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**

### Interfering Speech Training Data Synthesis
- **Two types of synthetic methods**
  - **Method 1**
    - Interfering segment insertion
    - \( <w> \) ‘what’s the weather’
    - \( <w> \) ‘play a song from frozen
  - **Method 2**
    - Complete interfering
    - \( <w> \) ‘what’s the weather’
    - \( <w> \) ‘play a song from frozen

### 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 \( (1 - \lambda) \) \( L_{CE} \) + \( \lambda \) \( L_{MCE} \).

---

**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** (5k 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>
<thead>
<tr>
<th>Model</th>
<th>Training Set</th>
<th>Test Set</th>
<th>WER</th>
<th>sub</th>
<th>del</th>
<th>WERR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td>normal 1.000</td>
<td>0.715</td>
<td>0.108</td>
<td>0.177</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Device-directed-only</td>
<td>hard 3.534</td>
<td>1.762</td>
<td>1.123</td>
<td>0.469</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Augmented</td>
<td>normal 3.215</td>
<td>1.223</td>
<td>0.038</td>
<td>1.954</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Device-directed-only</td>
<td>hard 4.208</td>
<td>1.777</td>
<td>0.246</td>
<td>2.185</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td><strong>Augmented</strong></td>
<td>normal 1.015</td>
<td>0.731</td>
<td>0.115</td>
<td>0.169</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Device-directed-only</td>
<td>hard 3.262</td>
<td>1.746</td>
<td>1.062</td>
<td>0.454</td>
<td>+2.8</td>
<td></td>
</tr>
<tr>
<td><strong>Augmented</strong></td>
<td>normal 1.015</td>
<td>0.700</td>
<td>0.108</td>
<td>0.207</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Device-directed-only</td>
<td>hard 2.854</td>
<td>1.569</td>
<td>0.723</td>
<td>0.562</td>
<td>+14.9</td>
<td></td>
</tr>
</tbody>
</table>

| Table 3. Mask-based Model: with and without mask supervision. |
|--------------------------|--------------------------|--------------------------|--------------------------|
| **Model**                | Training Set | Test Set | WER | sub | del | WERR (%) |
| **w/o** Supervision      | normal 1.348 | 0.725    | 0.096 | 0.527 | —  |
| **w/** Supervision       | hard 3.232   | 1.508    | 0.628 | 1.087 | +3.9 |
| **w/o** Supervision      | normal 1.030 | 0.715    | 0.115 | 0.200 | —  |
| **w/** Supervision       | 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% to +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.