References

[1] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182, 2016.

[3] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7304–7308. IEEE, 2013.

[4] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. Speech model pre-training for end-to-end spoken language understanding. *arXiv preprint arXiv:1904.03670*, 2019.

[5] Rakesh Gupta. Free-speech command classification for car navigation system, January 22 2013. US Patent 8,359,204.

[6] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*, 2018.

[7] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100, 2014.

[8] Ye-Yi Wang, Li Deng, and Alex Acero. Spoken language understanding. *IEEE Signal Processing Magazine*, 22(5):16–31, 2005.

[9] Mirco Ravanelli and Yoshua Bengio. Interpretable convolutional filters with sincnet. *arXiv preprint arXiv:1811.09725*, 2018.

[10] William Song and Jim Cai. End-to-end deep neural network for automatic speech recognition. *Standford CS224D Reports*, 2015.

[11] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*, 2016.

[12] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[13] Paul Lamere, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf. The cmu sphinx-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, volume 1, pages 2–5, 2003.

[14] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[15] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[16] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE, 2016.

[17] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM, 2006.

[18] Andrew Rosenberg, Kartik Audhkhasi, Abhinav Sethy, Bhuvana Ramabhadran, and Michael Picheny. End-to-end speech recognition and keyword search on low-resource languages. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5280–5284. IEEE, 2017.

[19] Özgür Cetin, Madelaine Plauché, and Udhaykumar Nallasamy. Unsupervised adaptive speech technology for limited resource languages: A case study for tamil. In *Spoken Languages Technologies for Under-Resourced Languages*, 2008.

[20] Jonas Lööf, Christian Gollan, and Hermann Ney. Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a polish speech recognition system. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[21] Ngoc Thang Vu, Franziska Kraus, and Tanja Schultz. Rapid building of an asr system for under-resourced languages based on multilingual unsupervised training. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[22] Julius Kunze, Louis Kirsch, Ilia Kurenkov, Andreas Krug, Jens Johannsmeier, and Sebastian Stober. Transfer learning for speech recognition on a budget. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 168–177, 2017.

[23] Ngoc Thang Vu, Florian Metze, and Tanja Schultz. Multilingual bottle-neck features and its application for under-resourced languages. In *SLTU*, 2012.

[24] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, 2018.

[25] Sibel Yaman, Li Deng, Dong Yu, Ye-Yi Wang, and Alex Acero. An integrative and discriminative technique for spoken utterance classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1207–1214, 2008.

[26] Jinfeng Rao, Ferhan Ture, and Jimmy Lin. Multi-task learning with neural networks for voice query understanding on an entertainment platform. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 636–645. ACM, 2018.

[27] Xiaodong He and Li Deng. Speech-centric information processing: An optimization-oriented approach. *Proceedings of the IEEE*, 101(5):1116–1135, 2013.

[28] Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. Towards end-to-end spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758. IEEE, 2018.

[29] Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters. From audio to semantics: Approaches to end-to-end spoken language understanding. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 720–726. IEEE, 2018.

[30] Yuan-Ping Chen, Ryan Price, and Srinivas Bangalore. Spoken language understanding without speech recognition. In *2018 IEEE International Confer-*

ence on Acoustics, Speech and Signal Processing (ICASSP), pages 6189–6193. IEEE, 2018.

[31] Chunxi Liu, Jan Trmal, Matthew Wiesner, Craig Harman, and Sanjeev Khudanpur. Topic identification for speech without asr. *Proc. Interspeech 2017*, pages 2501–2505, 2017.

[32] Santosh Kesiraju, Raghavendra Pappagari, Lucas Ondel, Lukáš Burget, Najim Dehak, Sanjeev Khudanpur, Jan Černockỳ, and Suryakanth V Gangashetty. Topic identification of spoken documents using unsupervised acoustic unit discovery. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5745–5749. IEEE, 2017.

[33] Matthew Wiesner, Chunxi Liu, Lucas Ondel, Craig Harman, Vimal Manohar, Jan Trmal, Zhongqiang Huang, Najim Dehak, and Sanjeev Khudanpur. Automatic speech recognition and topic identification from speech for almost-zero-resource languages. *Proc. Interspeech 2018*, pages 2052–2056, 2018.

[34] Kate M Knill, Mark JF Gales, Anton Ragni, and Shakti P Rath. Language independent and unsupervised acoustic models for speech recognition and keyword spotting. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[35] Lucas Ondel, Lukáš Burget, and Jan Černocký. Variational inference for acoustic unit discovery. *Procedia Computer Science*, 81:80–86, 2016.

[36] Chunxi Liu, Jinyi Yang, Ming Sun, Santosh Kesiraju, Alena Rott, Lucas Ondel, Pegah Ghahremani, Najim Dehak, Lukaš Burget, and Sanjeev Khudanpur. An empirical evaluation of zero resource acoustic unit discovery. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5305–5309. IEEE, 2017.

[37] Aren Jansen and Benjamin Van Durme. Efficient spoken term discovery using randomized algorithms. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 401–406. IEEE, 2011.

[38] Darshana Buddhika, Ranula Liyadipita, Sudeepa Nadeeshan, Hasini Witharana, Sanath Javasena, and Uthayasanker Thayasivam. Domain specific intent classification of sinhala speech data. In *2018 International Conference on Asian Language Processing (IALP)*, pages 197–202. IEEE, 2018.

[39] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[40] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, pages 498–502, 2017.

[41] Chia-ying Lee and James Glass. A nonparametric bayesian approach to acoustic model discovery. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 40–49. Association for Computational Linguistics, 2012.

[42] Yiren Wang and Fei Tian. Recurrent residual learning for sequence classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 938–943, 2016.

[43] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[44] Darshana Buddhika, Ranula Liyadipita, Sudeepa Nadeeshan, Hasini Witharana, Sanath Jayasena, and Uthayasanker Thayasivam. Voicer: A crowd sourcing tool for speech data collection. In *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 174–181. IEEE, 2018.

[45] James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8(1):014008, 2015.

[46] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

[47] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772, 2014.