Incorporating articulatory feature information in ASR transfer learning

Mahir Morshed

8 April 2020

Outline

Articulatory features

- Facets of phone production by which differences between such phones may be characterized
- Although two languages may lack a common phone, close equivalents may exist which differ in one or two characteristics
  - /ʈ/ in South Asian languages vs /t/ elsewhere
  - /e/ vs /ɛ/ elsewhere
  - /r/ vs /ɹ/ elsewhere
  - /a/ vs /ɑ/ elsewhere
  - ...
- (such near-equivalencies often manifest in language acquisition process)

Articulatory feature detection

Softmax output layer
4 x FC layer 2048
Max pooling 2x3
2x Convolational layer 3x2, 256
Max pooling 2x2
2x Convoluational layer 3x3, 256
Max pooling 2x2
2x Convolutational layer 3x3, 128
Max pooling 1x2
2x Convolutional layer 3x3, 128
Max pooling 1x2
2x Convolutional layer 3x3, 64
Mel Fbank layer
Input layer

Figure: CNN-based feature detector structure described by Merkx and Scharenborg.

- Binary feature detectors using fully-connected networks [1]
- CNN-based multiclass detection [2] (place and manner only)
- CTC-based multiclass detection [3]
- More recently, simultaneous feature detection using transformer-based systems [4]

End-to-end transfer learning

- Provide a better initial state from which to impart shared information between to a network
- Frequently based on retraining layers in a network already exposed to one language
    - More recent efforts see strapping of language models to new systems

End-to-end transfer learning

- Freezing lower layers of a system à la Wave2Letter for retraining an English system on German speech [5]
- More recently, fusing language-independent CTC-based model near the output of the target language network [6]

$$
\begin{aligned}
\boldsymbol{s}_u^{\mathrm{LM}} &= \boldsymbol{W}^{\mathrm{LM}}\boldsymbol{d}_u^{\mathrm{LM}} + \boldsymbol{b}^{\mathrm{LM}} \\
\boldsymbol{g}_u &= \sigma(\boldsymbol{W}^{\mathrm{g}}[\boldsymbol{s}_u^{\mathrm{S2S}}; \boldsymbol{s}_u^{\mathrm{LM}}] + \boldsymbol{b}^{\mathrm{g}}) \\
\boldsymbol{s}_u^{\mathrm{CF}} &= \boldsymbol{W}^{\mathrm{CF}}[\boldsymbol{s}_u^{\mathrm{S2S}}; \boldsymbol{g}_u \odot \boldsymbol{s}_u^{\mathrm{LM}}] + \boldsymbol{b}^{\mathrm{CF}} \\
P_{\mathrm{S2S}}(\boldsymbol{y}|\boldsymbol{x}) &= \mathrm{softmax}(\mathrm{ReLU}(\boldsymbol{W}^{\mathrm{out}}\boldsymbol{s}_u^{\mathrm{CF}} + \boldsymbol{b}^{\mathrm{o}}))
\end{aligned}
$$

Figure: 'Cold fusion' as described by Inaguma et al.

Corpora

- Large corpora of Bengali, Nepali, Sinhalese, Javanese, and Sundanese speech [7]
- In the absence of train/dev/test splits, created some myself (80/10/10)
- Due to differing corpora sizes, may not use all data in each split to aid comparability

Corpora splits

Language	Train (h)	Dev/Test (h)
bn	172.43	21.56
jv	236.70	29.59
ne	123.71	15.47
si	179.6	23
su	266.06	33.3

- Figures shown are upper limits of data used for training feature detectors.

Phonologies

- Festvox articulation information provided with each corpus
- Except for "schwa-like" (absent in Bengali and Nepali), all classes have at least one member in each language
    - Omitted from consideration since not uniquely contrastive in the five languages
    - Not all categories need have detectors before training low-resource language
- Lexicons also provided with phonetic transcriptions
    - Thrax G2P for each language available for out-of-vocabulary words

Articulatory feature classes

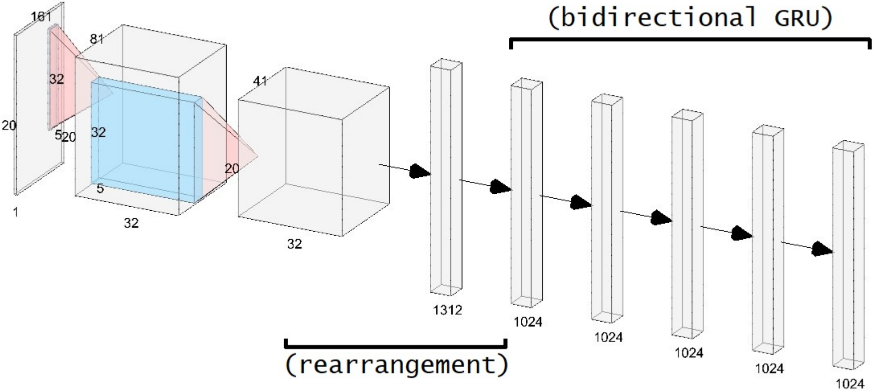
Language	features
manner	"stop", "affricate", "nasal", "approximant", "fricative"
place	"velar", "postalveolar", "alveolar", "dental", "labial", "glottal", "palatal"
voice	"unvoiced", "voiced"
height	"close", "close-mid", "mid", "near-open", "open"
length	"short"
frontness	"front", "central", "back"
round	"unrounded", "rounded"

- Articulatory feature set for Bengali.

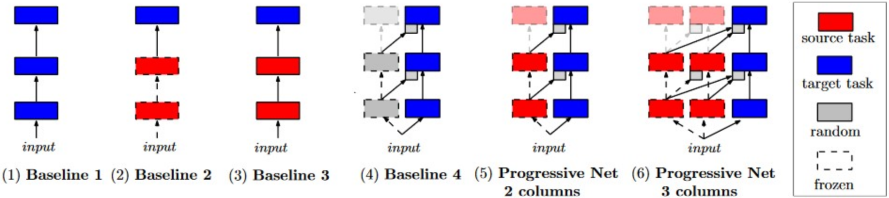
Networks

- Feature detectors à la DeepSpeech2
  - Some convolutional layers, then a series of recurrent layers, and a fully connected output layer
- Transfer learning environment: progressive networks [8]
  - Here similarly constructed pre-trained networks placed in parallel with a new network
  - Gates connect each recurrent layer of the pre-trained networks (kept constant) to their equivalent in the new network
  - (Originally developed for reinforcement learning activities)
- Input provided as log-spectrograms (20ms Hamming window, 10ms overlap); implementations in PyTorch

Feature detectors



Progressive networks



(1) **Baseline 1** (2) **Baseline 2** (3) **Baseline 3** (4) **Baseline 4** (5) **Progressive Net 2 columns** (6) **Progressive Net 3 columns**

- (Diagram from [9].)

Setups

- Test all pairings among the five languages (16 total) as follows:
  - Train feature detectors on a 'high-resource' language (40 epochs)
  - Connect detectors to comprehensive phone recognizer for 'low-resource' language
  - Train said recognizer on varied low-resource data sizes (1h, 5h, 10h for train/test) as part of progressive network
- Evaluation based on phone-error rates
- More info on this front to come...

- Alter the gating patterns in the progressive network
- Substitute other languages than the aforementioned five as low-resource languages
- Adjust feature detector architecture
- Consider using fewer detectors in the progressive network
- . . .

References

T. Bhowmik, A. Chowdhury, and S. K. Das Mandal, "Deep neural network based place and manner of articulation detection and classification for bengali continuous speech," vol. 125, pp. 895–901.

D. Merkx and O. Scharenborg, "Articulatory feature classification using convolutional neural networks," in *Interspeech 2018*. ISCA, pp. 2142–2146.

B. Abraham, S. Umesh, and N. M. Joy, "Articulatory feature extraction using CTC to build articulatory classifiers without forced frame alignments for speech recognition," pp. 798–802.

S. Li, C. Ding, X. Lu, P. Shen, T. Kawahara, and H. Kawai, "End-to-end articulatory attribute modeling for low-resource multilingual speech recognition," in *Interspeech 2019*. ISCA, pp. 2145–2149.

J. Kunze, L. Kirsch, I. Kurenkov, A. Krug, J. Johannsmeier, and S. Stober, "Transfer learning for speech recognition on a budget," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics, pp. 168–177.

H. Inaguma, J. Cho, M. K. Baskar, T. Kawahara, and S. Watanabe, "Transfer learning of language-independent end-to-end ASR with language model fusion," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6096–6100.

O. Kjartansson, S. Sarin, K. Pipatsrisawat, M. Jansche, and L. Ha, "Crowd-sourced speech corpora for javanese, sundanese, sinhala, nepali, and bangladeshi bengali," in *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*. ISCA, pp. 52–55.

L. Qu, C. Weber, E. Lakomkin, J. Twiefel, and S. Wermter, "Combining articulatory features with end-to-end learning in speech recognition," in *Artificial Neural Networks and Machine Learning ICANN 2018*, ser. Lecture Notes in Computer Science, V. Krkov, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis, Eds. Springer International Publishing, pp. 500–510.

A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.

Thank you!