Recent Advances in End-to-End Automatic Speech Recognition
Jinyu Li
• End-to-end (E2E) automatic speech recognition (ASR) fundamental

• E2E advances
  • Leveraging unpaired text
  • Multilingual ASR
  • Multi-talker ASR
  • Beyond ASR

• The next trend

• Conclusions
End-to-End Fundamental
Hybrid vs. End-to-End (E2E) Modeling

Hybrid
Separate models are trained, and then are used all together during testing in an ad-hoc way.

E2E
A single model is used to directly map the speech waveform into the target word sequence.
Advantages of E2E Models

• E2E models use a single objective function which is consistent with the ASR objective

• E2E models directly output characters or even words, greatly simplifying the ASR pipeline

• E2E models are much more compact than traditional hybrid models -- can be deployed to devices with high accuracy and low latency

Current Status

• E2E models achieve the state-of-the-art results in most benchmarks in terms of ASR accuracy.

• Practical challenges such as streaming, latency, adaptation capability etc., have been also optimized in E2E models.

• E2E models are now the mainstream models not only in academic but also in industry.
E2E Models

Connectionist Temporal Classification (CTC)

Attention-based encoder decoder (AED)

RNN-Transducer (RNN-T)
E2E Models

<table>
<thead>
<tr>
<th></th>
<th>CTC</th>
<th>AED</th>
<th>RNN-T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence assumption</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Attention mechanism</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Streaming</td>
<td>Natural</td>
<td>Additional work needed</td>
<td>Natural</td>
</tr>
<tr>
<td>Ideal operation scenario</td>
<td>Streaming</td>
<td>Offline</td>
<td>Streaming</td>
</tr>
<tr>
<td>Long form capability</td>
<td>Good</td>
<td>Weak</td>
<td>Good</td>
</tr>
</tbody>
</table>

RNN-T is the most popular E2E model in industry which requires streaming ASR.


Encoder is the Most Important Component

Connectionist Temporal Classification (CTC)

Attention-based encoder decoder (AED)

RNN-Transducer (RNN-T)
Encoder for RNN-T

\[ P(y_u|x_{1:t}, y_{1:u-1}) \]

softmax

Joint

\[ z_{t,u} \]

Prediction

\[ h_{u}^{pre} \]

Encoder

\[ h_{i}^{enc} \]

LSTM

Transformer

Conformer
Transformer

• Self-attention: computes the attention distribution over the input speech sequence
  \[ \alpha_{t,T} = \frac{\exp(\beta(W_q x_t)^T(W_k x_T))}{\sum_{T'} \exp(\beta(W_q x_t)^T(W_k x_{T'}))} \]

• Attention weights are used to combine the value vectors to generate the layer output
  \[ z_t = \sum_{T} \alpha_{t,T} W_v x_T = \sum_{T} \alpha_{t,T} v_T \]

• Multi-head self-attention: applies multiple parallel self-attentions on the input sequence

Vaswani et al. “Attention is all you need” NIPS 2017
Conformer

• Transformer: good at capturing global context, but less effective in extracting local patterns

• Convolutional neural network (CNN): works on local information

• Conformer: combines Transformer with CNN

Industry Requirement of Transformer Encoder

• Streaming with low latency and low computational cost

• Vanilla Transformer fails so because it attends the full sequence

• Solution: Attention mask is all you need
Attention Mask is All You Need

• Compute attention weight \( \{\alpha_{t,\tau}\} \) for time \( t \) over input sequence \( \{x_\tau\} \), binary attention mask \( \{m_{t,\tau}\} \) to control range of input \( \{x_\tau\} \) to use

\[
\alpha_{t,\tau} = \frac{m_{t,\tau} \exp(\beta (W_q x_t)^T (W_k x_\tau))}{\sum_{\tau'} m_{t,\tau'} \exp(\beta (W_q x_t)^T (W_k x_{\tau'}))} = \text{softmax}(\beta q_t^T k_\tau, m_{t,\tau})
\]

• Apply attention weight over value vector \( \{v_\tau\} \)

\[
z_t = \sum_{\tau} \alpha_{t,\tau} W_v x_\tau = \sum_{\tau} \alpha_{t,\tau} v_\tau
\]

Attention Mask is All You Need

• Offline (whole utterance)

Predicting output for $x_{10}$ Not streamable Attention Mask
Attention Mask is All You Need

• 0 lookahead, full history

Memory and runtime cost increase linearly
Attention Mask is All You Need

• 0 lookahead, limited history (3 frames)

In some scenario, small amount of latency is allowed
Attention Mask is All You Need

• Small lookahead (at most 2 frames), limited history (3 frames)

Look-ahead window [0, 2]

Predicting output for $x_{10}$
Live Caption in Windows 11
Advancing E2E Models

- unpaired text
- multi-talker
- multilingual
- speech translation
Unpaired Text
Leverage Unpaired Text

• Standard E2E models are trained with paired speech-text data, while hybrid models use large amount of text data for LM building.

• It is important to leverage unpaired text data for further performance improvement, especially in the domain adaptation task.
  • Adaptation with augmented audio
  • LM fusion
  • Direct adaptation with text data
Adaptation with Augmented Audio

• Adapt E2E models with the synthesized speech generated from the new domain text using TTS or from original ASR training data.

LM Fusion Methods

• Shallow Fusion
  ➢ A log-linear interpolation between the E2E and LM probabilities.

\[
\hat{Y} = \arg\max_Y \left[ \log P(Y|X; \theta_{E2E}^S) + \lambda_T \log P(Y; \theta_{LM}^T) \right]
\]

• Density Ratio Method
  ➢ Subtract source-domain LM score from Shallow Fusion score.

\[
\hat{Y} = \arg\max_Y \left[ \log P(Y|X; \theta_{E2E}^S) + \lambda_T \log P(Y; \theta_{LM}^T) - \lambda_S \log P(Y; \theta_{LM}^S) \right]
\]

• HAT/ILME-based Fusion
  ➢ Subtract internal LM score from Shallow Fusion score.

\[
\hat{Y} = \arg\max_Y \left[ \log P(Y|X; \theta_{E2E}^S) + \lambda_T \log P(Y; \theta_{LM}^T) - \lambda_I \log P(Y; \theta_{E2E}^I) \right]
\]

**Internal LM Estimation**

- **RNN-T**
  \[
P(\hat{y}_i | Y_{0:u-1}, X_{1:t_i}; \theta_{RNN-T}) = \text{softmax}(z_{t_i,u_i})
\]

- **Internal LM estimation of RNN-T**
  \[
P(y_u | Y_{0:u-1}; \theta_{pred}, \theta_{joint}) = \text{softmax}(z_u^{ILM,NB})
\]

- **Joint Network**
  - **Softmax**
  - **Prediction Network**
  - **Encoder**
  - **Acoustically-Conditioned LM**

- **Softmax**
  \[
z_{t_i,u_i}
\]

- **Joint Network**
  \[
z_u^{ILM,NB}
\]

- **Prediction Network**
  \[
y_{u-1}
\]

- **Encoder**
  \[
x_{1:t_i}
\]

- **Prediction Network**
  \[
h_u^{pred}
\]

- **Encoder**
  \[
h_{t_i}^{enc}
\]

- **Prediction Network**
  \[
h_{u_i}^{pred}
\]

- **Joint Network**
  \[
y_{u_i-1}
\]

**Internal LM probability**

- The output of the **acoustically-conditioned LM** after removing the contribution of the encoder

Factorized Neural Transducer

Functions as a neural LM. Can be adapted with text only data!

Multilingual ASR
Multilingual

- 40% people can speak only 1 language fluently.
- 43% people can speak only 2 languages fluently.
- 13% people can speak only 3 languages fluently.
- 3% people can speak only 4 languages fluently.
- <0.1% people can speak 5+ languages fluently.

- Human cannot recognize all languages. Can we build a single high quality multilingual model to serve all users?

Statistics are from: http://ilanguages.org/bilingual.php
Multilingual E2E Models

• Double-edged sword of pooling all language data
  • Maximum sharing between languages; One model for all languages
  • Confusion between languages
Configurable Multilingual Model

- **Universal module:** modeling the sharing across languages

- **Expert module:** modeling the residual from universal module for each language

```
0 1 0 1 1
```
Multi-talker ASR
Multi-talker ASR

• E2E ASR systems have high accuracy in single-speaker applications 😊

• Very difficult to achieve satisfactory accuracy in scenarios with multiple speakers talking at the same time 😞

• Solutions: E2E multi-talker models
Serialized Output Training (SOT)

how are you \(<sc>\) I am fine thank you \(<eos>\)

iterate for label $i$

softmax

decoder

encoder

Attention

how are you

I am fine thank you

Token-level Serialized Output Training (t-SOT)

Multi-talker transcription

<table>
<thead>
<tr>
<th>Virtual channel 1</th>
<th>hello how are</th>
<th>you</th>
<th>good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtual channel 2</td>
<td>i am</td>
<td>fine thank</td>
<td>you</td>
</tr>
</tbody>
</table>

Serialized transcription

Deserialization

Streaming E2E ASR

Recognition with low latency

Audio stream

Continuous audio input

Speaker 1

hello  how  are  you

Speaker 2

i  am  fine  thank  you

Speaker 3

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>good</td>
</tr>
</tbody>
</table>

Beyond ASR
E2E Speech Translation (ST)

- ASR is often the first step in a pipeline and is followed by
  - machine translation
  - speech synthesis (→ speech-to-speech translation)
  - natural language understanding / generation, etc.
Streaming Multilingual Speech Model (SM^2)

• Multilingual data is pooled together to train a streaming model to perform both ST and ASR functions.

• ST training is totally weakly supervised without using any human labeled parallel corpus.

• The model is very small, running on devices.

Simultaneous ST Demo
Foundation Models

Whisper from OpenAI

• Trained from 680k hours weakly supervised data collected from the web.

• A single model can perform multiple tasks: multilingual ASR + speech translation (to English), language identification, etc.

• Outstanding zero-shot capability

Universal Speech Understanding (USM) model

**Better Self-Sup Algorithm: BEST-RQ**

**Language and Domain expansion**

**Modality Expansion (Speech + Text)**

**Model Delivery (Adaptation, Modularization)**

- A 2B pre-trained encoder, finetuned by your favorite decoder
  - AED / CTC / RNN-T

- Expanded to 289 languages

- Maestro + Best-RQ

- Joint finetune from speech and text data.

- USM-CTC: 800x real-time on TPUv4i

---

Credit to Yu Zhang

What’s the Next Trend?
Large Language Model (LLM)-centric Integrative AI

Image credit to XD Huang
What does GPT-V mean to computer vision?

**Logo Recognition**

GPT-V: The image shows a close-up of a person’s foot wearing a white Nike Air Force 1 sneaker. The sneaker is a low-top style with a white rubber sole and white laces. ...

**Object Counting**

GPT-V: There are 7 apples in the image.

**Object Localization/Detection**

GPT-V: There is only one person in the image, so the bounding box for this person would be:
- Top left: (254, 118)
- Bottom right: (731, 693)

**Dense Captioning**

GPT-V: Geoffrey Hinton, computer scientist and cognitive psychologist, known for his work on artificial neural networks.

What does GPT-V mean to computer vision?

**Image Recognition**

**GPT-V** The object in the image is a traffic light.

**Celebrity Recognition**

**GPT-V** The person in the image is Jensen Huang, the CEO and co-founder of NVIDIA. He is holding and presenting a product from NVIDIA, likely a graphics processing unit (GPU).

**Landmark Recognition**

**GPT-V** The landmark in the image is the Space Needle, an observation tower located in Seattle, Washington. It stands 605 feet tall and offers panoramic views of the city and surrounding areas. …

**Food Recognition**

**GPT-V** This dish is called a Cajun Sampler. It is a sampler platter that consists of four small bowls of traditional Cajun dishes - gumbo, jambalaya, red beans and rice, and crawfish étouffée. …
**What does GPT-V mean to computer vision?**

**TASK**  
**Scene Text Recognition**

The scene text in the image are:
- Vacation - Beach - Smile
- Summer - Watermelon
- Flip flops - Lemonade

**GPT-V**

**TASK**  
**Document Reasoning**

Which country has the highest profit margin?

<table>
<thead>
<tr>
<th>Country</th>
<th>Sales Volume</th>
<th>Revenue</th>
<th>Profit</th>
<th>Profit Margin</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S.</td>
<td>$15,971,080</td>
<td>$3,006,421</td>
<td>19.3%</td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>$15,896,670</td>
<td>$3,032,162</td>
<td>19.1%</td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>$14,812,560</td>
<td>$2,864,036</td>
<td>19.4%</td>
<td></td>
</tr>
<tr>
<td>India</td>
<td>$10,608,174</td>
<td>$1,853,710</td>
<td>17.5%</td>
<td></td>
</tr>
<tr>
<td>South Korea</td>
<td>$10,494,948</td>
<td>$1,975,544</td>
<td>18.8%</td>
<td></td>
</tr>
<tr>
<td>Total/Avg</td>
<td>$141,282</td>
<td>$69,754,238</td>
<td>$12,816,772</td>
<td>18.8%</td>
</tr>
</tbody>
</table>

Based on the data provided, the country with the highest profit margin is **Australia**, with a profit margin of 19.4%.

**GPT-V**

**TASK**  
**Multilingual Image Description**


**GPT-V**

**TASK**  
**Multilingual Text Recognition**

The text in the image is in Chinese and translates to "Mapo Tofu." Mapo tofu is a popular Chinese dish made with tofu and minced meat in a spicy sauce.

**GPT-V**

Image credit to Lijuan Wang
### Video Activity Anticipation

**GPT-V**

The images show the process of making sushi rolls. The first image shows rice being mixed with vinegar in a bowl. The second image shows the rice being spread onto a sheet of seaweed. The third image shows the addition of salmon and cucumber to the roll. The next steps would be to **roll the sushi using a bamboo mat, cut the roll into smaller pieces, and serve with soy sauce and wasabi**.

### Video Action Recognition

**GPT-V**

The images show a person in a motion capture suit in a studio with multiple cameras and sensors. The person is **performing a push-up exercise**. In the first image, the person is in the starting position with their arms extended and body in a straight line. In the second image, the person is lowering their body towards the ground. In the third image, the person is at the bottom of the push-up with their chest close to the ground. In the fourth image, the person is pushing their body back up to the starting position.

*Image credit to Lijuan Wang*
VALL-E: Neural codec language model

- High quality Zero shot TTS: In context learning through prompts
  - “Steal voice from 3 second's prompt”

SpeechX – A versatile speech generation model

**Versatility:** able to handle a wide range of tasks from audio and text inputs.

**Robustness:** applicable in various acoustic distortions, especially in real-world scenarios where background sounds are prevalent.

**Extensibility:** flexible architectures, allowing for seamless extensions of task support.

---

VioLA: A multi-modal model with discrete audio inputs

- An extension to audio codec language model
- Naturally merge speech-language tasks
  - Speech recognition
  - Machine translation
  - Speech generation

Speech and text can freely serve as input and output

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
<th>Typical Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>Text</td>
<td>ASR, ST</td>
</tr>
<tr>
<td>Text</td>
<td>Text</td>
<td>MT, LM</td>
</tr>
<tr>
<td>Text</td>
<td>Speech</td>
<td>multilingual TTS</td>
</tr>
</tbody>
</table>

Advancing Speech-LLM For In-context Learning

- Trained tasks (EN only)
  - ASR
  - Speech-based Question Answering

- Emergent Capable tasks
  - 0-shot and 1-shot En->X ST
  - 1-shot domain adaptation
  - Instruction-followed ASR

Conclusions

• E2E models are now the mainstreaming ASR models.
  • Streaming Transformer Transducer with masks can achieve very high accuracy and low latency.

• To further advance E2E models, we have discussed several key technologies.
  • Leverage unpaired text: domain adaptation
  • Multilingual: configurable multilingual model
  • Multi-talker ASR: (token-level) serialized output training
  • Speech translation: streaming multilingual speech model

• Large language model (LLM) centric integrative AI may be the next trend.
Thank You!