REFERENCES

[1] Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost. 2018. Improving End-of-turn Detection In Spoken Dialogues By Detecting Speaker Intentions As A Secondary Task. In *ICAASP*.

[2] Zakaria Aldeneh, Soheil Khorram, Dimitrios Dimitriadis, and Emily Mower Provost. 2017. Pooling acoustic and lexical features for the prediction of valence. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, ACM, New York, NY, USA, 68–72.

[3] Harish Arsikere, Elizabeth Shriberg, and Umut Ozertem. 2014. Computationally-efficient endpointing features for natural spoken interaction with personal-assistant systems. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 3241–3245.

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599* (2017).

[5] Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Prediction of Turn-taking Using Multitask Learning with Prediction of Backchannels and Fillers. In *INTERSPEECH*. To appear.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. IEEE Computer Society, Washington, DC, USA, 1026–1034.

[7] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2014. Analysis of Respiration for Prediction of "Who Will Be Next Speaker and When?" in Multi-Party Meetings. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14)*. ACM, New York, NY, USA, 18–25.

[8] Yuichi Ishimoto, Takehiro Teraoka, and Mika Enomoto. 2017. End-of-Utterance Prediction by Prosodic Features and Phrase-Dependency Structure in Spontaneous Japanese Speech. In *Proceedings of Interspeech 2017*. 1681–1685.

[9] Kristiina Jokinen, Kazuaki Harada, Masafumi Nishida, and Seiichi Yamamoto. 2010. Turn-alignment using eye-gaze and speech in conversational interaction. In *Eleventh Annual Conference of the International Speech Communication Association*.

[10] Tatsuya Kawahara, Takuma Iwatate, and Katsuya Takanashi. 2012. Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations. In *Thirteenth Annual Conference of the International Speech Communication Association*.

[11] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[12] Chaoran Liu, Carlos Ishi, and Hiroshi Ishiguro. 2017. Turn-Taking Estimation Model Based on Joint Embedding of Lexical and Prosodic Contents. In *Proc. Interspeech 2017*. 1686–1690.

[13] Angelika Maier, Julian Hough, and David Schlangen. 2017. Towards Deep End-of-Turn Prediction for Situated Spoken Dialogue Systems. In *Proceedings of INTERSPEECH 2017*.

[14] Ryo Masumura, Taichi Asami, Hirokazu Masataki, Ryo Ishii, and Ryuichiro Higashinaka. 2017. Online End-of-Turn Detection from Speech based on Stacked Time-Asynchronous Sequential Networks. In *Proc. Interspeech 2017*. 1661–1665.

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[16] Antoine Raux and Maxine Eskenazi. 2009. A finite-state turn-taking model for spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 629–637.

[17] Antoine Raux and Maxine Eskenazi. 2012. Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing (TSLP)* 9, 1 (2012), 1.

[18] Emanuel A. Schegloff. 2006. Interaction: The infrastructure for social institutions, the natural ecological niche for language, and the arena in which culture is enacted. In *Roots of Human Sociality*, Nick J. Enfield and Stephen C. Levinson (Eds.). Berg, London, 70–96.

[19] Gabriel Skantze. 2017. Towards a General, Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 220–230.

[20] Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106, 26 (2009), 10587–10592.

[21] L. ten Bosch, N. Oostdijk, and Jan de Ruiter. 2004. Turn-taking in social talk dialogues: temporal, formal and functional aspects. In *SPECOM 2004*.

[22] Sho Ueno, Hirofumi Inaguma, Masato Mimura, and Tatsuya Kawahara. 2018. Acoustic-to-word attention-based model complemented with character-level CTC-based model. In *Proceedings of IEEE-ICASSP*. IEEE, 5804–5808.

[23] Nigel G. Ward and David DeVault. 2017. Challenges in Building Highly-Interactive Dialog Systems. *AI Magazine* 37, 4 (2017), 7–18.