RASwDA: Re-Aligned Switchboard Dialog Act Corpus for Dialog Act Prediction in Conversations

Run Chen, Eleanor Lin, Shayan Hooshmand, Mariam Mustafa, Rose Sloan, Ritika Nandi, Alicia Yang, Andrea Lopez, Ansh Kothary, Isaac Suh, Catherine Lyu, Eric Chen, Sophia Horng and Julia Hirschberg

Abstract The Switchboard Dialog Act (SwDA) corpus has been widely used for dialog act prediction and generation tasks. However, due to misalignment between the text and speech data in this corpus, models incorporating prosodic information have shown poor performance. In this paper, we report the misalignment issues present in the SwDA corpus caused by previous automatic alignment methods and introduce a re-aligned, improved version called RASwDA (Re-Aligned Switchboard Dialog Act Corpus). Our goal is to create the largest publicly available two-speaker dialogue act corpus which has correctly aligned transcripts and speech. Through manual realignment and validation of 537.5 conversations completed so far, we have exceeded the state-of-the-art dialog act recognition results trained on SwDA. As we continue to expand RASwDA by re-aligning the remaining conversations from SwDA, we anticipate further improvements in model performance, facilitated by a larger and more accurate dataset.

1 Introduction

Dialog Act (DA) prediction and production is of seminal importance today in research, government and industry, as more and more dialogue systems are being built to interact with people for training, education, decreasing human workload in call centers, and providing problem-solving advice. While many corpora have been developed and annotated for building machine learning models in DA prediction or generation tasks, only a few have been transcribed in speech. Many others were

Run Chen
Columbia University, 1214 Amsterdam Ave, New York, e-mail: runchen@cs.columbia.edu

Eleanor Lin
Columbia University, 1214 Amsterdam Ave, New York, e-mail: eml2221@columbia.edu

Julia Hirschberg
Columbia University, 1214 Amsterdam Ave, New York, e-mail: julia@cs.columbia.edu

annotated using domain-specific DAs, or they are limited in the number or length of conversations. Among the annotated DA corpora, only the Switchboard Dialog Act (SwDA) Corpus [16] includes domain-independent spoken conversations between two speakers, making it unique for modeling the type of interactions that are primary in most systems used today, such as information services and online chats.

Although the SwDA corpus is widely used for DA prediction and generation tasks, it suffers from a critical limitation: inaccurate alignments. The corpus consists of transcripts and speech derived from the larger Switchboard corpus [13], which were originally aligned using a GMM-HMM speech recognition system. However, these alignment results are unreliable, making it extremely difficult to use both speech and text data to accurately predict or generate DAs. There is very little evidence that the use of speech features from the currently aligned corpus significantly improves their results in any way and sometimes even leads to worse performance [28, 21, 29, 30].

Previous attempts to re-segment the Switchboard corpus, upon which SwDA is built, have resulted in completely different transcriptions and utterance boundaries that do not coincide with those in the SwDA [11]. While the NXT-format Switchboard Corpus links the transcriptions in SwDA with the alignments from [11], it does so for only 642 of the 1,155 conversations in SwDA [8]. To date, no one has produced a full realignment of all 1,155 SwDA conversations.

Our project aims to create an improved, Re-Aligned Switchboard Dialog Act (RASwDA) corpus for DA tasks by manual re-alignment and validation by experts of both sides of all SwDA conversations to correct the errors introduced by the early automatic alignment. Our goal is to 1) produce a more accurate RASwDA corpus for DA prediction and generation tasks and 2) set a new benchmark for identifying DAs using machine learning models that incorporate both text and speech features. We demonstrate that this new version of the SwDA corpus provides more useful information in both text and speech for DA identification models by comparing the new results to models built on the earlier version of the corpus. To encourage the wider community to make use of the fully re-aligned corpus, we will make it publicly available, thereby facilitating the current research efforts focused on modeling human-human and human-machine conversation. [1]

2 Related Work

2.1 Dialogue Act Labeled Corpora

Many corpora, including SwDA, have been annotated for DAs. They vary in domains, languages, types of interactions, and the number and type of annotated DAs. While some corpora were annotated using small tag sets, such as the DCIEM

[1] Data will be available through Linguistic Data Consortium (LDC), which currently provides many earlier versions of this corpus.

Map Task [4], the AMI Meeting [9] (under 20), and the Columbia Games Corpus (only 7) [14], others were annotated using tag sets with hundreds of tags, such as DIHANA [5] and NESPOLE [10]. Furthermore, some corpora, including the DCIEM Map Task, SwDA, SCHISMA [18], ICSI-MRDA [25], and AMI Meeting, utilized domain-independent tag sets suitable for annotating various corpora. On the other hand, corpora such as VERBMOBIL [17], NESPOLE, DIHANA, LEGO [24], TourSG [19], Ubuntu IRC [20], MultiWOz and its multiple updated versions [7, 12, 31], and Audio Visual Scene-Aware Dialog (AVSD) [1] were annotated using domain-dependent tag sets. Notably, many corpora did not include speech data, such as DSTC6 corpora (Twitter, WOCHAT) [15], Ubuntu IRC, and MultiWOz.

Among these DA corpora, SwDA is particularly valuable for investigating how speech and transcripts synergize to facilitate DA modeling in conversations. The corpus contains a substantial number of annotated segments and provides both speech and transcripts with domain-independent data, distinguishing itself from others with a limited number of annotations, such as SCHISMA and DCIEM Map Task. Although ICSI-MRDA and the AMI Meeting corpus also offer sizable annotated speech data in multi-participant meetings, only SwDA exclusively comprises dialogues between two individuals, making it particularly relevant for modeling the types of two-party interactions prevalent in conversational systems today. However, the limitation of SwDA lies in its inaccurate alignment of speech and transcripts, which cannot be used to identify or generate appropriate acoustic-prosodic features, such as pitch, intensity, speaking rate, and voice quality.

2.2 *The Switchboard Dialog Act Corpus*

The original Switchboard Corpus is a corpus of 2,400 two-sided telephone conversations, each between two native speakers of American English from different parts of the United States, and was collected in 1990-91 by Texas Instruments. The initial goal for this corpus collection was to develop speech processing algorithms, particularly speaker verification algorithms [13]. The SwDA corpus [16] was created from a portion of the Switchboard corpus, specifically LDC's Switchboard-1 Release 2 (LDC97S62) [13]. It consists of 1,155 conversations out of the original 2400 conversations, ranging from 1.5 to 10 minutes, comprising a total of 205,000 utterances and 1.4 million words.

SwDA was labeled with an augmented version of the Discourse Annotation and Markup System of Labeling (DAMSL) tag-set [2], the SWBD-DAMSL label set of 43 DA labels. The DA labels include items such as *Statement-non-opinion*, *Acknowledge*, and *Statement-opinion*, which represent over two thirds of the 43 DA items annotated; the full list is shown in Table 1.

The SwDA conversations were initially force-aligned with the participants' speech in the 1990s using a GMM-HMM Switchboard recognition system to identify the start and end times of speech segments [26, p. 454]. However, due to the limited reliability of ASR systems used during that era and various challenges posed by the recordings and the transcripts, much of this alignment contained major errors, so it is impossible to perform accurate prosodic analysis on the DAs from their poor alignment with the audio.

Problems with this speech aligner included misalignment of reduced and low-energy speech. Based on manual inspection of hundreds of audio files, we have also found that background noise from sources such as static, telephones ringing, children crying, music, radios, and TV's has also reduced the original alignment quality. Problems with the conversations' transcripts at the time included mis-transcribed or simply missing words (some had been excised in a previous transcription task as "not useful words"). Only a small subset of these alignments were corrected to create a small DEV test set. The rest of the corpus was left in its original, poorly aligned state.

3 SwDA Alignment Diagnosis

While the SwDA corpus has been widely used to build models to detect different DAs, studies have observed that incorporating the audio information from SwDA does not improve DA prediction or generation scores, and can sometimes even worsen them. This is likely due to the poor alignment of the audio with transcripts and dialog act labels. [28, 21] showed that integrating prosodic information with transcripts improved DA prediction accuracy only for a couple of selected DAs, while having negative or no effects on the rest. The DA recognition model that incorporates prosody reported a lower F1 score, compared to the model trained solely on transcripts [29]. Similarly, [30] found that removing pitch and energy features resulted in only a marginal decrease in accuracy (1% and 0.6%, respectively) for their end-to-end DAC model on the SwDA corpus.

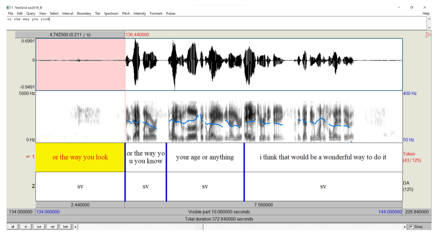
Transcripts and their aligned speech were often completely incorrect. We have found 27 conversations in which speakers were recorded on the wrong channel, resulting in incorrect speaker identifications when we attempt to match speaker audio with transcripts. Overlapping speech segments also cause confusion in the automatic alignment process. In many cases, shorter DAs such as *backchannel* or simple "yes" or "no" responses are missed entirely by the aligner. Furthermore, the presence of numerous simple timing errors in earlier parts of the conversations can propagate throughout the rest. These issues further highlight the challenges and limitations of DA modeling based on the SwDA corpus, underscoring the urgent need for its correction and improvement.

4 Re-alignment Methods

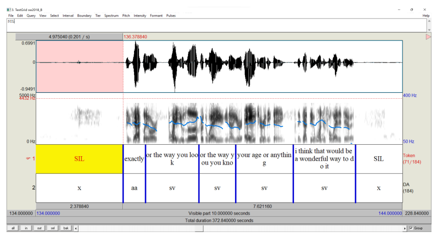
To produce high-quality alignments between the audio and transcripts of SwDA, we employ a two-step process. First, for conversations among the 642 conversations

DA	Description	Count (Full)	% (Full)	Count (RASwDA)	% (RASwDA)
sd	Statement-non-opinion	75145	34.26	32406	24.53
b	Acknowledge (Backchannel)	38298	17.46	16297	12.34
sv	Statement-opinion	26428	12.05	11762	8.90
%	Abandoned, Turn-Exit, or Uninterpretable	15550	7.09	6729	5.09
aa	Agree/Accept	11133	5.08	4973	3.76
x	Non-verbal	3630	1.65	3591	2.6
qy	Yes-No-Question	4727	2.15	2053	1.55
ba	Appreciation	4765	2.17	1799	1.36
ny	Yes answers	3034	1.38	1252	0.95
fc	Conventional-closing	2582	1.18	1056	0.80
qw	Wh-Question	1979	0.90	874	0.66
nn	No answers	1377	0.63	595	0.45
bk	Response Acknowledgement	1306	0.60	555	0.42
h	Hedge	1226	0.56	507	0.38
qy^d	Declarative Yes-No-Question	1219	0.56	472	0.36
bh	Backchannel in question form	1053	0.48	445	0.34
bf	Summarize/reformulate	952	0.43	444	0.34
^q	Quotation	983	0.45	427	0.32
fo_o_fw_"_by_bc	Other	883	0.40	408	0.31
na	Affirmative non-yes answers	847	0.39	351	0.27
qo	Open-Question	656	0.30	310	0.23
^2	Collaborative Completion	723	0.33	308	0.23
b^m	Repeat-phrase	688	0.31	283	0.21
ad	Action-directive	746	0.34	282	0.21
qh	Rhetorical-Questions	575	0.26	265	0.20
^h	Hold before answer/agreement	556	0.25	219	0.17
ar	Reject	346	0.16	141	0.11
ng	Negative non-no answers	302	0.14	137	0.10
br	Signal-non-understanding	298	0.14	137	0.10
no	Other answers	286	0.13	121	0.09
fp	Conventional-opening	225	0.10	117	0.09
qrr	Or-Clause	209	0.10	98	0.07
arp_nd	Dispreferred answers	207	0.09	91	0.07
^g	Tag-Question	92	0.04	53	0.04
oo_co_cc	Offers, Options, Commits	110	0.05	52	0.04
t1	Self-talk	103	0.05	44	0.03
bd	Downplayer	103	0.05	43	0.03
aap_am	Maybe/Accept-part	105	0.05	40	0.03
qw^d	Declarative Wh-Question	80	0.04	37	0.03
fa	Apology	79	0.04	34	0.03
t3	3rd-party-talk	117	0.05	32	0.02
ft	Thanking	78	0.04	28	0.02

Table 1: Comparison of the original SwDA DA counts ("Count (Full)") and our realigned corpus RASwDA DA counts ("Count (RASwDA)"). Original counts from [23].



(a) Automatic alignment.



(b) Automatic alignment + manual correction.

Fig. 1: A section of a SwDA transcript in the Praat interface (a) before and (b) after manual correction of the automatic alignment generated by *aeneas*. Praat allows aligners to view the waveform and spectrogram of the speech signal (top two sections of display) and a TextGrid transcript (bottom section of display) simultaneously.

which are included in the NXT-format Switchboard Corpus [8], we parse time-aligned SwDA transcripts from the NXT-provided XML files into TextGrid format. For conversations not included in the NXT-format Corpus, we parse each conversation's transcript into separate transcripts for each speaker. We also take advantage of the fact that speakers are recorded on separate channels to separate the audio for each conversation into two WAV files, one with each speaker's speech [27]. Then (for transcripts not sourced from NXT-format Switchboard) we compute the forced alignment for each utterance in each speaker transcript and conversation with the *aeneas* library [22], shown in Figure 1a. Based on manual inspection, we find that further manual realignment is still necessary to correct forced alignments generated with *aeneas*, as many of the issues that affected the original forced alignments (e.g. background noise) also affect the accuracy of the *aeneas* alignment.

Second, we manually correct the TextGrids produced both from the NXT-format Switchboard Corpus alignments and the *aeneas* forced alignments (Figure 1b). We use the Praat speech analysis interface, which allows expert aligners to easily manipulate audio and transcripts simultaneously [6]. Specifically, we convert each SwDA transcript into a TextGrid, a text file format commonly used for annotating audio in Praat.

In addition to correcting the transcript alignment, aligners are also instructed to mark speaker overlap and laughter with the special "SIL" and "⟨laughter⟩" tokens, and correct mis-transcriptions, segmentation errors, and omissions in the transcript. We attempted to resolve mis-transcriptions and segmentation errors marked by the original SwDA annotators themselves for correction at a later date [16]. Our aligners included 2 high school students, 15 undergraduates, and 8 graduate students in computer science, linguistics, and mathematics, some compensated for their time in either course credit or a stipend.

5 Results

Our Re-Aligned Switchboard Dialog Act (RASwDA) corpus currently consists of 537.5 manually realigned and validated conversations (1075 single speaker transcripts) from the 1155 SwDA conversations. Our final goal is to create a new, correctly aligned version of the entire SwDA corpus that is publicly available and to demonstrate the effect of adding correct acoustic-prosodic features for DA prediction.

Table 1 presents the counts of different DA tags in the original SwDA corpus as compared to our RASwDA. The original corpus consists of 203,801 dialog acts [23], while our realigned subset RASwDA contains 98,274 dialog acts and 42,231 silence segments.

6 DA Classification

By training dialog act classification (DAC) models on 55,049 utterances from RASwDA, we have achieved 59.53% accuracy on a 13,762-utterance validation set constructed from RASwDA, a 2.56% improvement over the 56.97% accuracy reported by [30] on a 4,088-utterance test set from the original SwDA corpus using their state-of-the-art end-to-end neural model trained on 192,768 utterances from the original SwDA corpus (Table 2).

Model	[30]	Ours
Dataset	SwDA	RASwDA
Accuracy	56.97	**59.53**
Train	192,768	55,049
Validation	3,196	13,762
Test	4,088	–

Table 2: Dialog act classification accuracy on speech from SwDA and RASwDA corpora, along with sizes of training, validation, and test splits in numbers of utterances.

Our model uses a convolutional neural network (CNN) and treats DAC as an image classification task on spectrograms of the speech signal, as this has proven a successful approach for applications such as emotion recognition [3]. The input to the CNN is a $256 \times 256 \times 3$ spectrogram of the speech signal, computed with matplotlib.[2] The CNN consists of three convolutional layers using 3×3 kernels, each followed by the application of the ReLU non-linearity, normalization, a max pooling layer with a 2×2 window, and another normalization. The first convolutional layer consists of 32 kernels with a stride of 2 pixels. The second convolutional layer consists of 64 kernels with a stride of 1. The third convolutional layer consists of 128 kernels with a stride of 1. After application of the ReLU non-linearity, normalization, and pooling, the output of the third convolutional layer is flattened into a 32768×1 vector and passed through a single fully connected layer to produce a 256-dimensional output. The output of the first fully connected layer is normalized and passed through a second and final fully connected layer. Finally, the softmax function is applied. We train on a 55,049-utterance subset of RASwDA and validate on a 13,762-utterance subset. We believe that as we continue to build RASwDA by realigning the rest of the SwDA conversations, the model performance will further improve with a larger, more accurate dataset.

7 Conclusions

We have identified inaccuracies in the current automatic alignments of the Switchboard Dialog Act (SwDA) corpus and have undertaken a manual realignment process for a subset of 537.5 out of 1155 conversations. Our Re-Aligned Switchboard Dialog Act (RASwDA) subset has already demonstrated improved performance of state-of-the-art models on the DA classification task. We plan to continue the realignment process for the remainder of the SwDA corpus and make it publicly available for the wider speech community.

Acknowledgements We express our gratitude to the students who contributed to this project by assisting with annotation and validation.

[2] `https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.specgram.html`

References

1. Alamri, H., Hori, C., Marks, T.K., Batra, D., Parikh, D.: Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7. In: DSTC7 at AAAI2019 Workshop, vol. 2 (2018)
2. Allen, J., Core, M.: Draft of damsl: Dialog act markup in several layers (1997). URL http://www.fb10.uni-bremen.de/anglistik/ling/ss07/discourse-materials/DAMSL97.pdf
3. Badshah, A.M., Ahmad, J., Rahim, N., Baik, S.W.: Speech emotion recognition from spectrograms with deep convolutional neural network. In: 2017 International Conference on Platform Technology and Service (PlatCon), pp. 1–5 (2017). DOI 10.1109/PlatCon.2017.7883728
4. Bard, E., Sotillo, C., Anderson, A., Taylor, M.: The dciem map task corpus: spontaneous dialogue under sleep deprivation and drug treatment. In: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, vol. 3, pp. 1958–1961 vol.3 (1996). DOI 10.1109/ICSLP.1996.608019
5. Benedí, J.M., Lleida, E., Varona, A., Castro, M.J., Galiano, I., Justo, R., López de Letona, I., Miguel, A.: Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). European Language Resources Association (ELRA), Genoa, Italy (2006). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/504_pdf.pdf
6. Boersma, P., Van Heuven, V.: Speak and unspeak with praat. Glot Int. **5**(9/10), 341–347 (2001)
7. Budzianowski, P., Wen, T.H., Tseng, B.H., Casanueva, I., Ultes, S., Ramadan, O., Gašić, M.: MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 5016–5026. Association for Computational Linguistics, Brussels, Belgium (2018). DOI 10.18653/v1/D18-1547. URL https://aclanthology.org/D18-1547
8. Calhoun, S., Carletta, J., Brenier, J.M., Mayo, N., Jurafsky, D., Steedman, M., Beaver, D.: The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Language resources and evaluation **44**, 387–419 (2010)
9. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al.: The ami meeting corpus: A pre-announcement. In: Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers 2, pp. 28–39. Springer (2006)
10. Costantini, E., Burger, S., Pianesi, F.: NESPOLE!'s multilingual and multimodal corpus. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02). European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain (2002). URL http://www.lrec-conf.org/proceedings/lrec2002/pdf/156.pdf
11. Deshmukh, N., Ganapathiraju, A., Gleeson, A., Hamaker, J., Picone, J.: Resegmentation of switchboard. In: ICSLP. Syndey (1998)
12. Eric, M., Goel, R., Paul, S., Sethi, A., Agarwal, S., Gao, S., Kumar, A., Goyal, A., Ku, P., Hakkani-Tur, D.: MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 422–428. European Language Resources Association, Marseille, France (2020). URL https://aclanthology.org/2020.lrec-1.53
13. Godfrey, J., Holliman, E., McDaniel, J.: Switchboard: telephone speech corpus for research and development. In: [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 517–520 vol.1 (1992). DOI 10.1109/ICASSP.1992.225858
14. Gravano, A., Hirschberg, J.: Backchannel-inviting cues in task-oriented dialogue. In: Tenth Annual Conference of the International Speech Communication Association (2009)
15. Hori, C., Perez, J., Higashinaka, R., Hori, T., Boureau, Y.L., Inaba, M., Tsunomori, Y., Takahashi, T., Yoshino, K., Kim, S.: Overview of the sixth dialog system technology challenge: Dstc6. Comput. Speech & Lang. **55**, 1–25 (2019)

16. Jurafsky, D., Shriberg, E., Biasca, D.: 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)
17. Kay, M., Gawron, M., Norvig, P.: Verbmobil: A Translation System for Face-to-Face Dialog. Center for the Study of Language and Information Publication Lecture Notes. Cambridge University Press (1994). URL https://books.google.com/books?id=3ezWxwEACAAJ
18. Keizer, S.: Dialogue act classification: Experiments with the schisma corpus. Tech. rep., Technical report, University of Twente (2002)
19. Kim, S., D'Haro, L.F., Banchs, R.E., Williams, J.D., Henderson, M.: The Fourth Dialog State Tracking Challenge, pp. 435–449. Springer Singapore, Singapore (2017). DOI 10.1007/978-981-10-2585-3_36. URL https://doi.org/10.1007/978-981-10-2585-3_36
20. Lowe, R., Pow, N., Serban, I., Pineau, J.: The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 285–294. Association for Computational Linguistics, Prague, Czech Republic (2015). DOI 10.18653/v1/W15-4640. URL https://aclanthology.org/W15-4640
21. Ortega, D., Thang Vu, N.: Lexico-acoustic neural-based models for dialog act classification. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6194–6198 (2018). DOI 10.1109/ICASSP.2018.8461371
22. Pettarin, A.: Aeneas. https://www.readbeyond.it/aeneas/ (2017)
23. Potts, C.: The switchboard dialog act corpus. https://compprag.christopherpotts.net/swda.html (2011)
24. Schmitt, A., Ultes, S., Minker, W.: A parameterized and annotated spoken dialog corpus of the CMU let's go bus information system. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pp. 3369–3373. European Language Resources Association (ELRA), Istanbul, Turkey (2012). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/333_Paper.pdf
25. Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., Carvey, H.: The ICSI meeting recorder dialog act (MRDA) corpus. In: Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004, pp. 97–100. Association for Computational Linguistics, Cambridge, Massachusetts, USA (2004). URL https://aclanthology.org/W04-2319
26. Shriberg, E., Stolcke, A., Jurafsky, D., Coccaro, N., Meteer, M., Bates, R., Taylor, P., Ries, K., Martin, R., Van Ess-Dykema, C.: Can prosody aid the automatic classification of dialog acts in conversational speech? Lang. and speech **41**(3-4), 443–492 (1998)
27. Sox - sound exchange. https://sox.sourceforge.net/ (2015)
28. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C.V., Meteer, M.: Dialogue act modeling for automatic tagging and recognition of conversational speech. Comput. linguistics **26**(3), 339–373 (2000)
29. Tran, T.: Neural models for integrating prosody in spoken language understanding. PhD thesis, University of Washington, Seattle, WA (2020)
30. Wei, K., Knox, D., Radfar, M., Tran, T., Müller, M., Strimel, G.P., Susanj, N., Mouchtaris, A., Omologo, M.: A neural prosody encoder for end-to-end dialogue act classification. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7047–7051. IEEE (2022)
31. Zang, X., Rastogi, A., Sunkara, S., Gupta, R., Zhang, J., Chen, J.: MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In: Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, pp. 109–117. Association for Computational Linguistics, Online (2020). DOI 10.18653/v1/2020.nlp4convai-1.13. URL https://aclanthology.org/2020.nlp4convai-1.13