SLT 2018 recap

18-21 December, Athens, Greece
Talks

YUN-NUNG (VIVIAN) CHEN: Towards Open-Domain Conversational AI

Dirk Hovy: Bias in Language Data

Pınar Yolum: Privacy, Trust, and Ethics

Dimitra Vergyri: Speech-In-The-Wild Analytics in the Era of Deep Learning: Recent Advancements and Remaining Challenges

Karen Livescu: Acoustic (and acoustically grounded) word embeddings

http://www.slt2018.org/presentations/
Poster sessions

Deep Learning for Speech Synthesis
ASR I, ASR II, ASR III (End-to-End), ASR IV
Spoken Language Understanding
Corpora and Evaluation Methodologies Detection, Paralinguistics and Coding
Dialogue
Speaker Recognition/Verification
Natural Language Processing
Voice Conversion and TTS
Learning Noise-Invariant Representations for Robust Speech Recognition
Davis Liang, Zhiheng Huang, Amazon AI; Zachary Lipton, Carnegie Mellon University

- a clean example and its superficially perturbed counterparts shouldn't merely map to the same class - they should map to the same representation
- enforce similarity in hidden representations
- penalize L2/cosine distance between the intermediate representations at the encoder or encoder+decoder
IMPROVING LF-MMI USING UNCONSTRAINED SUPERVISIONS FOR ASR
Hossein Hadian, Sharif University of Technology; Daniel Povey, Johns Hopkins University; Hossein Sameti, Sharif University of Technology; Jan Trmal, Sanjeev Khudanpur, Johns Hopkins University

- improving the numerator graph
- scheme for creating unconstrained numerator graphs by removing time constraints from the baseline numerator graphs
- relax the supervision time constraints in each chunk, so that the numerator graph is not acyclic anymore (it will have self-loops)
- this leads to much smaller graphs
- 2x speed-up in preparing supervisions for DNN training

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI-SDM</td>
<td>dev</td>
<td>37.1</td>
<td><strong>36.8</strong></td>
<td>0.8%</td>
</tr>
<tr>
<td></td>
<td>eval</td>
<td>40.7</td>
<td><strong>40.5</strong></td>
<td>0.5%</td>
</tr>
<tr>
<td>Switchboard</td>
<td>eval2000</td>
<td>13.1</td>
<td><strong>12.8</strong></td>
<td>2.3%</td>
</tr>
<tr>
<td></td>
<td>RT03</td>
<td>15.4</td>
<td><strong>15.0</strong></td>
<td>2.6%</td>
</tr>
<tr>
<td>TEDLIUM</td>
<td>dev</td>
<td>7.4</td>
<td><strong>7.3</strong></td>
<td>1.4%</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>7.7</td>
<td><strong>7.6</strong></td>
<td>1.3%</td>
</tr>
<tr>
<td>Librispeech</td>
<td>dev</td>
<td>3.3</td>
<td><strong>3.3</strong></td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>dev-other</td>
<td>8.8</td>
<td><strong>8.7</strong></td>
<td>1.1%</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>3.8</td>
<td><strong>3.8</strong></td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>test-other</td>
<td>9.0</td>
<td><strong>8.8</strong></td>
<td>2.2%</td>
</tr>
<tr>
<td>WSJ</td>
<td>dev93</td>
<td>4.3</td>
<td><strong>4.4</strong></td>
<td>-2.1%</td>
</tr>
<tr>
<td></td>
<td>eval92</td>
<td>2.5</td>
<td><strong>1.5</strong></td>
<td>0%</td>
</tr>
<tr>
<td>Mini-librispeech</td>
<td>dev</td>
<td><strong>8.5</strong></td>
<td>8.6</td>
<td>-1.1%</td>
</tr>
</tbody>
</table>

Table 4. Summary of results on various databases. Last column shows the relative improvement in word error rates.
Unsupervised Representation Learning of Speech for Dialect Identification
Suwon Shon, Wei-Ning Hsu, James Glass

- factorized hierarchical variational autoencoder (FHVAE) model to learn an unsupervised latent representation for dialect identification (DID)
- separate static and dynamic factors
- language related information (tone, accent, rhythm) are mostly encoded in the dynamic factors

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>EER</th>
<th>$C_{avg}$ *100</th>
</tr>
</thead>
<tbody>
<tr>
<td>i-vector</td>
<td>57.44</td>
<td>24.43</td>
<td>23.79</td>
</tr>
<tr>
<td>End-to-end (MFCC)</td>
<td>65.55</td>
<td>20.24</td>
<td>19.92</td>
</tr>
<tr>
<td>End-to-end (FBANK)</td>
<td>64.81</td>
<td>20.22</td>
<td>19.91</td>
</tr>
<tr>
<td><strong>End-to-end (FHVAE)$z_1$</strong></td>
<td><strong>67.98</strong></td>
<td><strong>18.62</strong></td>
<td><strong>18.32</strong></td>
</tr>
<tr>
<td>End-to-end (FHVAE)$z_2$</td>
<td>54.55</td>
<td>27.39</td>
<td>27.35</td>
</tr>
</tbody>
</table>
E2E ASR with n-gram context
- Contextual Listen, Attend and Spell (CLAS) jointly optimizes the ASR components along with embeddings of the context n-grams
- Comparison with a more traditional contextualization approach, which performs shallow-fusion between independently trained LAS and contextual n-gram models during beam search
- 68% relative WER, indicating the advantage of joint optimization over individually trained components
LEVERAGING SEQUENCE-TO-SEQUENCE SPEECH SYNTHESIS FOR ENHANCING
ACOUSTIC-TO-WORD SPEECH RECOGNITION
Masato Mimura, Sei Ueno, Hirofumi Inaguma, Shinsuke Sakai, Tatsuya Kawahara, Kyoto University

- word-based models commonly suffer from the out-of-vocabulary (OOV) word problem
- explore how to leverage the current speech synthesis technology to tailor the ASR system for a
  target domain by preparing only a relevant text corpus
- from a set of target domain texts, generate speech features using a sequence-to-sequence speech
  synthesizer
- the artificial speech features together with real speech features from conventional speech corpora
  are used to train an attention-based A2W model
BACK-TRANSLATION-STYLE DATA AUGMENTATION FOR END-TO-END ASR
Tomoki Hayashi, Nagoya University; Shinji Watanabe, Johns Hopkins University; Yu Zhang, Google; Tomoki Toda, Nagoya University; Takaaki Hori, Mitsubishi Electric Research Laboratories; Ramon Astudillo, INESC-ID-Lisboa; Kazuya Takeda, Nagoya University

- back-translation -- a pre-trained target-to-source translation model is used to generate source text from unpaired target text
- use text not paired with speech for data augmentation
- after training, the text-to-encoder model generates the hidden states from a large amount of unpaired text, then E2E-ASR decoder is retrained using the generated hidden states as additional training data.
- targeting the states of the speech encoder, rather than speech itself makes it possible to achieve faster attention learning and reduce computational cost, thanks to sub-sampling present in E2E-ASR encoder

Fig. 3: Flowchart of proposed retraining.
Kullback-Leibler divergence (KLD) regularization and multi-task learning (MTL)

The MTL adaptation improves the baseline SI word CTC model by up to 8.8% and 4.0% relative WERRs for supervised and unsupervised adaptation, and obtains up to 9.6% and 3.8% WERRs over the baseline SI mix-unit CTC model, respectively.

Results show that the MTL adaptation performs better than KLD adaptation for both word and mix-unit CTC models.
- **Improving Unsupervised Style Transfer** in End-to-End Speech Synthesis with End-to-End Speech Recognition

- Speech Chain for Semi-Supervised Learning of Japanese-English Code-Switching ASR and TTS

- Transliteration Based Approaches to Improve Code-Switched Speech Recognition Performance

- **Confidence Estimation and Deletion Prediction** Using Bidirectional Recurrent Neural Networks

- **Dialog-Context Aware** End-to-End Speech Recognition

- **Quaternion** Convolutional Neural Networks for Theme Identification of Telephone Conversations

- A Comparison of Techniques for **Language Model Integration** in Encoder-Decoder Speech Recognition

- Efficient **Implementation** of Recurrent Neural Network Transducer in Tensorflow