.6
+ session word history	37.0
+ speaker change history	35.5
+ speaker overlap history	35.0

AM Framework: Hybrid HMM/DNN



[Yu et al., 2010; Dahl et al., 2011]

	CallHome	Switchboard
DNN	21.9%	13.4%

1st pass decoding

Record performance in 2011 [Seide et al.]

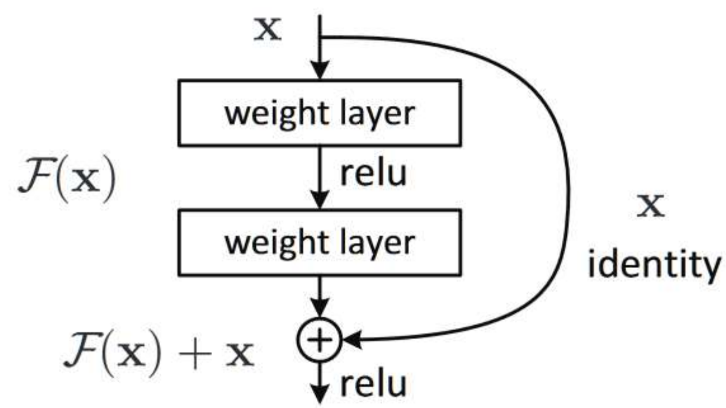
Hybrid HMM/NN approach still standard
But DNN model now obsolete (!)
- Poor spatial/temporal invariance

Acoustic Modeling: ResNet

Add a non-linear offset to linear transformation of features
Similar to fMPE in Povey et al., 2005
See also Ghahremani & Droppo, 2016

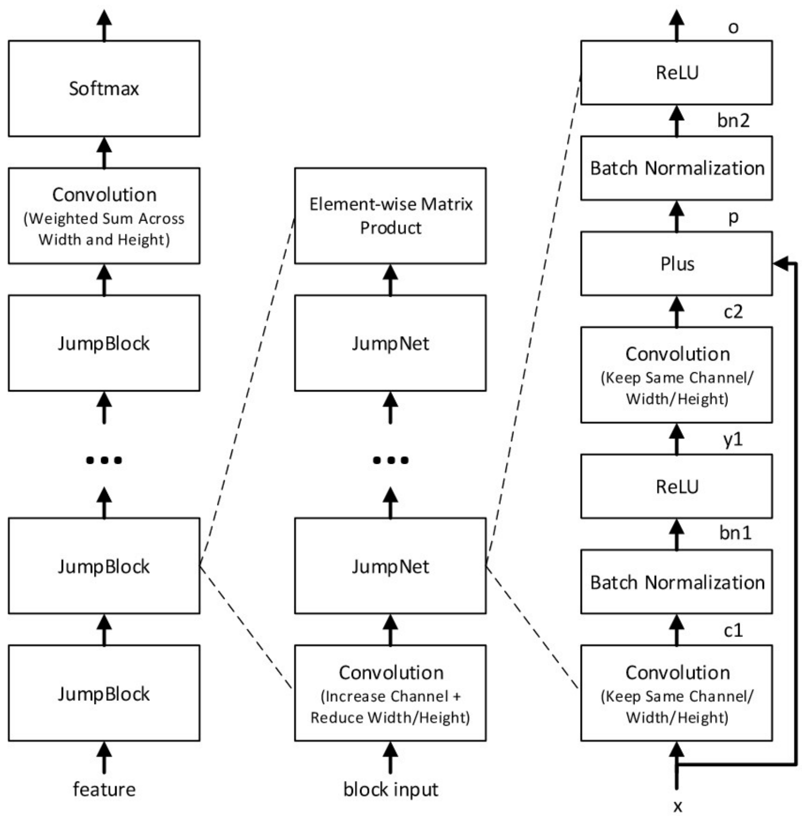
	CallHome	Switchboard
DNN	21.9%	13.4%
ResNet	17.3%	11.1%

1st pass decoding



[He et al., 2015]

Acoustic Modeling: LACE CNN

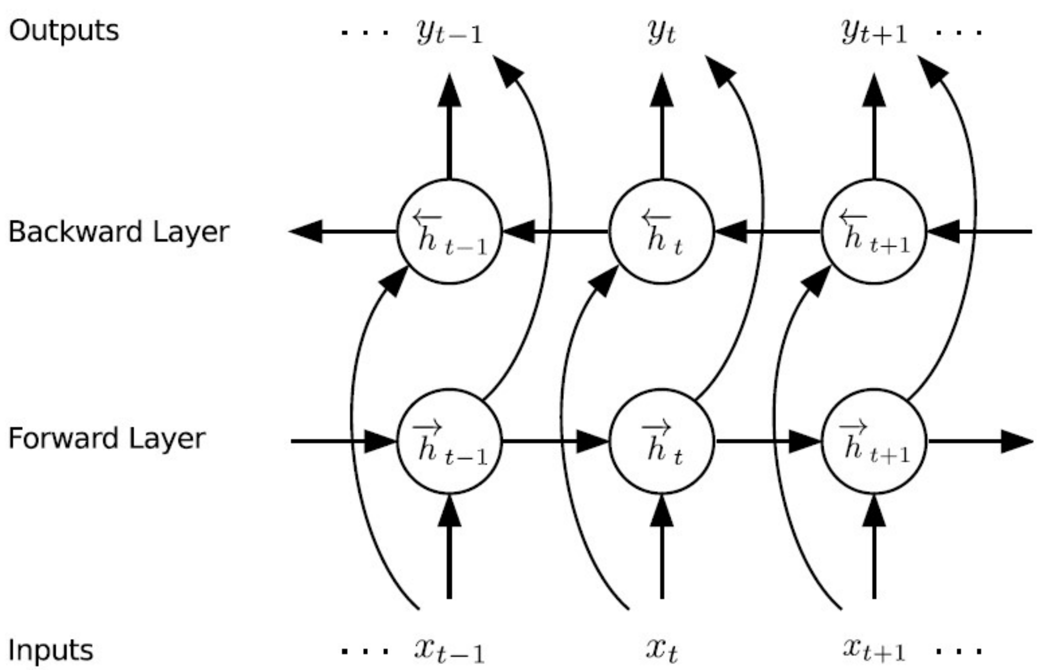


	CallHome	Switchboard
DNN	21.9%	13.4%
ResNet	17.3%	11.1%
LACE	16.9%	10.4%

1st pass decoding

CNNs with **batch normalization**, **Resnet jumps**, and **attention masks** [Yu et al., 2016]

Acoustic Modeling: Bidirectional LSTMs



[Graves & Jaitly '14]

	CallHome	Switchboard
DNN	21.9%	13.4%
ResNet	17.3%	11.1%
LACE	16.9%	10.4%
BLSTM	17.3%	10.3%

Stable form of recurrent neural net
Robust to temporal shifts

[Hochreiter & Schmidhuber, 1997,
Graves & Schmidhuber, 2005; Sak et al., 2014]

Acoustic Modeling: CNN-BLSTM

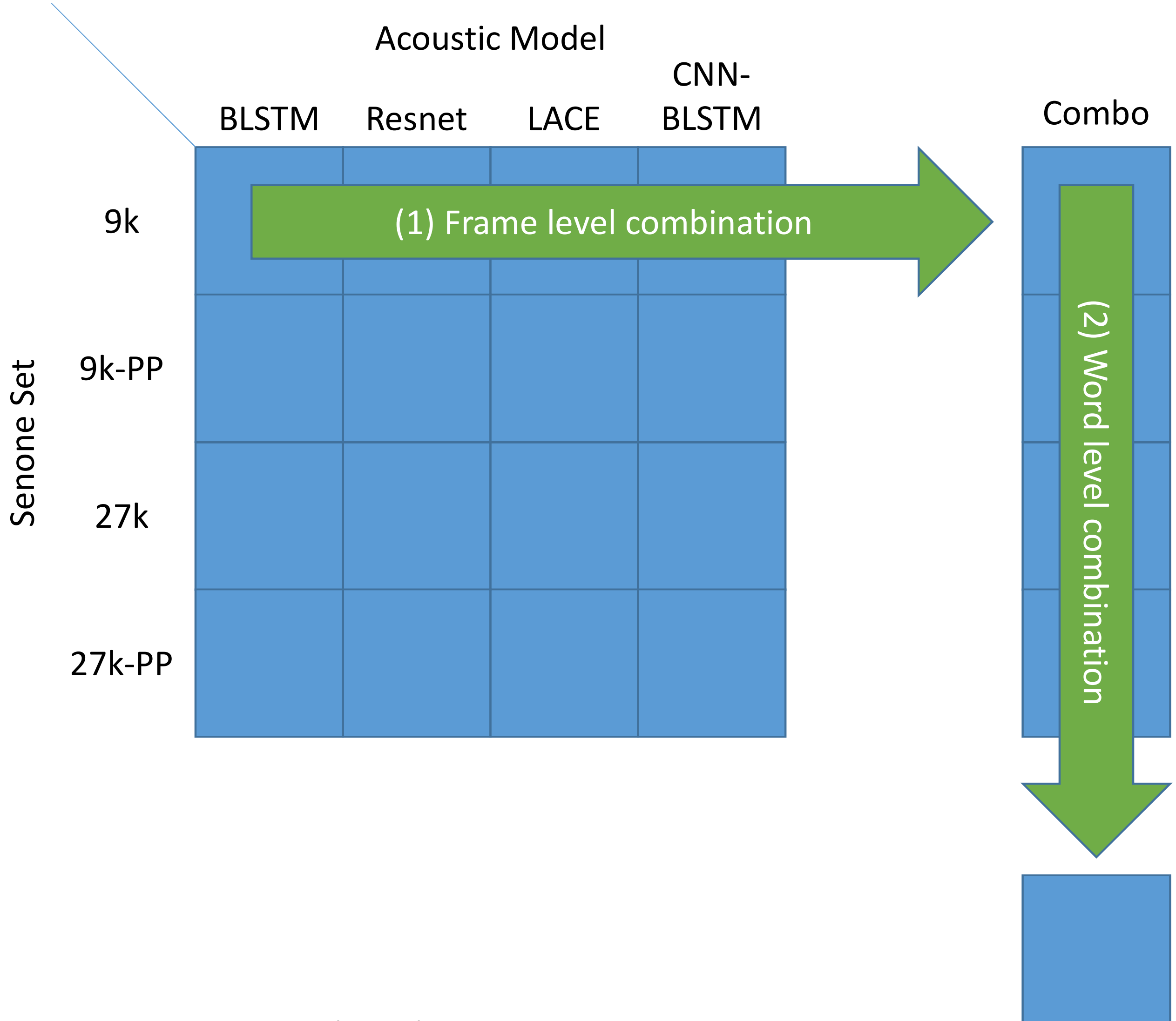
- Combination of convolutional and recurrent net model
  [Sainath et al., 2015]
- Three convolutional layers
- Six BLSTM recurrent layers

Acoustic model combination

Step 0: create 4 different versions of each acoustic model by clustering phonetic model units (**senones**) differently

Step 1: combine **different models** for **same senone** set at the **frame level** (posterior probability averaging)

Step 2: after LM rescoring, combine **different senone** systems at the **word level** (confusion network combination)



Results

Word error rates (WER)

Senone set	Acoustic models	SWB WER	CH WER
1	BLSTM	6.4	12.1
2	BLSTM	6.3	12.1
3	BLSTM	6.3	12.0
4	BLSTM	6.3	12.8
1	BLSTM + Resnet + LACE + CNN-BLSTM	5.4	10.2
2	BLSTM + Resnet + LACE + CNN-BLSTM	5.4	10.2
3	BLSTM + Resnet + LACE + CNN-BLSTM	5.6	10.2
4	BLSTM + Resnet + LACE + CNN-BLSTM	5.5	10.3
1+2+3+4	BLSTM + Resnet + LACE + CNN-BLSTM	5.2	9.8
	+ Confusion network rescoring	**5.1**	**9.8**

Frame-level combination

Word-level combination

Human vs. Machine

Human Performance on Switchboard

- The goal of reaching “human parity” in automatic CTS transcription raises the question of what should be considered human accuracy on this task.

Microsoft Human Error Estimate (2015)

- Skype Translator has a weekly transcription contract
  - For quality control, training, etc.
- Initial transcription followed by a second checking pass
  - Two transcribers on each speech excerpt
- One week, we added NIST 2000 CTS evaluation data to the pipeline
  - Speech was pre-segmented as in NIST evaluation

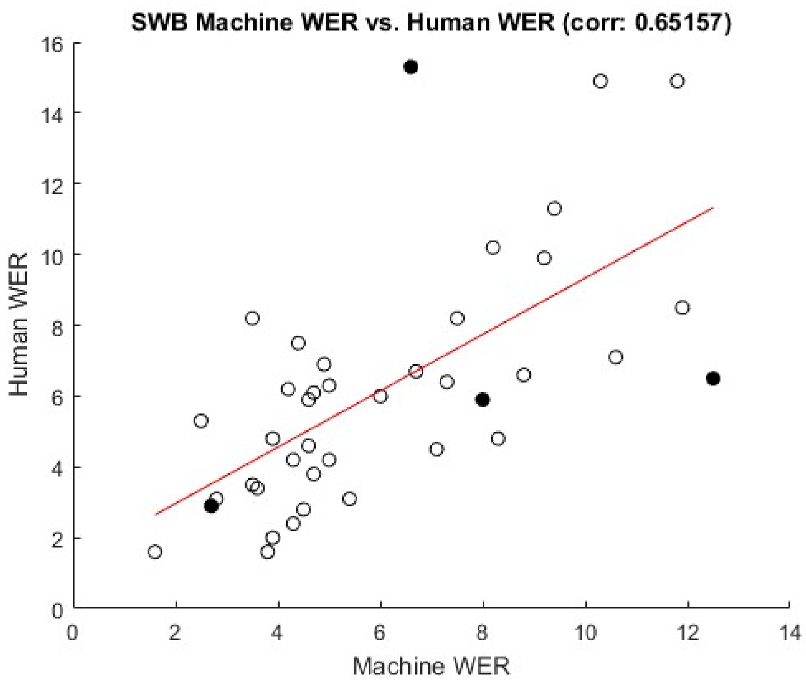
Human Error Estimate: Results

- Applied NIST scoring protocol (same as ASR)
- Switchboard: **5.9%** error rate
- CallHome: **11.3%** error rate
- SWB in the 4.1% - 9.6% range expected based on NIST study
- CH is *difficult for both people and machines*
  - Machine error about 2x higher
  - High ASR error not just because of mismatched conditions

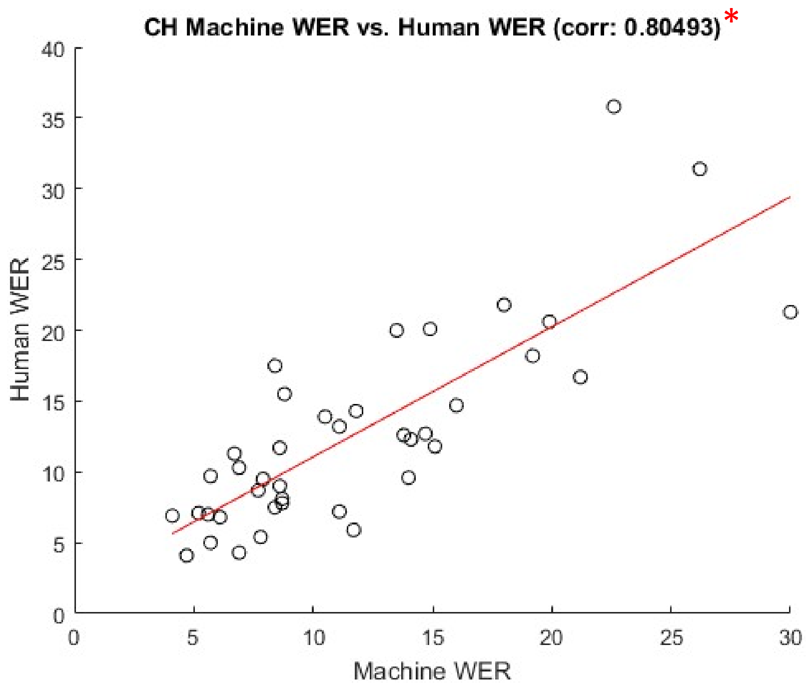
New questions:

- Are human and machine errors correlated?
- Do they make the same type of errors?
- Can humans tell the difference?

Correlation between human and machine errors?



$\rho = 0.65$

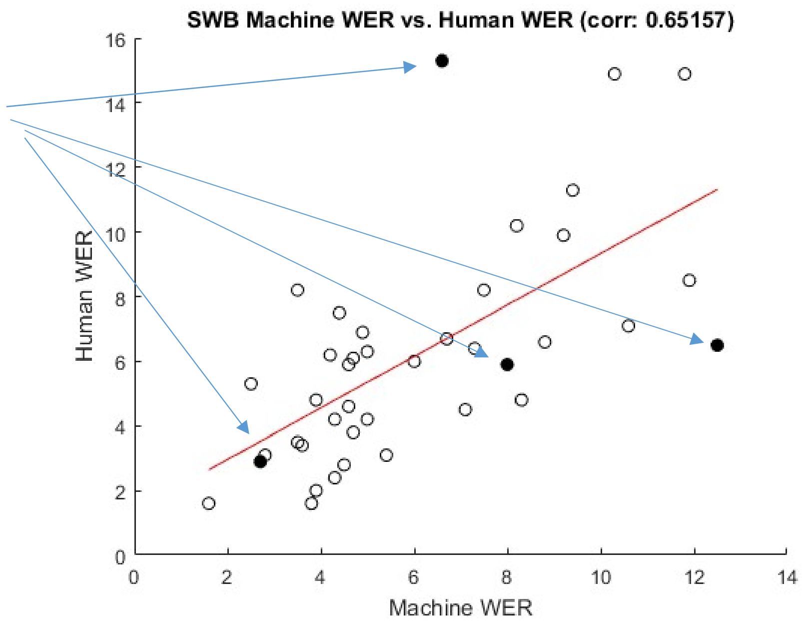


$\rho = 0.80$

*Two CallHome conversations with multiple speakers per conversation side removed, see paper for full results

Does the machine benefit from seeing test speakers in its training data?

- It has been suggested that the 2000 Switchboard test set is so “easy” because most of the speakers also occur in the training set (a corpora shortcoming)
- The filled dots are the *unseen* speakers
- This doesn’t seem to be the case:
  - Machine WER on unseen speakers is within the normal range
  - For the most part (3 of 4), machine WER predicts the human WER



Humans and machines: different error types?

Top word substitution errors (≈ 21k words in each test set)

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

Overall similar patterns: short function words get confused (also: inserted/deleted)

One outlier: machine falsely recognizes backchannel "uh-huh" for filled pause "uh"

- These words are acoustically confusable, have opposite pragmatic functions in conversation
- Humans can disambiguate by prosody and context

Top Insertion and Deletion Errors

Deletions

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

Insertions

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

Both humans and machines insert “I” and “and” a lot.
Short function words dominate the list for both.

Can humans tell the difference?

- Attendees at a major speech conference played "Spot the Bot"
- Showed them human and machine output side-by-side in random order, along with reference transcript
- Turing-like experiment: tell which transcript is human/machine
- Result: it was hard to beat a random guess
  - 53% accuracy (188/353 correct)
  - Not statistically different from chance ($p \approx 0.12$, one-tailed)

Conclusions

- Human transcription performance is around 5-6%, but also varies greatly with the function of the amount of effort!
  - Multiple independent transcription passes with reconciliation would lower this further, as done by NIST for their reference transcriptions
- State-of-the-art ASR technology based on neural net acoustic and language models has reached human-level accuracy <u>on this task</u>
- Human and machine transcription performance is highly correlated
  - “Hard” versus “easy” speakers
  - Word types involved in most frequent errors
  - Humans are better at recognizing pragmatically relevant words (“uh” vs. “uh-huh”)

Outlook

- Speech recognition is not solved!
- Need to work on
  - Robustness to acoustic environment (e.g., far-field mics, overlap)
  - Speaker mismatch (e.g., accented speech)
  - Style mismatch (e.g., planned vs. spontaneous, single vs. multiple speakers)
- Computational challenges
  - Inference too expensive for mobile devices
  - Static graph limits what can be expressed → Dynamic networks

The Future:
More Challenging Environments

- A Challenging Task
  - Unsupervised Single-channel Overlapped Speech Recognition
  - Permutation Invariant Training (baseline)
- Methods
  - Modular Initialization
  - Transfer Learning Based Joint Training
  - Temporal Correlation Modeling
  - Multi-output Sequence Discriminative Training
- Experiments

Overlapped ASR

- Received speech is linear combination of multiple independent speech signals.

$$O_u^{(m)} = \sum_{n=1}^{N} O_{un}^{(r)}$$

- Recognition task is to produce posterior over several label sequences.

$$P(L_{u1}, \ldots, L_{uN})$$

Overlapped ASR

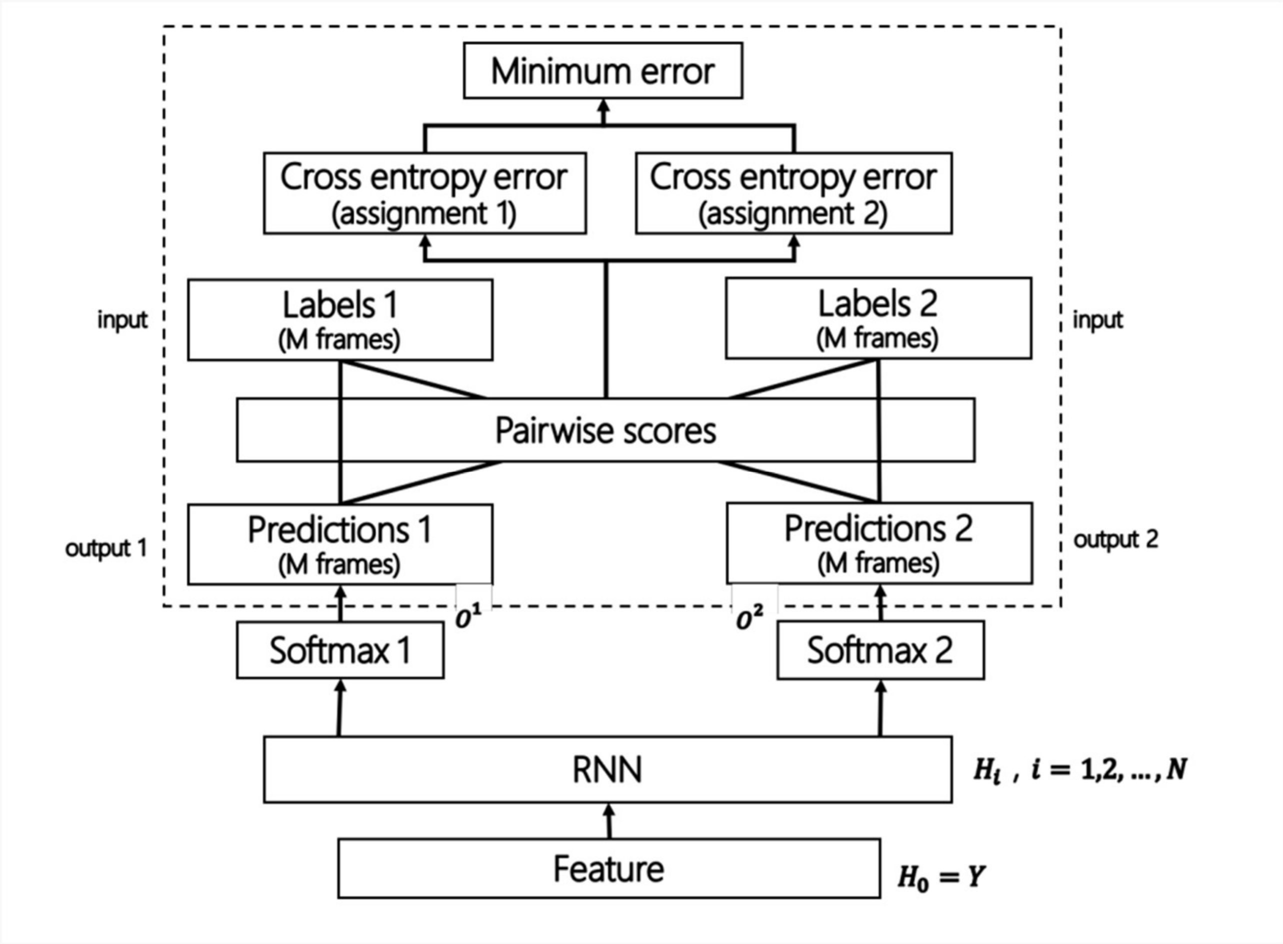
- Possible solutions:

$$P(L_{u1}, \dots, L_{uN}) \approx \prod_{n=1}^{N} P\left(L_{un} \middle| O_u^m\right) \approx \prod_{n}^{N} P\left(L_{un} \middle| O_{un}^{\hat{r}}\right)$$

- Speech Separation followed by Speech-to-text
  - Computational Auditory Scene Analysis (CASA)
  - Deep Clustering (DPCL)
  - Permutation Invariant Training for Speech Separation (PIT-SS or PIT-MSE)

- Joint Modeling
  - Permutation Invariant Training for ASR (PIT-ASR)

Permutation Invariant Training for ASR

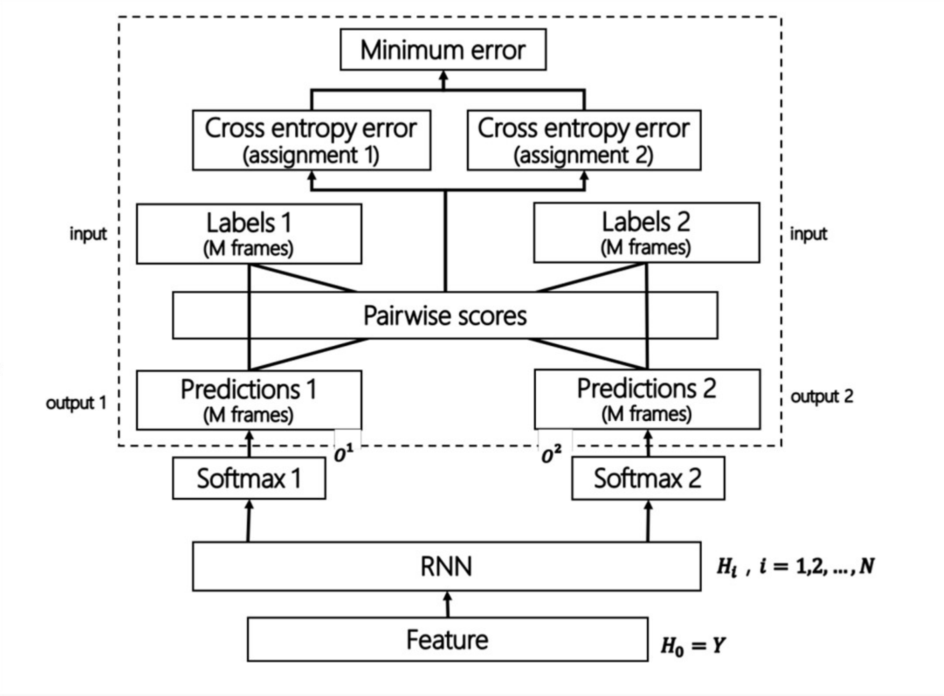


Permutation Invariant Training for ASR

$$P(\mathbf{L}_{u1}, ..., \mathbf{L}_{uN} | \mathbf{O}_u^{(m)}) \approx \prod_{n=1}^{N} P(\mathbf{L}_{un}^{(r)} | \mathbf{O}_u^{(m)}) \qquad (2)$$

$$\mathcal{J}_{\text{CE-PIT}} = \sum_u \min_{s' \in \mathbf{S}} \sum_t \frac{1}{N} \sum_{n \in [1,N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)}) \qquad (4)$$

- Disadvantages
  - Model solves three hard problems in one step
    - Separation, tracing, and recognition.
  - Frame CE applied to solve sequential problem.
  - Doesn't incorporate linguistic information.
- Result
  - WER more than 50%



- Methods
  - **Modular Initialization** 4-10%
  - Transfer Learning Based Joint Training 20%
  - Temporal Correlation Modeling 8%
  - Multi-outputs Sequence Discriminative Training 8%

Modular Initialization

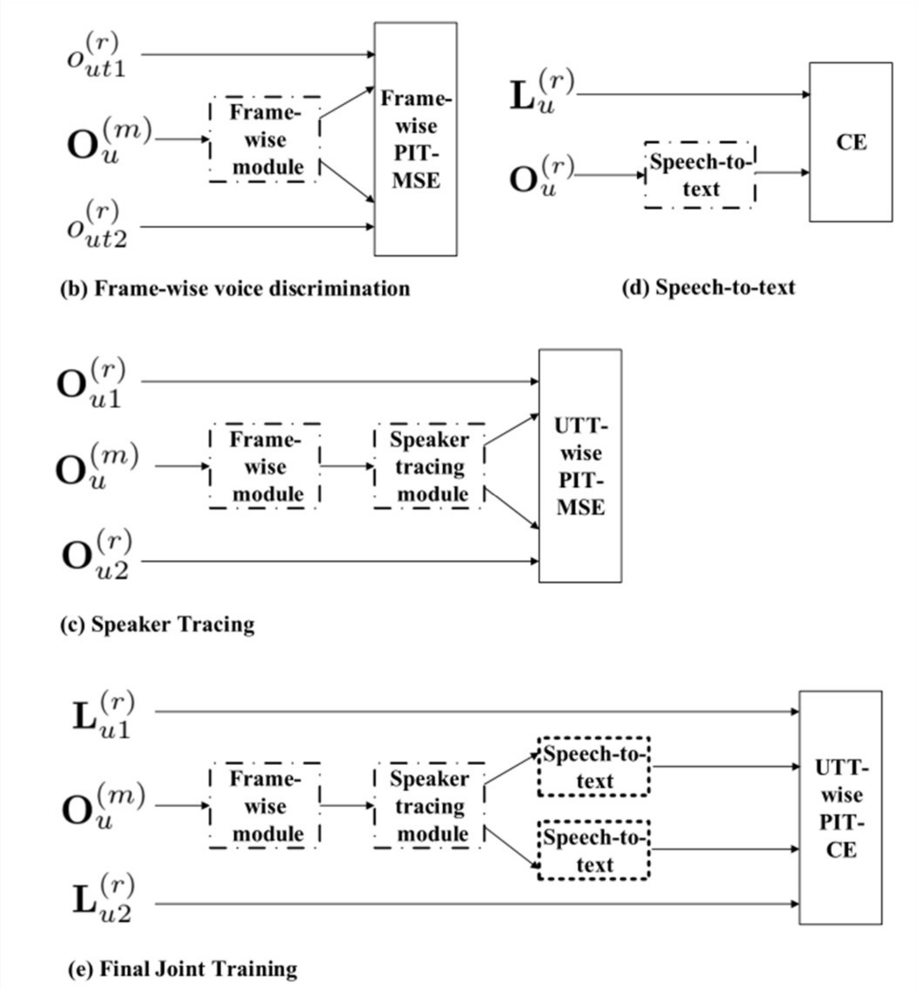
- Frame-wise **interpreting** (swapped segments)
  - Local feature extraction → CNN
- Speaker **Tracing** (no swap)
  - Temporal modeling → RNN
- Speech-to-text

$$\mathcal{J}_{\text{F-PIT}} = \sum_u \sum_t \frac{1}{N} \min_{s' \in \mathbf{S}} \sum_{n \in [1,N]} MSE(o_{utn}^{(s')}, o_{utn}^{(r)}) \quad (5)$$

$$\mathcal{J}_{\text{U-PIT}} = \sum_u \min_{s' \in \mathbf{S}} \sum_t \frac{1}{N} \sum_{n \in [1,N]} MSE(o_{utn}^{(s')}, o_{utn}^{(r)}) \quad (6)$$

Modular Initialization

- Progressive joint training
  - Curriculum learning theory
  - The harder task, the larger NN (stacking)
- Less Model Complexity
  - Speed of convergence
  - Better local minima
- Data Efficiency
- Combine with other tech.
  - Sequence disc. training on speech-to-text
  - Integrate LM



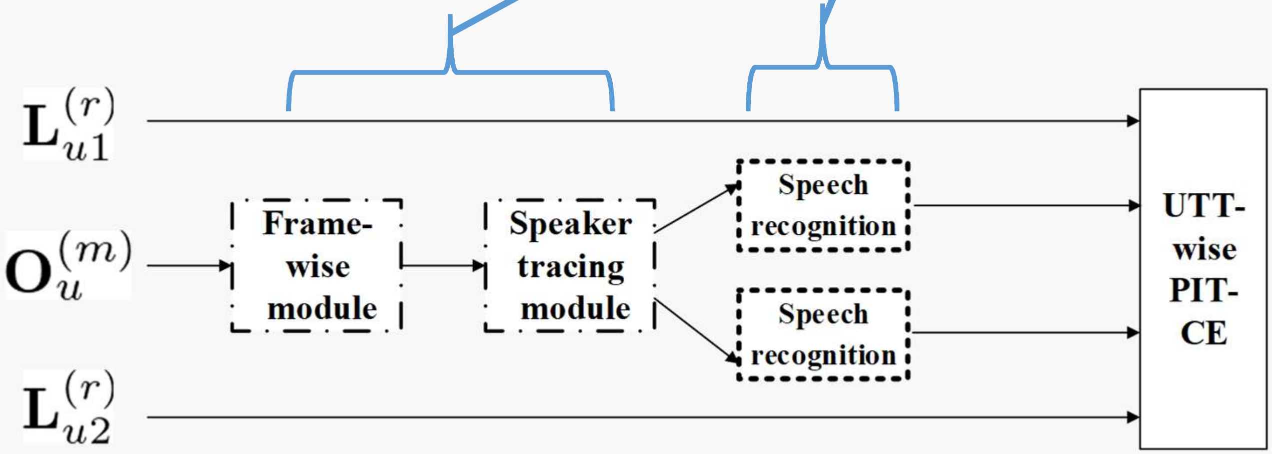
Experiments

- Data
  - Artificially overlapped Switchboard
    - 300 hours source material creates 150 hours of overlapped speech
    - The hub5e-swb test set maps from 1831 to 915 utterances
- Models
  - All speech recognition models have 9000 dimensional senone posterior output
  - Baseline 1: 10 layer, 768 cells BLSTM PIT-ASR model
  - Baseline 2: 6 layer, 768 cells BLSTM PIT-SS model + 4 layer 768 cells BLSTM ASR model

Experiments - Modularization

- Better model generalization

Layers	Modular	Fine-tune ST	Fine-tune ASR	WER	Rel. (%)
$10 \cdot 0$	×	×	×	57.5	0
$\boxed{6} \cdot \boxed{4}$	×	×	×	52.8	-8.2
	√	×	×	93.4	+62.4
	√	√	×	51.3	-10.7
	√	√	√	50.2	-12.7



Experiments - Modularization

- Better model generalization

Layers	Modular	Fine-tune ST	Fine-tune ASR	WER	Rel. (%)
10 · 0	×	×	×	57.5	0
6 · 4	×	×	×	52.8	-8.2
	√	×	×	93.4	+62.4
	√	√	×	51.3	-10.7
	√	√	√	50.2	-12.7

Better structure for ASR

Progressive joint training

- Methods
  - Modular Initialization 4-10%
  - **Transfer Learning Based Joint Training** 20%
  - Temporal Correlation Modeling 8%
  - Multi-outputs Sequence Discriminative Training 8%

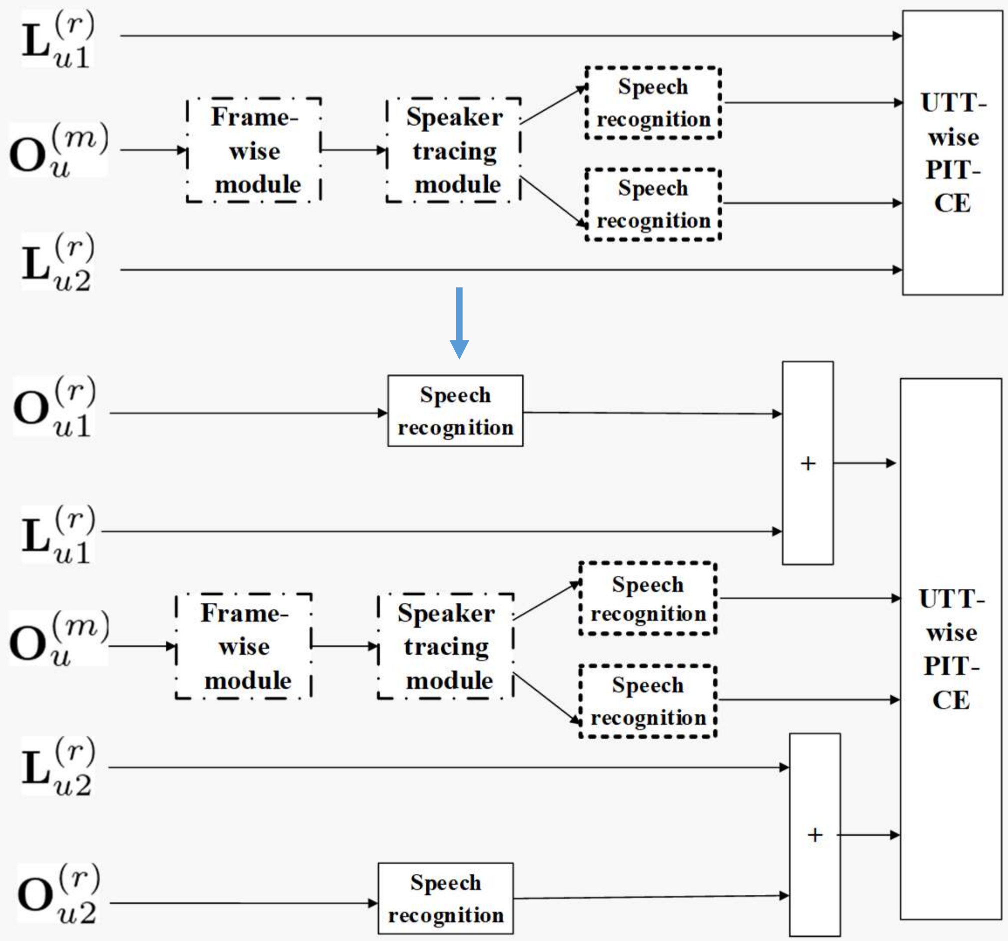
Transfer Learning based Joint Training

$$\mathcal{J}_{\text{CE-PIT}} = \sum_u \min_{s' \in \mathbf{S}} \sum_t \frac{1}{N} \sum_{n \in [1,N]} \boxed{CE}(l_{utn}^{(s')}, l_{utn}^{(r)}) \quad (4)$$

$$\mathcal{J}_{\text{KLD-PIT}} = \sum_u \min_{s' \in \mathbf{S}} \sum_t \frac{1}{N} \sum_{n \in [1,N]} \boxed{KLD}(P(l_{utn}^{(c)} | \mathbf{O}_{un}^{(r)}), P(l_{utn}^{(s')} | \mathbf{O}_u^{(m)})) \quad (8)$$

Clean infer.

PIT model infer.



Experiments – Transfer Learning

L

Learn from ensemble

eacher

ASR From scratch v.s. Domain adaptation

Layers	Modular	teacher	WER	Rel. (%)
10·0	×	×	57.5	0
	×	9·1 ⊕ 6·4 ⊕ 3·7	55.0	-4.4
	×	clean	52.5	-8.7
6·4	×	×	52.8	-8.2
	×	clean	47.1	-18.0
	√	clean	38.9	-32.4
	√	MMI clean	35.8	-37.7

- Methods
  - Modular Initialization 4-10%
  - Transfer Learning Based Joint Training 20%
  - **Temporal Correlation Modeling** 8%
  - Multi-outputs Sequence Discriminative Training 8%

- Methods
  - Modular Initialization 4-10%
  - Transfer Learning Based Joint Training 20%
  - Temporal Correlation Modeling 8%
  - **Multi-output Sequence Discriminative Training** 8%

Experiments – Seq. Disc. Training

Performance Summary in SWBD 50 Hours Dataset

Neural network	Model	WER	Rel. (%)
10·0 BLSTM	PIT-CE	57.5	0
6·4 BLSTM	progressive joint training	50.2	-13
	+ clean teacher	38.9	-32.4
	+ MMI clean teacher	35.8	-37.7
	+ LF-DC-bMMI	35.2	-38.8
1 LACE + 5·4 BLSTM	progressive joint training	47.4	-17.5
	+ clean teacher	36.0	-37.4
	+ MMI clean teacher	34.6	-39.8
	+ LF-DC-bMMI	34.0	-40.9

Conclusion

Human Parity and Beyond

- Today’s systems can transcribe English conversational telephone speech at least as well as humans.
- There remain interesting areas where humans are still superior:
  - Distant speech
  - Overlapped speech
  - Accented speech
  - Multilingual speech
  - Language expansion
  - Speech understanding
- Solving these problems should keep the field busy for years to come.

END