Evaluating Automatic Speech Recognition Quality and Its Impact on Counselor Utterance Coding

Do June Min, Verónica Pérez-Rosas, Rada Mihalcea
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI, USA
dojmin@umich.edu, vrncapr@umich.edu, mihalcea@umich.edu

Abstract

Automatic speech recognition (ASR) is a crucial step in many natural language processing (NLP) applications, as often available data consists mainly of raw speech. Since the result of the ASR step is considered as a meaningful, informative input to later steps in the NLP pipeline, it is important to understand the behavior and failure mode of this step. In this work, we analyze the quality of ASR in the psychotherapy domain, using motivational interviewing conversations between therapists and clients. We conduct domain agnostic and domain-relevant evaluations using evaluation metrics and also identify domain-relevant keywords in the ASR output. Moreover, we empirically study the effect of mixing ASR and manual data during the training of a downstream NLP model, and also demonstrate how additional local context can help alleviate the error introduced by noisy ASR transcripts.

1 Introduction

Evaluating the quality of psychotherapy is an essential step in assessing the fidelity of treatment and providing feedback to practitioners. In psychotherapy practice, this is usually done through a process called behavioral coding that consists of manually analyzing recordings of therapy conversations and then labeling specific behaviors from participants.

Recent efforts have addressed the automatic analysis and evaluation of psychotherapy quality, including the study of conversational dynamics between therapists and clients, the analysis of empathy and emotional responses, and the automatic assessment of therapist's skills (Althoff et al., 2016; Zhang and Danescu-Niculescu-Mizil, 2020a; Pérez-Rosas et al., 2017).

Most of these research studies have been conducted using small collections of manually transcribed counseling conversations due to the need of an accurate representation of what is being said during the conversation. However, the use of manual transcription restricts the inclusion of a larger number of conversations into the analysis as it is a costly and slow process, making it challenging to apply data hungry machine learning approaches. As an alternative, some studies have explored the use of automatic speech recognition (ASR) systems that are able to quickly transcribe a large number of conversations (Flemotomos et al., 2021). However, there are several open questions regarding the feasibility of using automatic transcriptions in the evaluation of psychotherapy (Miner et al., 2020).

In this work, we study the quality of ASR in counseling conversations and its impact on the task of behavioral coding. We use an existing dataset of behavioral counseling conversations consisting of audio recordings and manual transcriptions as well as annotations of ten behaviors related to therapists' counseling skills. We start by generating automatic transcriptions using a commercially available ASR system (Google, 2020). Using the resulting parallel corpus of manual and ASR transcriptions, we conduct an assessment of the ASR quality using three main approaches. First, we use automatic evaluation metrics such as word error rate (WER) and semantic distance to conduct domain agnostic evaluations of the ASR performance across conversation participants. Second, we conduct a domain-specific examination of the ASR output by identifying domain-relevant keywords using behavioral codes and keywords identified using the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001). Finally, we study the effect of the noisy ASR on the downstream behavioral coding task and empirically show that additional local context in the form of neighboring utterances can help alleviate the impact of ASR errors.

We believe that studying the role of ASR systems in the NLP pipeline is an important step to develop and evaluate robust systems for better understanding of counseling dialogues.

2 Related Work

As the overall accuracy of ASR systems keeps improving, the ability of producing accurate transcriptions of conversational data has enabled the development of NLP applications in health. Particularly, in the psychotherapy domain, where a large fraction of therapy sessions are conducted in spoken language, ASR can help reduce the burden of manual transcription, potentially allowing for large-scale analysis of interactions between counselors and patients.

There have been several efforts on applying NLP on conversation analysis and utterance coding tasks in the psychotherapy domain. NLP was used to evaluate counselor behaviors and strategies (Zhang and Danescu-Niculescu-Mizil, 2020b; Pérez-Rosas et al., 2019; Xiao et al., 2015), or to provide feedback by generating appropriate responses to client utterances (Shen et al., 2020).

While most of previous work was conducted on manual transcriptions, there are only a few cases where automatically generated transcripts have been used, limiting the use of computational methods in psychiatry (Imel et al., 2015). The main reason behind this is the need for reliable ASR systems that are able to produce accurate transcriptions as the error introduced by transcribing words incorrectly can have a great impact on the performance of the overall application.

It has been pointed out by previous research that automatic evaluation metrics such as word error rate alone are not a good indicator of accuracy in speech understanding (Park et al., 2008). Our work is similar to Miner et al. (2020) recent work in that we use both agnostic and domain-relevant approaches to assess ASR systems in the mental health domain. However, we additionally investigate how the ASR error, both domain-agnostic and domain-relevant, propagates through the common NLP pipeline, in training and inference times, and provide an advice for researchers.

Finally, Mani et al. (2020) recently framed post-processing ASR error correction as a machine translation task from noisy transcription to ground truth transcription, and trains a sequence to sequence error correction model. Although this approach can provide a modular solution to mitigate ASR errors in many speech understanding systems, we note that building such a parallel corpus can be prohibitive for many researchers.

	Average	Std
Session Length		
Duration (min)	21.03	9.33
Length (words)	3320.02	1494.68
Words Spoken per Session (n)		
Therapist	2002.24	1024.63
Client	1317.77	858.25

Table 1: Session statistics

3 Dataset

3.1 Data Source

We evaluate utterances and behavioral codes from 213 counseling sessions compiled by Pérez-Rosas et al. (2016). The sessions were originally drawn from various sources, including two studies on smoking cessation and medication adherence. The full set comprises a total of 97.8 hours of audio with average session duration of 20.8 minutes. All the sessions were manually anonymized to remove identifiable information such as counselor and patient names and references to counseling sites' location. The sessions were transcribed using manual and crowd-sourced methods. The transcription set consist of 707,165 words distributed across 52,658 utterances and 39,637 talk-turns. More detailed statistics on words and utterances per session are provided in Table 1. The average conversation in the dataset has a duration of 21 minutes and a length of 3320 words.

The dataset also includes utterance-level annotations for ten behavioral codes from the Motivational Interviewing Treatment Integrity (MITI) coding scheme, the current gold standard for evaluating MI fidelity. MITI is focused on therapist language only and measures how well the therapist adhered to MI strategies by counting behaviors such as asking questions, using reflective language, seeking collaboration and emphasizing autonomy, among others. The dataset annotations were conducted by annotators with previous MI experience and trained on the use of MITI system. In addition to the MITI coding, our study uses two additional categories for utterances that are not labeled in the original dataset. The first includes therapist's speech that is not labeled under any MITI code (NAT) and the second includes client's utterances (NAC). Table 2 list the different behavioral codes, their count and their average word length.

Code	Count	Avg Len.
Question (QUEST)	6269	14.55
Simple reflection (SR)	2564	14.33
Complex reflection (CR)	3354	16.95
Seeking collaboration (SEEK)	927	20.34
Emphasizing autonomy (AUTO)	170	17.68
Affirm (AF)	550	17.71
Confront (CON)	139	12.97
Persuading without permission (PWOP)	1046	20.62
Persuading with permission (NPWP)	378	20.27
Giving Information (NGI)	1894	20.59
Non-coded Therapist (NAT)	12814	10.60
Non-coded Counselor (NAC)	22553	12.63

Table 2: Statistics for MITI behaviors coded in the dataset

3.2 Preprocessing

Alignment. Since the manual transcriptions provided in the dataset consist of transcribed speech without corresponding timestamps, we used forced alignment to automatically align speakers' speech with its corresponding transcription. We used *Gentle* (Ochshorn and Hawkins), a forced speech aligner implemented using the Kaldi toolkit for speech recognition (Povey et al., 2011). Note that this is a necessary step to enable comparisons between manual and automatic transcriptions for the same audio segments.

Automatic Transcription. To automatically transcribe each counseling session, we first spliced its audio into smaller segments using the obtained timestamps. Next, we individually transcribed each segment using the Google's Speech-to-Text recognition system (Google, 2020).[1] Again, our choice of transcribing segments rather than full conversations is motivated by the need of comparable units so we can avoid potential misalignment generated by ASR segmentation.

4 Domain-agnostic Evaluation

We start by conducting a domain-agnostic evaluation of the automatic transcription process that considers that the accuracy of the ASR system is equally important for all speech in the conversation. To this end, we focus on two automatic evaluation metrics: word error rate and semantic distance. The first one evaluates transcription error at the word-level; the second one aims to evaluate transcription error considering the semantic distance between the ASR output and the ground truth i.e., human transcription.

Word Error Rate (WER) We calculate WER using the equation below, where S, D, I each denote the number of substitutions, deletions, and insertions respectively required to make the reference sequence identical to the ASR sequence. C refers to the number of correct words, whereas N is the number of words in the reference.

$$WER = \frac{S+D+I}{S+D+C} = \frac{S+D+I}{N} \quad (1)$$

We use the Python Jiwer package[2] to automatically calculate WER for all conversations in the dataset. Our calculations are done by aggregating transcriptions by the corresponding speaker and averaging across sessions.

Semantic Distance. Although recent works show that averaging WERs over large benchmark sets can provide good estimation of model performance (Likhomanenko et al., 2020), there have been criticisms against relying solely on WERs, on the grounds that some important aspects of transcription quality are ignored when focusing on word overlaps (Kong et al., 2016; Szymański et al., 2020). For instance, "This is a cap" and "This is a cat" will have a low score of WER because of the low edit distance between the sentences, while their semantic contents are about two distant concepts (Kim et al., 2021). We use semantic distance to complement WER as semantics play an important role in understanding psychotherapy language and the meaning of a particular utterance could be greatly affected by substitutions done during the ASR process.

More specifically, we measure the difference in semantic content between the ground truth and ASR transcriptions. Our calculations are conducted at the utterance level and aggregated overall all conversations. We define the semantic distance between a manually transcribed utterance Utt_{MAN} and an automatically transcribed utterance Utt_{ASR} as the cosine distance between the sentence embeddings of each utterance:

$$\begin{aligned}&\text{Semantic Distance}(Utt_{MAN}, Utt_{ASR})\\ &= 1 - \frac{emb(Utt_{MAN}) \cdot emb(Utt_{ASR})}{\|emb(Utt_{MAN})\|\|emb(Utt_{ASR})\|} \quad (2)\end{aligned}$$

Thus, lower semantic distance between a manual transcription and an ASR transcription would indicate lower degree of transcription error.

[1] We use the Google Cloud speech-to-text enhanced model

[2] https://pypi.org/project/jiwer/

	n	WER	Semantic Distance
Aggregated	426	0.35± 0.09	0.28±0.06
Speaker Role			
Therapist	213	0.35±0.10	0.27±0.07
Client	213	0.40±0.16	0.30±0.10
Speaker Gender			
Female	344	0.35±0.09	0.27±0.06
Male	82	0.46±0.17	0.34±0.11
Therapist Gender			
Female	195	0.34±0.09	0.27±0.07
Male	18	0.40±0.11	0.31±0.05
Client Gender			
Female	149	0.37±0.14	0.28±0.08
Male	64	0.48±0.18	0.35±0.11

Table 3: WER and Semantic Distance statistics by speaker role and gender for manual and automatic transcriptions. Plus and minus values denote standard deviation.

Code	WER	Semantic Distance
AF	0.36±0.23	0.18±0.16
AUTO	0.34±0.29	0.18±0.17
CON	0.38±0.40	0.13±0.12
CR	0.32±0.14	0.18±0.16
NGI	0.33±0.27	0.16±0.15
NPWP	0.35±0.57	0.17±0.16
PWOP	0.29±0.14	0.15±0.14
QUEST	0.31±0.19	0.18±0.17
SEEK	0.32±0.43	0.17±0.15
SR	0.36±0.19	0.20±0.18
NAT	0.48±0.20	0.37±0.26
NAC	0.40±0.16	0.30±0.10

Table 4: WER and Semantic Distance statistics for ten MITI codes and non-annotated utterances in the dataset by therapists (NAT) and clients (NAC). Plus and minus values denote standard deviation.

For the $emb(\cdot)$ function we use sentence transformer embeddings (Reimers and Gurevych, 2019). We chose the sentence transformer over alternative methods of sentence embeddings such as BERT or word2vec, since recent research has shown that off-the-shelf transformer models without fine-tuning often lead to representations that perform poorly on semantic similarity tasks (Li et al., 2020).

4.1 Results

Table 3 summarizes the results obtained by speaker's role (i.e., therapist, client) and gender (i.e., male, female). Overall, transcription of therapist's speech shows significantly lower error than client speech in terms of WER, but not on semantic distance (two tailed Mann-Whitney U-test, $p < .05$). We also observe significant differences in female and male speech recognaition for both WER and semantic distance ($p < .05$, two tailed Mann-Whitney U-test). The difference between genders is also confirmed when the speaker roles are considered. This result is aligned with previous findings that ASR systems tend to perform better on female speakers due to being more consistent to standard pronunciations than male speakers (Adda-Decker and Lamel, 2005; Goldwater et al., 2008). However, it is important to mention that other work on ASR evaluation have encountered the opposite trend, where transcription of female speakers speech obtained higher WER than of males (Tatman, 2017). A factor that potentially affected our analysis is that due to the unavailability of identity data for speakers in the dataset, we treated each session as featuring a unique set of speakers. This might have been caused by the over-representation of speakers who appear multiple times in the dataset.

5 Domain-relevant Evaluation

Although the domain-agnostic evaluation can provide insights into the aggregate performance of an ASR system, a domain informed evaluation can help to better understand the quality of derived transcriptions and its potential impact on downstream tasks. In the counseling domain, incorrect transcription of words or phrases related to emotion, mental state, addiction, or medication can cause more harm than the incorrect transcription of other types of words. Seeking to evaluate the role of domain on ASR quality in our automatically transcribed conversations, we focus on speech that is relevant to counseling quality. To identify such speech, we use the behavioral coding provided in the dataset and also word categories from the Linguistic Inquiry and Word Count (LIWC) lexicon (Pennebaker et al., 2001).

Behavioral codes. We measure WER and semantic distance on utterances coded with the ten counselor behaviors included in the dataset and also examined transcription error in uncoded utterances from both, therapists and clients. For WER, we first concatenated all the utterances labeled with a given code in each single conversation, and then averaged the obtained WER across all conversations. Semantic distances for each utterance are averaged over all utterances in the dataset.

LIWC Categories. LIWC is a psycholinguistic lexicon that maps words and its stems to a set of categories related to psychological processes. There are 69 predefined categories that cover four high-level topics: psychological processes, personal concerns, linguistic dimensions, and linguistic fillers.

For our analysis, we identify and select a subset of categories from psychological processes and personal concerns as they have been found relevant to psychotherapy conversations. For words in the different categories appearing in the ground truth utterances, we evaluated whether the ASR system was able to correctly transcribed them. We calculate the true positive, false negative, and false positive rates as well the standard metrics of recall and precision.

5.1 Results

Table 4 shows the average WER and semantic distance of transcription for behavior codes and also for non-coded ("Non-coded Client", "Non-coded Therpapist") language in the conversations.

In general, we find that non-coded language tends to have higher transcription error than coded-language (two-tailed Mann-Whitney U-test, $p < 0.05$ for both WER and semantic distance). Within non-applicable codes, we note that NAT shows higher WER and semantic distance. Since in Table 3 we saw that client language tends to have higher error overall than therapist language, this may indicate that transcription error is correlated to speech content or topic, because NAC covers all client utterances, while NAT is only applied for non-MITI labeled utterances.

When the ASR system is evaluated in terms of transcribing keywords that are relevant to psychotherapy and counseling, results from Table 5 indicate that correctly retrieving keywords is harder for ASR systems than avoiding incorrect insertion of keywords in the transcription, as precision values are concentrated near 1.0, while recall values are more diverse. Table 6 gives an example of how omission errors can change the semantic content of the utterance for LIWC categories such as "DEATH, BODY". In the context of mental health and psychotherapy, these results suggest that aggregate metrics that compare whole ground truth utterances and ASR transcriptions to compute error rate are not granular enough to capture such cases of ASR failure where mistrancriptions of keywords might result in clinicians or counselors missing signs of patient distress or danger.

6 The Role of ASR on the Automatic Evaluation of Psychotherapy

Beyond studying the domain-agnostic and domain-relevant error patterns of the automatic transcription, we also study the relationship between the speech transcription step and the later behavior code classification, where ASR transcriptions are fed as input.

6.1 Model Performance

To explore whether the use of noisy ASR transcriptions affects the automatic evaluation of psychotherapy, we focus on a behavioral coding task where we seek to label participants' utterances into a set of predefined codes relevant to counseling quality using transcripts that are either manual or automatically generated.

We use the utterance-level annotations provided with the dataset described in Section 3, which consist of ten codes for therapist language plus two additional codes for annotated language from therapists and clients. We thus conduct a multi-label classification task to assign each utterance in the conversation to any of these 12 labels.

Our experiments are performed using a BERT model as our baseline classifier (Devlin et al., 2019) and our evaluated are conducted using 5-fold cross-validation. BERT is a transformer-based model that has been widely used in NLP. We chose this model since pretrained parameters fine-tuned on large natural language corpora are readily available, and also because due to its design the additional context input could easily supplied through the use of separate token type ids. We used the version implemented in (Wolf et al., 2020) with a learning rate of 2e-5. The input to the model is a sequence of token-level embeddings of each utterance in the conversation and the predicted label is assigned using a multilayer perceptron. The experiments are run on a GeForce RTX 2080 Ti.

We first conduct a set of experiments where we train and test multi-class utterance classifiers using either manual or automatic transcripts. In our first experiment, we aim to measure the model accuracy when using high quality training data i.e., manual transcripts for both, testing and training sets. Second, we substitute the train set for its automatically transcribed version and test on a manually transcribed set to evaluate the potential performance loss when training with noisy transcripts. Third, we again train on manual transcripts but this time test on automatic transcripts to evaluate whether a model built with accurate transcripts (i.e., produced by humans) would be effective while testing on transcriptions that are automatically obtained. Fi-

Category	N	TP	FN	FP	Recall	Precision	Avg Word Len	Std Word Len
FAMILY	926	827	99	8	89.31	99.04	5.10	1.38
FEEL	2470	2191	279	47	88.70	97.90	4.12	0.43
POSFEEL	8614	7568	1046	454	87.86	94.34	4.03	0.20
HOME	1550	1360	190	14	87.74	9.898	4.87	1.20
LEISURE	1966	1690	276	21	85.96	98.77	4.92	1.29
JOB	2077	1778	299	21	85.60	98.83	4.98	1.38
OPTIM	791	669	122	13	84.58	98.09	4.53	1.39
SELF	43100	36433	6667	3338	84.53	91.61	1.34	0.66
SOCIAL	54504	45866	8638	3907	84.15	92.15	3.31	1.02
ANX	268	224	44	2	83.58	99.12	6.07	3.00
POSEMO	20712	17152	3560	840	82.81	95.33	3.73	1.29
ANGER	412	341	71	8	82.77	97.71	4.18	1.71
AFFECT	23044	18993	4051	867	82.42	95.63	3.83	1.39
BODY	1721	1398	323	27	81.23	98.11	4.62	1.47
PHYSCAL	4042	3224	818	57	79.76	98.26	4.87	1.49
MONEY	674	534	140	4	79.23	99.26	4.95	1.30
EATING	2063	1633	430	27	79.16	98.37	5.35	1.78
NEGEMO	2130	1684	446	27	79.06	98.42	4.61	1.80
SAD	755	587	168	13	77.75	97.83	4.90	1.35
SCHOOL	492	379	113	2	77.03	99.48	5.04	1.40
SLEEP	212	163	49	2	76.89	98.79	3.99	1.01
DOWN	552	415	137	3	75.18	99.28	3.36	0.77
DEATH	152	112	40	1	73.68	99.12	3.72	0.73
FRIENDS	110	81	29	0	73.64	100	5.72	0.83
SEXUAL	253	186	67	5	73.52	97.38	4.04	0.45
RELIG	234	166	68	2	70.94	98.81	3.30	0.65

Table 5: Performance on LIWC-identified Keywords

Category: DEATH, BODY / Error Type: Omission
Manual: And that’s losing all the **weight**, and I really felt like I was **dying**
ASR: And to Annette loosen all the way. And I really felt like I was there.
Category: MONEY / Error Type: Insertion
Manual: Oh money to buy the cigarettes, and not to buy medicine Exactly Because it’s expensive.
ASR: Money to buy cigarettes, but no **money** for the medicine exactly six months ago

Table 6: Sample ASR errors for LIWC-identified keywords

nally, we evaluate a fully automatic pipeline, where both, train and test sets are obtained using ASR models. Results for these experiments are shown in Table 9.

6.2 Performance Trade-off

As results in Table 7 indicate, the choice of transcription method for both training and testing sets has a significant impact on the classification performance. Here, we see that even the model trained on the same manually transcribed training data can have drastically different reported performance, depending on the transcription method of the testing set. On the other hand, we also note that using ASR transcription as training set leads to a large decrease in performance when tested using manual testing data.

Since manual transcription is the most accurate representation of speech data, working with manual transcriptions would be the optimal choice. However, manual transcription can be expensive, especially for situations where a large amount of data has been collected. Thus, in many cases ASR technologies provide a faster and much more affordable transcription method. However, supervised learning with noisy ASR transcripts may result in the model learning spurious correlations, rather than the desired relationship between certain linguistic patterns and the predicted variables. This in turn leads to lower performances as shown in our experiments, where we observe performance losses up to 15%. Furthermore, consider a real case reported by Miner et al. (2020), where the word “depressed” was incorrectly transcribed into “the preston” in a self-harm counseling session. If an emotion detector were to be trained on the automatically transcribed data, the obvious correlation between “depressed” and “sad, blue” emotions will be lost, and replaced with a spurious one.

These considerations raise the question of what would be the best trade-off between the use of manual and automatic transcription methods in the psychotherapy domain.

To answer this question, we conduct a set of ex-

Train	Test	Acc.	F-score							
			QUEST	CR	SR	NAT	NAC	SEEK	NGI	PWOP
Manual	Manual	0.6940	0.6071	0.5334	0.0794	0.6919	0.8758	0.3058	0.5186	0.0048
Automatic	Manual	0.5520	0.4529	0.3642	0.0010	0.2587	0.7587	0.0815	0.3789	0.0127
Manual	Automatic	0.5289	0.4135	0.2483	0.0002	0.2263	0.7382	0.1060	0.2915	0.0173
Automatic	Automatic	0.5645	0.5268	0.3538	0.0020	0.2688	0.7765	0.1209	0.4341	0.0189

Table 7: Classification results for behavioral coding in MI sessions. AF, CON, NPWP, AUTO are not reported as their F-scores are zero

% of Manual Data	Acc.	F-score							
		QUEST	CR	SR	NAT	NAC	SEEK	NGI	PWOP
0%	0.5520	0.4529	0.3642	0.0010	0.2587	0.7587	0.0815	0.3789	0.0127
20%	0.6173	0.5820	0.5053	0.0076	0.4360	0.8132	0.1821	0.4865	0.011
40%	0.6734	0.5988	0.5225	0.0241	0.6397	0.8601	0.2981	0.4943	0.0276
60%	0.6827	0.5966	0.5298	0.0336	0.6700	0.8678	0.2314	0.4996	0.0021
80%	0.6914	0.6061	0.5340	0.0810	0.6866	0.8726	0.3073	0.5119	0.0534
100%	0.6940	0.6071	0.5334	0.0794	0.6919	0.8758	0.3058	0.5186	0.0048
Majority Class Classifier	0.4321	0.0	0.0	0.0	0.6034	0.0	0.0	0.0	0.0

Table 8: Classification results for behavioral coding for incremental fraction of manual transcripts in training set. The majority class classifier outputs the majority label in the training dataset for each instance

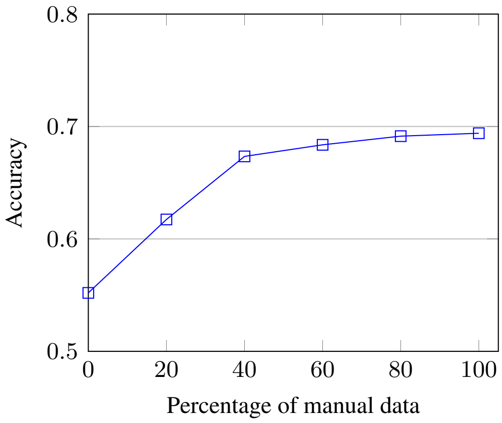

Figure 1: Classification accuracy as the fraction of manually transcribe data increases in the training set

periments where we gradually mix manually and automatically transcribed data during the training phase of the classification model. To ensure that the model is learning fairly, we ensured that each utterance only appears once in the entire dataset, without appearing both in the manual or ASR sets. By progressively adding more manual data in the training set, we emulate practical settings where only a fraction of data can be manually transcribed due to cost or time constrains. More specifically, we start with a full training set using ASR transcription, and increase the percentage of manual data at 20% increments. Note that reported accuracy is measured in a manually transcribed testing set.

As shown in Figure 1, the performance of the trained system does increase as the fraction of manual data increases. However, this is not shown as a linear relationship, as most of the performance gain occurs in the first few additions of the manual data. Although further study is warranted to explain how the small fraction of manual transcription leads to a noticeable increase in performance, this result indicates that even a small amount of manual transcription effort can improve the system performance in a meaningful way, and thus manual transcription is more cost-effective in its early stages than its later stages. For example, in the context of this experiment, practitioners can expect approximately 85% of the performance improvement of full manual transcription at the price of manually transcribing only 40% of the dataset.

6.3 Can (noisy) Local Context Help?

ASR error correction is an ongoing research topic in signal processing and natural language processing communities, and several techniques, including post-editing and domain adaptation, have been proposed (Mani et al., 2020). However, in this paper, we explore a simpler strategy based on context augmentation considering the distributional hypothesis in semantic theory, which states that words appearing in the same contexts tend to have similar meaning (Harris, 1954). We thus hypothesize that augmenting the target utterance with local context consisting of neighboring utterances can alleviate the effect of noisy transcription.

To this end, we compare BERT-based classifiers with different amounts of local context in addition to the target utterance (Devlin et al., 2019). The results shown in Table 9 are averaged over the re-

	Accuracy	Macro F1
No Context	0.5645	0.2085
Context = 1	0.5762	0.2297
Context = 2	0.5772	0.2290

Table 9: Classification results for behavioral coding when using local context

sult of five-fold cross validation. The “No Context” model is given a single utterance as input, and the final label by computing softmax after the final linear layer. For the “Context = n” models, n previous and following utterances surrounding the target utterance are also provided to the BERT model, as a concatenation. Note that through the use of separate token type ids, BERT allows practitioners to separately designate a sequence of context tokens, distinct from the target tokens. Overall, models that integrate context information outperform the base model in terms of average accuracy and Macro F1 with small but consistent performance gains, thus suggesting that the system’s performance can be improved using this simple strategy as opposed to conducting expensive manual transcription.

7 Limitations

Our work has several limitations that should be addressed through future work. First, our study only considers Google’s ASR and although this a popular choice there are several other commercial and open source alternatives. Initially, we also explored the use of Amazon Transcribe Medical[3]; however initial experiments did not show much variation with respect to the use of Google ASR. Nonetheless, further analysis is needed to evaluate how well the findings of this work will generalize to other ASR systems. Second, the computed WER and semantic distance are noisy, since the timestamps we used to align manual and automatic transcriptions were obtained through forced alignment. Furthermore, we did not evaluate the speaker diarization performance of the ASR system in identifying speaker’s role. Current ASR systems, including Google’s speech-to-text, offer the functionality to automatically assign speaker identities to transcribed utterances, and this feature might be useful for automatically assigning speaker roles to each utterance. Finally, we limited our focus to the behavioral coding task.

8 Conclusion and Lessons Learned

In this work, we conducted an evaluation of automatic speech recognition in the counseling domain using conversations between counselors and clients. To measure the degree of transcription error introduced by the use of an ASR system, we conducted domain-agnostic and domain-relevant evaluations using WER and semantic distance. Our analysis showed that while WER and semantic distance are in the 35 to 40% range when conducting a domain agnostic evaluation, the transcription error is slightly lower when considering transcription segments that are relevant to the domain i.e., utterances identified as important in evaluating the quality of counseling.

Moreover, we examined how the ASR step fits in and impacts the larger pipeline of an NLP system for behavioral coding in psychotherapy by comparing how the use of ASR data in place of manually transcribed data affects the performance of the downstream NLP system. Finally, we empirically showed that augmenting the system input with local context may alleviate the impact of noisy transcription. Given the results and analyses of this work, we conclude with the following lessons we learned in this study, on using ASR for NLP applications in psychotherapy and counseling: (1) Aggregate error measures are not sufficient by themselves, and must be complemented with domain-specific evaluations. (2) ASR error rates and performances differ across speaker roles and demographics as well as utterance content/topics. (3) Even a relatively small amount of manual transcription effort can help counteract noisy ASR and improve performance during the training of NLP models for psychotherapy applications.

Acknowledgment

This material is based in part upon work supported by the Precision Health initiative at the University of Michigan, by the National Science Foundation (grant #1815291), and by the John Templeton Foundation (grant #61156). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the Precision Health initiative, the National Science Foundation, or John Templeton Foundation.

[3] https://aws.amazon.com/transcribe/medical/

References

M. Adda-Decker and L. Lamel. 2005. Do speech recognizers prefer female speakers? In *Interspeech*.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Nikolaos Flemotomos, Victor R. Martinez, Zhuohao Chen, Karan Singla, Victor Ardulov, Raghuveer Peri, Derek D. Caperton, James Gibson, Michael J. Tanana, Panayiotis Georgiou, Jake Van Epps, Sarah P. Lord, Tad Hirsch, Zac E. Imel, David C. Atkins, and Shrikanth Narayanan. 2021. Automated evaluation of psychotherapy skills using speech and language technologies.

Sharon Goldwater, Dan Jurafsky, and Christopher D. Manning. 2008. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase ASR error rates. In *Proceedings of ACL-08: HLT*, pages 380–388, Columbus, Ohio. Association for Computational Linguistics.

Google. 2020. Google. cloud speech-to-text

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Zac E. Imel, M. Steyvers, and David C. Atkins. 2015. Computational psychotherapy research: scaling up the evaluation of patient-provider interactions. *Psychotherapy*, 52 1:19–30.

Suyoun Kim, Abhinav Arora, Duc Le, Ching-Feng Yeh, Christian Fuegen, Ozlem Kalinli, and Michael L. Seltzer. 2021. Semantic distance: A new metric for asr performance analysis towards spoken language understanding.

Xiang Kong, Jeung-Yoon Choi, and Stefanie Shattuck-Hufnagel. 2016. Evaluating automatic speech recognition systems in comparison with human perception results using distinctive feature measures.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models.

Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Paden Tomasello, Jacob Kahn, Gilad Avidov, Ronan Collobert, and Gabriel Synnaeve. 2020. Rethinking evaluation in asr: Are our models robust enough?

Anirudh Mani, Shruti Palaskar, Nimshi Venkat Meripo, Sandeep Konam, and Florian Metze. 2020. Asr error correction and domain adaptation using machine translation.

Adam S. Miner, Albert Haque, Jason Alan Fries, S. L. Fleming, D. Wilfley, G. Terence Wilson, A. Milstein, D. Jurafsky, B. Arnow, W. Stewart Agras, Li Fei-Fei, and N. Shah. 2020. Assessing the accuracy of automatic speech recognition for psychotherapy. *NPJ Digital Medicine*, 3.

Robert M Ochshorn and Max Hawkins. gentle forced-aligner.

Youngja Park, Siddharth Patwardhan, K. Visweswariah, and S. C. Gates. 2008. An empirical analysis of word error rate and keyword error rate. In *INTERSPEECH*.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count*. Lawerence Erlbaum Associates, Mahwah, NJ.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51, San Diego, CA, USA. Association for Computational Linguistics.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J. Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137, Valencia, Spain. Association for Computational Linguistics.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Nagendra Goel, Mirko Hannemann, Yanmin Qian, Petr Schwarz, and Georg Stemmer. 2011. The kaldi speech recognition toolkit. In *In IEEE 2011 workshop*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20, 1st virtual meeting. Association for Computational Linguistics.

Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 2020. WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3290–3295, Online. Association for Computational Linguistics.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

B. Xiao, Zac E. Imel, P. Georgiou, David C. Atkins, and Shrikanth S. Narayanan. 2015. "rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLoS ONE*, 10.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020a. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5276–5289, Online. Association for Computational Linguistics.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020b. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In *Proceedings of ACL*.