Bibliography

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, J. Chen, J. Chen, Z. Chen, M. Chrzanowski, A. Coates, G. Diamos, K. Ding, N. Du, E. Elsen, and Z. Zhu, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *ArXiv*, 12 2015.

[2] L. Lugosch, M. Ravanelli, P. Ignoto, V. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *ArXiv*, vol. abs/1904.03670, 2019.

[3] Y.-A. Chung, H. Tang, and J. Glass, "Vector-quantized autoregressive predictive coding," in *INTERSPEECH*, 2020.

[4] A. H. Liu, Y.-A. Chung, and J. Glass, "Non-autoregressive predictive coding for learning speech representations from local dependencies," *ArXiv*, vol. abs/2011.00406, 2020.

[5] Y.-P. Chen, R. Price, and S. Bangalore, "Spoken language understanding without speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6189–6193.

[6] J. Poncelet and H. Van hamme, "Multitask learning with capsule networks for speech-to-intent applications," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 05 2020, pp. 8494–8498.

[7] Y. Karunanayake, U. Thayasivam, and S. Ranathunga, "Sinhala and tamil speech intent identification from english phoneme based asr," *2019 International Conference on Asian Language Processing (IALP)*, pp. 234–239, 2019.

[8] B. Zhen, X. Wu, Z. Liu, and H. Chi, "On the importance of components of the mfcc in speech and speaker recognition," in *INTERSPEECH*, vol. 37, 01 2000, pp. 487–490.

[9] V. Këpuska and H. Elharati, "Robust speech recognition system using conventional and hybrid features of mfcc, lpcc, plp, rasta-plp and hidden markov model classifier in noisy conditions," *Journal of Computer and Communications*, vol. 03, pp. 1–9, 01 2015.

[10] K. Aida-zade, C. Ardil, and S. Rustamov, "Investigation of combined use of mfcc and lpc features in speech recognition systems," *Signal Processing*, 01 2007.

[11] S. Shum, N. Dehak, R. Dehak, and J. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," *Odyssey*, 01 2010.

[12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 04 2018, pp. 5329–5333.

[13] Y. Shi, Q. Huang, and T. Hain, "H-vectors: Utterance-level speaker embedding using a hierarchical attention model," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 05 2020, pp. 7579–7583.

[14] A. Senior and I. Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 05 2014, pp. 225–229.

[15] C. yu, A. Ogawa, M. Delcroix, T. Yoshioka, T. Nakatani, and J. Hansen, "Robust i-vector extraction for neural network adaptation in noisy environment," in *INTERSPEECH*, 09 2015.

[16] S. Garimella, A. Mandal, N. Strom, B. Hoffmeister, S. Matsoukas, and S. H. K. Parthasarathi, "Robust i-vector based adaptation of dnn acoustic model for speech recognition," in *INTERSPEECH*, 2015.

[17] Y. Miao, H. Zhang, and F. Metze, "Towards speaker adaptive training of deep neural network acoustic models," in *INTERSPEECH*, 2014.

[18] X. Cui, V. Goel, and G. Saon, "Embedding-based speaker adaptive training of deep neural networks," in *INTERSPEECH*, 08 2017, pp. 122–126.

[19] N. Tomashenko, A. Caubrière, and Y. Estève, "Investigating Adaptation and Transfer Learning for End-to-End Spoken Language Understanding from Speech," in *Proc. Interspeech 2019*, 2019, pp. 824–828. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2158

[20] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. R. Glass, "An unsupervised autoregressive model for speech representation learning," *ArXiv*, vol. abs/1904.03240, 2019.

[21] Y.-A. Chung and J. R. Glass, "Generative pre-training for speech with autoregressive predictive coding," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3497–3501, 2020.

[22] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayan, "Paralinguistics in speech and language - state-of-the-art and the challenge," *Computer Speech and Language, Special Issue on Paralinguistics in Naturalistic Speech and Language*, 01 2013.

[23] B. Belean, "Comparison of formant detection methods used in speech processing applications," *AIP Conference Proceedings*, vol. 1565, pp. 85–89, 11 2013.

[24] J. P. Teixeira and A. Gonçalves, "Algorithm for jitter and shimmer measurement in pathologic voices," *Procedia Computer Science*, vol. 100, pp. 271 – 279, 2016.

[25] X. Li and X. Wu, "Modeling speaker variability using long short-term memory networks for speech recognition," in *INTERSPEECH*, 2015.

[26] Y. zhao, J. Li, X. Wang, and Y. Li, "The speechtransformer for large-scale mandarin chinese speech recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 05 2019, pp. 7095–7099.

[27] Z. Fan, J. Li, S. Zhou, and B. Xu, "Speaker-aware speech-transformer," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 222–229, 2019.

[28] J. Pan, D. Liu, G. Wan, J. Du, Q. Liu, and Z. Ye, "Online speaker adaptation for lvcsr based on attention mechanism," *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 183–186, 2018.

[29] W.-N. Hsu, Y. Zhang, and J. R. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *NIPS*, 2017.

[30] W.-N. Hsu and J. R. Glass, "Extracting domain invariant features by unsupervised learning for robust automatic speech recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5614–5618, 2018.

[31] S. Feng and T. Lee, "Improving unsupervised subword modeling via disentangled speech representation learning and transformation," in *INTERSPEECH*, 2019.

[32] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 2041–2053, 2019.

[33] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *INTERSPEECH*, 2019.

[34] A. T. Liu, S.-W. Li, and H. yi Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *ArXiv*, vol. abs/2007.06028, 2020.

[35] P.-H. Chi, P.-H. Chung, T. Wu, C.-C. Hsieh, S.-W. Li, and H. yi Lee, "Audio albert: A lite bert for self-supervised learning of audio representation," *ArXiv*, vol. abs/2005.08575, 2020.

[36] A. T. Liu, S. Yang, P.-H. Chi, P.-C. Hsu, and H. yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6419–6423, 2020.

[37] E. Morais, H. Kuo, S. Thomas, Z. Tuske, and B. Kingsbury, "End-to-end spoken language understanding using transformer networks and self-supervised pretrained features," *ArXiv*, vol. abs/2011.08238, 2020.

[38] E. Palogiannidi, I. Gkinis, G. Mastrapas, P. Mizera, and T. Stafylakis, "End-to-end architectures for asr-free spoken language understanding," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7974–7978, 2020.

[39] M. Radfar, A. Mouchtaris, and S. Kunzmann, "End-to-end neural transformer based spoken language understanding," in *INTERSPEECH*, 2020.

[40] Y. Karunanayake, U. Thayasivam, and S. Ranathunga, "Transfer learning based free-form speech command classification for low-resource languages," in *ACL*, 2019.

[41] A. Larcher, K.-A. Lee, and S. Meignier, "An extensible speaker identification sidekit in python," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5095–5099, 2016.

[42] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Ng, "Deepspeech: Scaling up end-to-end speech recognition," *ArXiv*, 12 2014.

[43] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC*, 2020.

[44] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[45] D. Buddhika, R. Liyadipita, S. Nadeeshan, H. Witharana, S. Jayasena, and U. Thayasivam, "Voicer: A crowd sourcing tool for speech data collection," in *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2018, pp. 174–181.