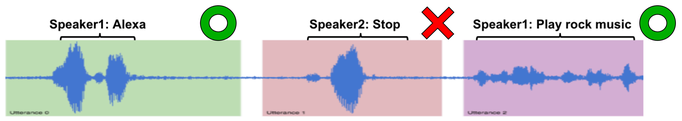
End-to-End Anchored Speech Recognition

Yiming Wang[1], Xing Fan[2], I-Fan Chen[2], Yuzong Liu[2], Tongfei Chen[1], Björn Hoffmeister[2]
[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
[2]Alexa Automatic Speech Recognition
{yiming.wang, tongfei}@jhu.edu, {fanxing, ifanchen, liuyuzon, bjorhn}@amazon.com

Problem Definition

Anchored Speech Recognition

- To distinguish target speaker from interfering speakers and background speech/noises and only recognize speech from the target speaker



- Interfering speech: A challenging problem in far-field Automatic Speech Recognition (ASR) which causes
  - Undesired insertion and misrecognition errors
  - End-pointing delay

Previous Work

- **Feature based Anchored Speech Recognition**
  - Feed in speech recognition system with additional speaker representation features extracted from anchored words
  - Speaker Representations
    - mean-variance normalization, maximum likelihood linear regression (MLLR), i-vector, D-vector, X-vector, anchored mean subtraction (AMS), encoder-decoder network, etc.
  - **Pros**: Easy to implement, decent performance
  - **Cons**: Does not really distinguish speaker differences between target and interfering speech due to limited information capacity of the conventional ASR model architecture

End-to-End Anchored Speech Recognition

Attention-based Encoder-Decoder Model

- Attention mechanism enables ASR systems to focus on recognizing only speech from target speakers
  - **Multi-Source Attention**
  - **Mask-based Attention**

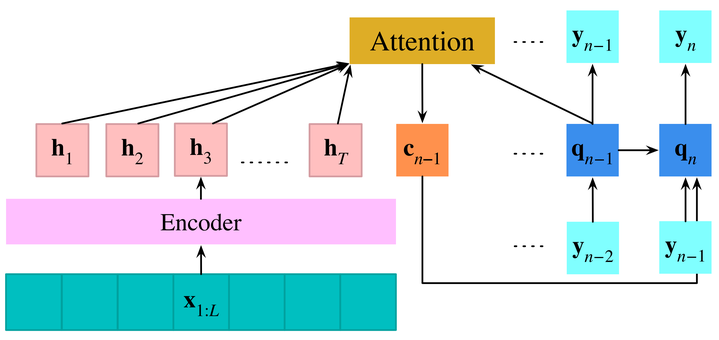


Fig. 1. Attention-based Encoder-Decoder Model. It is an illustration in the case of a one-layer decoder. If there are more layers, as in our experiments, an updated context vector $\mathbf{c}_n$ will also be fed into each of the upper layers in the decoder at the time step n.

- **Conventional attention-based encoder-decoder model**

$$\mathbf{h}_{1:T} = \text{Encoder}(\mathbf{x}_{1:L}) \quad (1)$$
$$\alpha_{n,t} = \text{Attention}(\mathbf{q}_n, \mathbf{h}_t) \quad (2)$$
$$\mathbf{c}_n = \sum_t \alpha_{n,t}\mathbf{h}_t \quad (3)$$
$$\mathbf{q}_n = \text{Decoder}(\mathbf{q}_{n-1}, [y_{n-1}; \mathbf{c}_{n-1}]) \quad (4)$$
$$y_n = \arg\max_v(\mathbf{W}^f\mathbf{q}_n + \mathbf{b}^f) \quad (5)$$
$$\omega_{n,t} = \mathbf{v}^\top \tanh(\mathbf{W}^q\mathbf{q}_n + \mathbf{W}^h\mathbf{h}_t + \mathbf{b}) \quad (6)$$
$$\alpha_{n,t} = \text{softmax}(\omega_{n,t}) \quad (7)$$

Bahdanau Attention

- **Multi-Source Attention**
  - Speaker representation encoder: **S-Encoder**

Anchored Word

$$\tilde{\mathbf{w}} = \text{Pooling}(\text{S-Encoder}(\overrightarrow{\mathbf{w}_{1:L'}})) \quad (8)$$
$$\mathbf{u}_{1:T} = \text{S-Encoder}(\mathbf{x}_{1:L}) \quad (9)$$
$$\phi_t = \text{Similarity}(\mathbf{u}_t, \tilde{\mathbf{w}}) \quad (10)$$

Eq (7) → $\alpha_{n,t}^{\text{anchor\_aware}} = \text{softmax}(\omega_{n,t} + g \cdot \phi_t) \quad (11)$

Eq (3) → $\mathbf{c}_n = \sum_t \alpha_{n,t}^{\text{anchor\_aware}}\mathbf{h}_t \quad (12)$

- **Mask-based Attention**

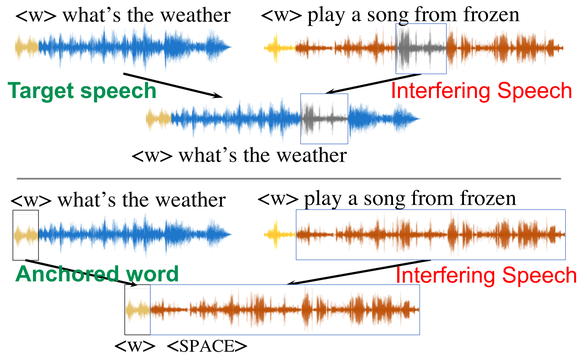
$$\phi_t = \text{sigmoid}(g \cdot \text{Similarity}(\mathbf{u}_t, \tilde{\mathbf{w}})) \quad (13)$$
$$\mathbf{h}_t^{\text{masked}} = \phi_t\mathbf{h}_t \quad (14)$$

Eq (6) → $\omega_{n,t} = \mathbf{v}^\top \tanh(\mathbf{W}^q\mathbf{q}_n + \mathbf{W}^h\mathbf{h}_t^{\text{masked}} + \mathbf{b}) \quad (15)$

Eq (3) → $\mathbf{c}_n = \sum_t \alpha_{n,t}\mathbf{h}_t^{\text{masked}} \quad (16)$

Interfering Speech Training Data Synthesis

- Two types of synthetic methods
  - **Method 1**: Interfering segment insertion
  - **Method 2**: Complete interfering



Multi-task Training for Mask-based Attention Model

- For synthesized training data, we have ground truth for the mask of target speech – which can be used to train mask-based attention model in a supervised manner.
- Combine the normal ASR CE loss with Mask CE loss with interpolation weight $(1-\lambda)\mathcal{L}_{ASR} + \lambda\,\mathcal{L}_{Mask}$

Experiments

Experimental Setup

- Dataset
  - Training: 1200-hour manual transcribed English Amazon Echo live data with same wake word. Mostly clean condition utterances
  - Test datasets
    - **Normal set** (25k words) – similar to training data condition (clean)
    - **Hard set** (5.4k words) – live data containing interfering speech
- E2E ASR systems
  - Input: 64-dim LFBE feature; Output: Graphemes for beam search (beam size = 15) with vocabulary
  - **Baseline**
    - **Enc**: 3 Conv Layers (with down samplings) + 3 BLSTM Layers; **Dec**: 3 uni-LSTM (320-dim) layers
  - **Multi-Source Attention** – **S-Enc**: 3-Conv layers (same as **Enc**)
  - **Mask-based Attention** – **S-Enc**: 3-Conv layers + 1 BLSTM layer

Table 2. Augmented vs. Device-directed-only training data.

Model	Training Set	Test Set	WER	sub	ins	del	WERR(%)
Baseline	Device-directed-only	normal	1.000	0.715	0.108	0.177	—
		hard	3.354	1.762	1.123	0.469	—
	Augmented	normal	3.215	1.223	0.038	1.954	-221.5
		hard	4.208	1.777	0.246	2.185	-30.9
Mul-src. Attn.	Device-directed-only	normal	1.015	0.731	0.115	0.169	**-1.5**
		hard	3.262	1.746	1.062	0.454	+2.8
	Augmented	normal	1.015	0.700	0.108	0.207	**-1.5**
		hard	2.854	1.569	0.723	0.562	**+14.9**

Table 3. Mask-based Model: with and without mask supervision.

Model	Training Set	Test Set	WER	sub	ins	del	WERR(%)
w/o Supervision	Augmented	normal	1.348	0.725	0.096	0.527	-34.8
		hard	3.223	1.508	0.628	1.087	+3.9
w/ Supervision	Augmented	normal	1.030	0.715	0.115	0.200	**-3.0**
		hard	2.931	1.586	0.809	0.536	**+12.6**

Conclusion

- Two approaches for E2E anchored speech recognition are proposed: **Multi-source Attention** and **Mask-based Attention**
- Two ways of interfering speech training data synthesis are proposed addressing training data sparsity issue in anchored speech recognition task – provides **~12% relative improvement** (+2.8% → +14.9%)
- A multi-task training scheme for Mask based model is also proposed
- **15% WER reduction** on test data with interfering background speech; while with only a minor degradation of 1.5% on clean speech.