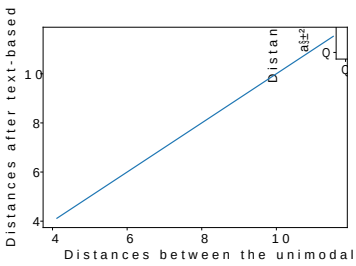
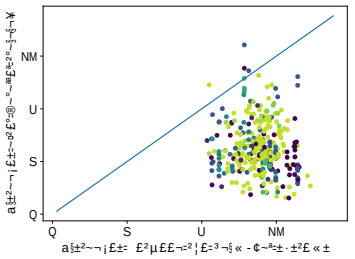


(a) Speech after text



(b) Speech after parallel

Fig. 3: Distance comparison between text and speech embeddings before and after multiview training (each point represents an utterance)

sufficient amount of speech data, the unimodal speech-only training achieves above 60% accuracy. Our observations from the first experiments still hold for this case, i.e. text-based pretraining of the classifier and then learning the speech encoder ("Speech after text") helps improving the performance and in the very low-resource case (less than 30 hours), additional fine-tuning (Speech after parallel) with the parallel data helps further increasing the accuracy.

In Table 2, we report the true text and ASR-text based testing of the multiview model for the second set of experiments performed on ASR-based speech embeddings. In terms of the results, the major difference between the previous experiment and the current one is that here, speech-only results approach to the true text based performance and they are significantly better than ASR-text based testing. This shows that we can achieve better performance than an ASR+NLP system with speech-only training when ASR-based speech embeddings are used.

When we compare "NP speech" and "Speech after parallel" setups, for the low-resource case, we get between 5-20% relative improvement in accuracy after fine-tuning with parallel data. The largest gain is observed when we have 10 hours of non-parallel speech data (0.516 to 0.620). Although the relative improvements are not as large as the first experiments, the absolute accuracies are much better in this case. If we compare ASR text-based testing to the speech-only testing case, we achieve about 15% improvement in accuracy (roughly from 0.55 to 0.63).

5. CONCLUSIONS

In this work, we have proposed a technique to train SLU task with non-parallel speech and text data, using speech-only dialog act recognition as an example. We showed how classification accuracy can be improved using non-parallel data. To handle the lack of parallel data, we proposed a multiview approach that consists of two branches each of which contains an encoder and a classifier. By sharing the classifier between two branches, we constrain the encodings from text and speech to be similar. One of the main advantages of this architecture is that it allows testing the system in a unimodal fashion. In our experiments on the SWDA corpus, we showed that text-based pretraining of the classifier in the multiview system helps improving speech-only classification accuracy (up to 32%) and also that additional fine-tuning on parallel data helps further (up to 40%) in the cases where we have less than 30 hours of speech. Since the text branch uses BERT features from a pretrained model, we also experimented with the case where the speech features come from a pretrained ASR model. In these experiments, speech accuracy approached to that of the text and also performed significantly better than ASR-text based testing of the text branch (up to 15%).

Dialog act classification is highly context dependent, as similar words can imply different acts depending on the history of the conversation. Therefore, one future goal is to incorporate context information into training. Another direction could be to investigate other multiview learning techniques such as deep canonical correlation analysis in this framework.

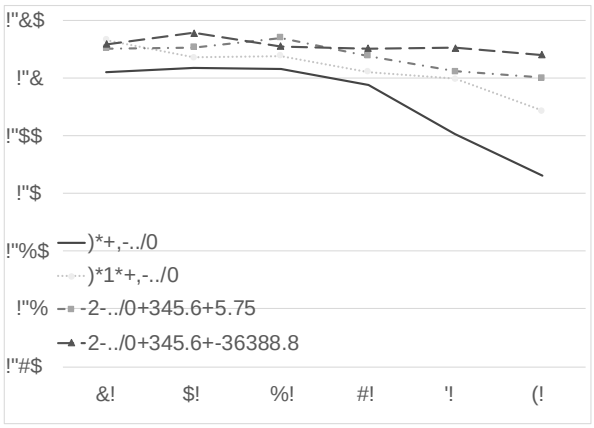


Fig. 4: Classification accuracy versus the amount of non-parallel (NP) speech data when inputs are ASR based embeddings

Table 2: Amount of non-parallel data (hr) to pretrain the branches and the accuracy of the text-only, speech-only and ASR-text based testing of the multiview model for the ASR embedding-based setup

Training condition (in hr)			Test Accuracy		
Text	Speech	Parallel	Text	Speech	ASR-text
60	60	14.5	0.677	0.630	0.549
70	50	14.5	0.688	0.640	0.549
80	40	14.5	0.672	0.628	0.535
90	30	14.5	0.681	0.626	0.558
100	20	14.5	0.682	0.627	0.556
110	10	14.5	0.672	0.620	0.543

6. ACKNOWLEDGMENT

The first and third authors are supported by the National Science Foundation under Grant No. NSF IIS 19-10319. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The second author was funded by IBM, and computational resources for this research were provided by the IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM AI Horizons Network.

7. REFERENCES

[1] Christian Raymond and Giuseppe Riccardi, "Generative and discriminative algorithms for spoken language understanding," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[2] Patrick Haffner, Gokhan Tur, and Jerry H Wright, "Optimizing svms for complex call classification," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*. IEEE, 2003, vol. 1, pp. I–I.

[3] Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélene Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, et al., "Breaking the unwritten language barrier: The bulb project," *Procedia Computer Science*, vol. 81, pp. 8–14, 2016.

[4] Odette Scharenborg, Francesco Ciannella, Shruti Palaskar, Alan Black, Florian Metze, Lucas Ondel, and Mark Hasegawa-Johnson, "Building an asr system for a low-resource language through the adaptation of a high-resource language asr system: Preliminary results," *Proceedings of ICNLSSP, Casablanca, Morocco*, 2017.

[5] 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)*. IEEE, 2018, pp. 5754–5758.

[6] 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)*. IEEE, 2018, pp. 720–726.

[7] Antoine Caubrière, Natalia Tomashenko, Antoine Laurent, Emmanuel Morin, Nathalie Camelin, and Yannick Estève, "Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability," *arXiv preprint arXiv:1906.07601*, 2019.

[8] 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.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] J.R. Searle, *Expression and meaning: Studies in the theory of speech acts*, Cambridge University Press, 1979.

[11] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[12] Dilek Hakkani-Tür, Gökhan Tür, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang, "Multi-domain joint semantic frame parsing using bi-directional rnn-lstm.," in *Interspeech*, 2016, pp. 715–719.

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

[14] Qian Chen, Zhu Zhuo, and Wen Wang, "Bert for joint intent classification and slot filling," *arXiv preprint arXiv:1902.10909*, 2019.

[15] Yao Qian, Rutuja Ubale, Vikram Ramanaryanan, Patrick Lange, David Suendermann-Oeft, Keelan Evanini, and Eugene Tsuprun, "Exploring asr-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 569–576.

[16] Yuan-Ping Chen, Ryan Price, and Srinivas Bangalore, "Spoken language understanding without speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6189–6193.

[17] Shiliang Sun, "A survey of multi-view machine learning," *Neural computing and applications*, vol. 23, no. 7-8, pp. 2031–2038, 2013.

[18] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*, 2013, pp. 1247–1255.

[19] Ryo Masumura, Mana Ihori, Tomohiro Tanaka, Atsushi Ando, Ryo Ishii, Takanobu Oba, and Ryuichiro Higashinaka, "Improving speech-based end-of-turn detection via cross-modal representation learning with punctuated text data," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019.

[20] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[21] Leda Sarı, Mark Allan Hasegawa-Johnson, S Kumaran, Georg Stemmer, and Krishnakumar N Nair, "Speaker adaptive audio-visual fusion for the open-vocabulary section of avicar," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, vol. 2018, pp. 3524–3528.

[22] Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca, "Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13," Tech. Rep. 97-02, University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO, 1997.

[23] Elizabeth Shriberg, Rebecca Bates, Paul Taylor, Andreas Stolcke, Daniel Jurafsky, Klaus Ries, Noah Coccaro, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech?," *Language and Speech*, vol. 41, no. 3–4, pp. 439–487, 1998.

[24] Vipul Raheja and Joel Tetreault, "Dialogue act classification with context-aware self-attention," *arXiv preprint arXiv:1904.02594*, 2019.

[25] Kartik Audhkhasi, George Saon, Zoltán Tüske, Brian Kingsbury, and Michael Picheny, "Forget a bit to learn better: Soft forgetting for ctc-based automatic speech recognition," *Proc. Interspeech 2019*, pp. 2618–2622, 2019.