TABLE II
PERFORMANCE OF THE POLYPHONY CLASSIFICATION NETWORK FOR DIFFERENT FEATURE SETS, IN TERMS OF ACCURACY.

Features	Notation	Accuracy (%)
logFBE	$[\mathbf{X}^{(\mathbf{s}_0)}]$	95.59
GCC	$[\mathbf{GCC}^{(\mathbf{s}_0,\mathbf{s}_1)};\mathbf{GCC}^{(\mathbf{s}_0,\mathbf{s}_2)}]$	98.68
logFBE + GCC	$[\mathbf{X}^{(\mathbf{s}_0)};\mathbf{GCC}^{(\mathbf{s}_0,\mathbf{s}_1)};\mathbf{GCC}^{(\mathbf{s}_0,\mathbf{s}_2)}]$	**99.27**

to boost the overall performance. Nevertheless, the fusion schemes still achieve improvements over the baseline.

Table II provides the polyphony level classification accuracy of the proposed polyphony network for various choices of feature sets. We observe that while all feature sets achieve good performance, the best option is to combine the logFBE features with the GCC-based ones, leading to 99.27% classification accuracy. With such performance, it is guaranteed that our pipeline will be applied almost only on overlapped instances during testing.

IV. CONCLUSIONS

In this paper, we examined the combination of sound source separation with overlapped sound event classification in a multi-channel setup with a large variety of event classes. Our results showcase the potential of incorporating separation methods in SEC systems, albeit high reverberation scenarios can be a limiting factor for the performance of the proposed pipeline. In future work, we plan to explore scenarios with polyphony of higher degree (≥ 3 simultaneous events). Also we will investigate the perspective of sound separation via a distributed microphone network, which could potentially further improve the separation quality.

REFERENCES

[1] S. Krstulovic, "Audio event recognition in the smart home," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. Plumbley, and D. Ellis, Eds., pp. 335–371. Springer, 2018.

[2] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos, "Multi-microphone fusion for detection of speech and acoustic events in smart spaces," in *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 2375–2379.

[3] D. Hollosi, J. Schröder, S. Goetze, and J.-E. Appell, "Voice activity detection driven acoustic event classification for monitoring in smart homes," in *Proc. 3rd International Symposium in Applied Sciences in Biomedical and Communication Technologies (ISABEL)*, 2010, pp. 1–5.

[4] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," *ACM Trans. Multimedia Computing, Communications, and Applications*, vol. 4, no. 2, pp. 1–23, 2008.

[5] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: a systematic review," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, pp. 1–46, 2016.

[6] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6460–6464.

[7] *DCASE: Detection and classification of acoustic scenes and events*, http://dcase.community/.

[8] I. Choi, K. Kwon, S. H. Bae, and N. S. Kim, "DNN-based sound event detection with exemplar-based approach for noise reduction," in *Proc. Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE)*, 2016, pp. 16–19.

[9] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *Signal Processing Letters (SPL)*, vol. 24, no. 3, pp. 279–283, 2017.

[10] E. Çakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 5, no. 6, pp. 1291–1303, 2017.

[11] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution augmented transformer for semi-supervised sound event detection," in *Proc. Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE), Tech. Rep.*, 2020.

[12] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. International Conference on Spoken Language Processing (Interspeech)*, 2016, pp. 545–549.

[13] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," *arXiv preprint arXiv:1910.06379*, 2019.

[14] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *arXiv preprint arXiv:2002.08933*, 2020.

[15] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," *arXiv preprint arXiv:2003.01531*, 2020.

[16] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. Ellis, "Improving universal sound separation using sound classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 96–100.

[17] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, "Universal sound separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 175–179.

[18] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8677–8681.

[19] N. Turpault, S. Wisdom, H. Erdogan, J. Hershey, R. Serizel, E. Fonseca, P. Seetharaman, and J. Salamon, "Improving sound event detection in domestic environments using sound separation," *arXiv preprint arXiv:2007.03932*, 2020.

[20] W. Xue, T. Ying, Z. Chao, and D. Guohong, "Multi-beam and multi-task learning for joint sound event detection and localization," in *Proc. Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE), Tech. Rep.*, 2019.

[21] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM Int. Conference Multimedia*, 2015, pp. 1015–1018.

[22] L. Christoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmüller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. Int. Conference Language Resources and Evaluation (LREC)*, 2014, pp. 2629–2634.

[23] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," *arXiv preprint arXiv:1910.14104*, 2020.

[24] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S. C. Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019, pp. 260–267.

[25] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of CRNN models," in *Proc. Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE), Tech. Rep.*, 2019.

[26] F.-R. Stöter, S. Chakrabarty, B. Edler, and E. A. P. Habets, "CountNet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 268–282, 2019.

[27] Y. Cao, T. Iqbal, Q. Kong, M. B. Galindo, W. Wang, and M. D. Plumbley, "Two-stage sound event localization and detection using intensity vector and generalized cross-correlation," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Tech. Rep.*, 2019.

[28] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised filterbank learning using convolutional restricted Boltzmann machine for environmental sound classification," in *Proc. International Conference on Spoken Language Processing (Interspeech)*, 2017, pp. 3107–3111.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "WHAMR!: Noisy and reverberant single-channel speech separation.," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 696–700.