TRAINING SPOKEN LANGUAGE UNDERSTANDING SYSTEMS WITH NON-PARALLEL SPEECH AND TEXT

Leda Sarı[1], *Samuel Thomas*[2], *Mark Hasegawa-Johnson*[1]

[1]Department of Electrical and Computer Engineering,
University of Illinois at Urbana-Champaign, IL, US
[2]IBM Research AI, NY, US

ABSTRACT

End-to-end spoken language understanding (SLU) systems are typically trained on large amounts of data. In many practical scenarios, the amount of labeled speech is often limited as opposed to text. In this study, we investigate the use of non-parallel speech and text to improve the performance of dialog act recognition as an example SLU task. We propose a multiview architecture that can handle each modality separately. To effectively train on such data, this model enforces the internal speech and text encodings to be similar using a shared classifier. On the Switchboard Dialog Act corpus, we show that pretraining the classifier using large amounts of text helps learning better speech encodings, resulting in up to 40% relatively higher classification accuracies. We also show that when the speech embeddings from an automatic speech recognition (ASR) system are used in this framework, the speech-only accuracy exceeds the performance of ASR-text based tests up to 15% relative and approaches the performance of using true transcripts.

***Index Terms*—** Dialog act recognition, spoken language understanding, multiview training, non-parallel data

1. INTRODUCTION

Speech understanding is a major component of human-machine interactions and its quality affects the user experience. Conventional speech understanding systems rely on a two-step approach where the speech signals are converted into text using an automatic speech recognition (ASR) system and then a natural language processing (NLP) system is applied to understand intents, to fill the slots or to detect named entities [1, 2]. However, this two-step approach suffers from error-propagation due to imperfect ASR systems and also from non-optimality as ASR and NLP systems are trained separately with different objectives. Moreover, for many of the world languages, there is not sufficient data to train reliable ASR sytems [3, 4]. Therefore, there is an interest in approaches which can directly use speech input to achieve the understanding task without using intermediate ASR transcripts [5, 6, 7, 8].

Given that the variability in speech signals is larger than that of the text inputs, and also the fact that recent text-based embeddings such as BERT [9] achieve state-of the art performance in NLP tasks, the performance of text based SLU systems are usually better than corresponding speech based systems. To improve the performance of speech-only based systems, it would therefore be useful to utilize the complementary information present in text based representations. For many training approaches, although parallel speech and text data would be required to integrate such information, with multiview based techniques we can train systems with non-parallel data. Systems trained in this fashion also have an advantage of being able to use any one of the two modalities at test time.

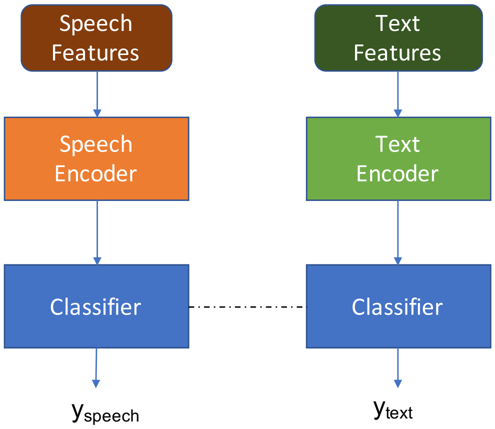


Fig. 1: Multiview training for SLU systems

In this work, we focus on two goals for learning SLU systems with non-parallel data using speech-only dialog act recognition as an example task. First, we propose a multimodal (speech and text) approach for dialog act recognition based on a multiview training approach. In many practical scenarios, we have large amounts of external text data but limited amounts of parallel data with the corresponding text and speech for dialog act recognition. Therefore, our first goal is to show how we can improve speech-only performance by incorporating text information during training, especially in the non-parallel text case in the multiview approach. This problem is tackled by using a multiview system shown in Fig. 1 where we try to tie the speech and text encodings using a shared classifier. Second, if we are given an ASR system during training time, we try to identify the best way of utilizing information in the ASR model to train a speech-based dialog act recognition system.

Rest of the paper is organized as follows: We first review prior work and contrast them with our contributions in this work. In Section 3, we give an overview of the proposed multiview system. Next, we describe the experimental setups in detail and present our results. In Section 5, we conclude the paper by summarizing this work.

2. RELATION TO PRIOR WORK

Dialog act recognition is a form of utterance classification in which each utterance is an action [10], and the label encodes the type of action, e.g. acceptance, appreciation, open-question, negative answer, thanking, etc. Early systems for dialog act recognition usually extract lexical, prosodic, or word n-gram features, and use statistical modeling techniques such as hidden Markov models [11] to classify the features. Alternatively, CRFs [1] or SVMs [2] are used to classify the representations obtained from ASR output lattices. Recently, neural network based approaches have also been used for this task. Most of these works only focus on classification of the text rather than the corresponding speech signal. Our goal in this work, however, is to perform recognition directly on speech rather than text. There are also NLP studies that focus on text-based SLU which are usually based on classifying word representations. For example, in [12], 1-hot vectors or embeddings such as word2vec [13] are used for SLU. Recently, more powerful embeddings such as BERT embeddings [9] are used for joint intent classification and slot-filling [14].

End-to-end (E2E) approaches for spoken language understanding (SLU) include [5, 6, 7, 8]. Most of these approach require large amounts of labeled speech data to achieve good performance. In [5], authors attempt to predict intent labels directly from log-mel features. Although the speech-only accuracy is lower than a cascaded ASR+SLU system performance, the ASR+SLU degrades when tested with ASR based text. In [15], the authors aim at finding compact speech representations instead of using acoustic features directly to improve speech-only SLU. An encoder-decoder framework is used in [6], where the decoder is conditioned on the audio transcript. The authors conclude that having an intermediate text representation performs better than simply classifying acoustic features without any constraint. In our experiments, we also make similar observations and therefore use a text-based classifier pretraining to guide a subsequent speech-only training.

None of the mentioned studies tackle the problem of having non-parallel text and speech. In [7], an E2E approach for slot filling is introduced; because of data scarcity, the authors pre-train the system as an ASR system, then adapt it to the tasks of named entity recognition and slot-filling using small parallel corpora. Another transfer learning approach is used in [8] where the authors first train a word recognizer and use it as a feature extractor or fine-tune those layers on the slot-filling task. Although the word recognizer and the SLU classifier can be trained on different datasets, the recognition system still requires large amounts of parallel speech and text. In our ASR embedding based experiments, we use a similar idea but our features are extracted at an earlier layer rather than the pre-softmax layer. In [16], a cascaded approach is used where grapheme posteriors are generated from speech features and then the posterior features are classified. Similar to the previous method, although the graphemic part can be separately trained on an ASR corpus, and the SLU part on a text based dataset, this model still requires large amounts of parallel data.

As mentioned above, the primary contribution of the proposed speech-to-dialog act detection system is the handling of non-parallel speech and text data for training using a multiview architecture. A brief review of various multiview training approaches is presented in [17]. A more recent method for multiview training, namely deep canonical correlation analysis, is proposed in [18]. In [19], a shared decoder is used for multiview learning. However, to our knowledge, this work is the first study that uses multiview learning for SLU.

3. MULTIVIEW TRAINING

As observed in earlier studies, achieving good performance in a speech-only E2E SLU is difficult especially with limited amounts of data. It has also been shown that multimodal approaches usually improve results as compared to unimodal systems [20, 21]. When labeled text data is available for a task, we therefore hypothesize that it will be useful to improve the speech-based system.

One direct way of utilizing two modalities is to append the features in the system either at the input or at an intermediate level. However, training such a system requires parallel data corresponding to the same sample both at training and test time but especially during test time, we do not have access to text data for speech-based dialog act recognition and it is therefore not usable if, during test time, we do not have access to text.

To handle the non-parallel data case, we propose a multiview learning technique which consists of two unimodal branches which are coupled. The unimodal systems take either text or speech as input and produce dialog act labels. They consist of an encoder and a classifier. In this work, we used BERT [9] embeddings upon their recent success as text features and MFCCs or ASR-derived acoustic embeddings as speech features to the unimodal systems.

Our proposed model shown in Fig. 1, processes speech and text information separately using two branches. We try to force the learned embeddings to be similar by using a shared classifier on both branches. The system can be thought of as an inverted Siamese network because of the shared classification parts in the two branches. This structure allows us to partially train the model using one modality without parallel speech and text data. This model is also practical as it allows to use speech data during test time.

Training of the multiview model is summarized in Algorithm 1. We start with training the encoder and classifier with the rich-resource (text) modality. Then we freeze the classifier and train the encoder on the other modality (speech) and in the final step we fine-tune both branches using parallel data while still sharing the classifier between the two branches. If there is no parallel data available, we skip the fine-tuning step. At the end, we report the speech branch accuracy.

Algorithm 1 Training steps of the multiview system

```
Input: Labeled text-only, speech-only and parallel data
Output: Dialog act labels per utterance and overall accuracy
Train the text branch using text-only data
Freeze the classifier
Train the speech encoder with fixed classifier on speech-only data
if parallel data exists, then
    Fine-tune the encoders and the classifier on parallel data
end if
Test the speech branch alone
return Speech branch accuracy
```

4. EXPERIMENTS AND RESULTS

Experiments are performed on the Switchboard Dialog Act Corpus (SWDA) [22, 23]. The labels in the dataset are originally associated with text rather than speech. To use both speech and text modalities, we first create a matching speech corpus by finding the corresponding speech segments from the original Switchboard dataset based on forced alignments. We simulated the non-parallel setting by splitting the training data into text-only, speech-only and parallel portions where the amounts of total training, heldout and test sets are determined based on the division of [11].

In the first set of experiments, we used MFCCs with delta and double delta features as speech input. For text input, we extracted

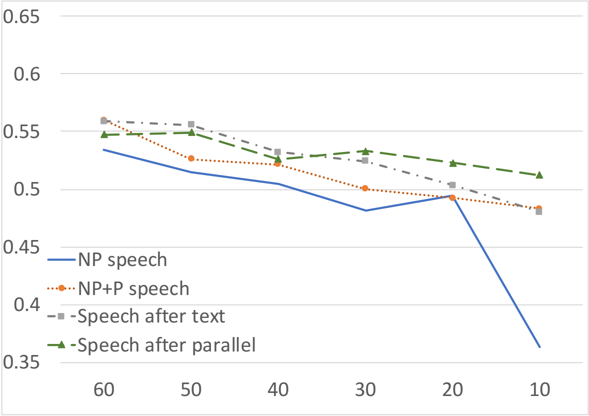


Fig. 2: Classification accuracy versus the amount of non-parallel (NP) speech data when inputs are MFCCs

Table 1: 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 MFCC-based setup

Training condition (in hr)			Test Accuracy		
Text	Speech	Parallel	Text	Speech	ASR-text
60	60	14.5	0.675	0.547	0.541
70	50	14.5	0.685	0.549	0.548
80	40	14.5	0.679	0.526	0.552
90	30	14.5	0.673	0.533	0.539
100	20	14.5	0.677	0.523	0.546
110	10	14.5	0.654	0.512	0.536

BERT embeddings [9] from a pretrained model on the true transcripts.

In multiview systems, the speech encoder consists of 3 bidirectional LSTM (BLSTM) layers each of size 128 followed by 2 fully-connected layers of size 64. The text encoder consists of 2 BLSTM layers each with 128 units followed by a single fully-connected layer. In both branches transition from the BLSTM layer to the fully-connected layers is achieved by averaging over time. The classifier has 3 fully-connected layers with rectified linear unit nonlinearity.

Fig. 2 compares the classification accuracies of four speech-only systems depending on the amount of non-parallel (NP) speech data used in training. The baseline is the case where we train the speech branch on low amounts of NP speech data (NP speech). Next, we combine the NP speech data with the speech portion of the parallel (P) data and train the speech branch on that set (NP+P speech). As the amount of data is larger in this situation, we achieve higher accuracy than the baseline. In multiview training, we first train the rich-resource text branch with NP data. We then freeze the classifier and train speech encoder on non-parallel speech data ("Speech after text", corresponds to the model at the end of Step 3 in Algorithm 1). As seen from the figure, pretraining the classifier on text and then learning the speech encoder on NP data performs better than training the speech model on NP+P data. We then fine-tune both text and speech branches using the limited amount of parallel text and speech ("Speech after parallel", corresponds to the model at the end of Step 5 in Algorithm 1). For the cases where we have more than 30 hours of speech, fine-tuning step does not bring any benefit. However, when we have less than 30 hours of speech, fine-tuning with parallel data improves the accuracy as compared to "Speech after text". In the fine-tuning stage we adjust both the encoders and the classifier whereas in "Speech after text", we only learn the encoder with classifier fixed.

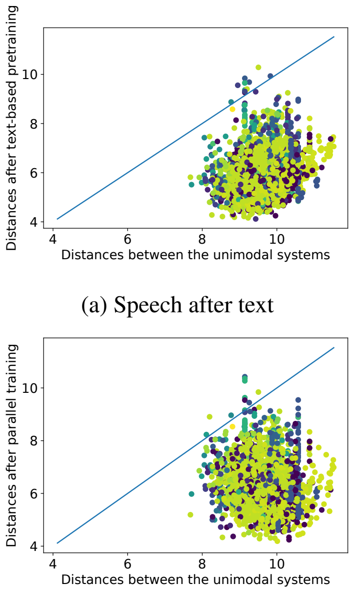
Since multiview system allows us to test the system using unimodal data, we also report the text-only performance of the systems. Table 1 shows the classification accuracy of both speech and text branch after all training steps. The speech accuracies in the table correspond to "Speech after parallel" curve in Fig. 2. Although training is performed on true transcript text, in practical scenarios we usually do not have the true text during test time but only ASR outputs. Therefore, we also show the results of testing the text branch with ASR-based text. We see how the mismatch between noisy and clean text affects the classification accuracy. We see that although true-text based testing gives above 65% performance, ASR-text based testing lowers the accuracy to that of the speech-only testing. Another disadvantage of ASR based testing is that it requires a language model in addition to an acoustic model whereas in the speech-only E2E classification, all we need is the acoustic features.

When we compare "NP speech" and "Speech after parallel" setups, for the low-resource case, we get between 5-40% relative improvement in accuracy after fine-tuning with parallel data. The gain reaches to 40% (0.363 to 0.512) when we have only 10 hours of non-parallel speech at the beginning.

For the conditions achieving 40% relative improvement, which is the 10 hours of non-parallel speech scenario, we plot the Euclidean distance between the text and speech embeddings to see if the proposed approach can tie them together using a shared classifier. If the hypothesis holds, then the distances after training should be smaller than the distance of the unimodal systems. As shown in Fig. 3, after applying either "Speech after text" or the "Speech after parallel" method, we get smaller distances between embeddings as they are mostly below the diagonal, e.g. the point (10, 4) implies that after multiview approach, the Euclidean distance between the average speech embedding and average text embedding for an utterance is reduced from 10 to 4. The comparison between both methods do not show a significant difference as shown in Fig. 2 implying that the main contribution comes from "Speech after text", i.e. pretraining the classifier with text rather than the fine-tuning stage.

These results confirm several hypotheses. First, simple acoustic features are harder to classify than text embeddings such as BERT. Second, although text-based system works well if tested on true text, in practice we do not have access to that information and hence need to resort to ASR-based noisy text which deteriorates the results to the level of speech-only testing. Third, having non-parallel text data can be used to guide learning speech encodings and it helps improving speech-only performance. Although we do not have the state of the art results on the text branch [24], we can still improve the speech-only performance in the proposed multiview architecture. Our speech-only performance on the other hand achieves better than the best speech-only system reported in [11], which is at 38.9%.

Another way of increasing the performance is to improve the speech features fed into the system. Note that text representations come from a pretrained BERT model however in the first set of experiments, speech features were MFCCs. Even though ASR text-based testing performs poorly, in the cases where we have access to a neural network based acoustic model, we can utilize it as a feature extractor. In the second set of experiments, we took an off-the-shelf ASR model trained on the Switchboard dataset [25], and extracted speech features from the LSTM output of that acoustic model. We then repeated the first set of experiments using these ASR-based speech features. As shown in Fig. 4, when we have



(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%).

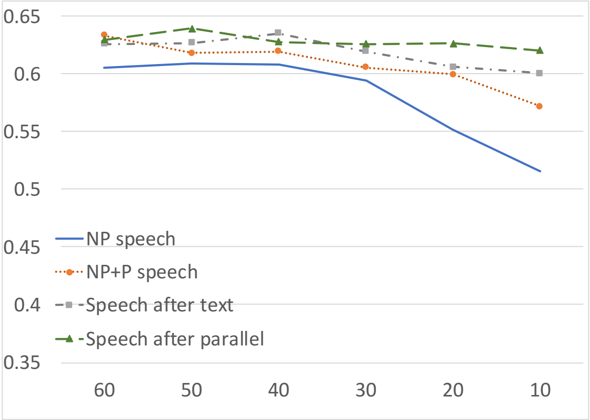


Fig. 4: Classification accuracy versus the amount of non-parallel (NP) speech data when inputs are ASR based embeddings

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

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.

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.