Detecting Institutional Dialog Acts in Police Traffic Stops

Vinodkumar Prabhakaran
Stanford Univ., CA

Camilla Griffiths
Stanford Univ., CA

Hang Su
UC Berkeley, CA

Prateek Verma
Stanford Univ., CA

Nelson Morgan
ICSI Berkeley, CA

Jennifer L. Eberhardt
Stanford Univ., CA

Dan Jurafsky
Stanford Univ., CA

{vinodkpg,camillag,jleberhardt,jurafsky}@stanford.edu,
{suhang3240,prateek119}@gmail.com, morgan@uprise.org

Abstract

We apply computational dialog methods to police body-worn camera footage to model conversations between police officers and community members in traffic stops. Relying on the theory of *institutional talk*, we develop a labeling scheme for police speech during traffic stops, and a tagger to detect institutional dialog acts (Reasons, Searches, Offering Help) from transcribed text at the turn (78% F-score) and stop (89% F-score) level. We then develop speech recognition and segmentation algorithms to detect these acts at the stop level from raw camera audio (81% F-score, with even higher accuracy for crucial acts like conveying the reason for the stop). We demonstrate that the dialog structures produced by our tagger could reveal whether officers follow law enforcement norms like introducing themselves, explaining the reason for the stop, and asking permission for searches. This work may therefore inform and aid efforts to ensure the procedural justice of police-community interactions.

1 Introduction

Improving the relationship between police officers and the communities they serve is a critical societal goal. We propose to study this relationship by applying NLP techniques to conversations between officers and community members in traffic stops. Traffic stops are one of the most common forms of police contact with community members, with 10% of U.S. adults pulled over every year (Langton and Durose, 2013). Yet past research on what people experience during these traffic stops has mainly been limited to self-reported behavior and post-hoc narratives (Lundman and Kaufman, 2003; Engel, 2005; Brunson, 2007; Epp et al., 2014).

The rapid adoption of body-worn cameras by police departments in the U.S. (laws in 60% of states in the U.S. encourage the use of body cameras) and across the world has provided unprecedented insight into traffic stops.[1] While footage from these cameras is used as evidence in contentious cases, the unstructured nature and immense volume of video data means that most of this footage is untapped.

Recent work by Voigt et al. (2017) demonstrated that body-worn camera footage could be used not just as evidence in court, but as data. They developed algorithms to automatically detect the degree of respect that officers communicated to drivers in close to 1,000 routine traffic stops captured on camera. It was the first study to use machine learning techniques to extract insights from this footage.

This footage can be further used to unearth the structure of police-community interactions and gain a more comprehensive picture of the traffic stop as an every day institutional practice. For instance, knowing which requests the officer makes, whether and when they introduce themselves or explain the reason for the stop is a novel way to measure *procedural justice*; a set of fairness principles recommended by the President's Task Force on 21st Century Policing,[2] and endorsed by police departments across the U.S.

[1] https://en.wikipedia.org/wiki/Body_worn_video_(police_equipment)

[2] http://www.theiacp.org/TaskForceReport

Transactions of the Association for Computational Linguistics, vol. 6, pp. 467–481, 2018. Action Editor: Jordan Boyd-Graber .
Submission batch: 11/2017; Revision batch: 2/2018; Published 7/2018.

We propose automatically extracting dialog structure from body camera footage to contribute to our understanding of police-community interactions. We rely on the notion of *institutional talk* (Heritage, 2005), which posits that dialog acts, topics, and narrative are heavily defined by the institutional context. Traffic stops are a kind of institutional talk; as are, for example, doctor-patient interactions, counseling conversations, and citizen calls for help from police. We introduce a model of *institutional acts* for traffic stop conversations. Since the officer holds a position of power within this institutional context, their dialog behavior has a greater influence in shaping the conversation (Coupland et al., 1991; Gnisci, 2005); hence, we focus on the institutional acts performed by the officer in this paper.

Contributions of our paper: 1) A typology of institutional dialog acts to model the structure of police-driver interactions during traffic stops. 2) An institutional act tagger that works from transcribed words (78% F-score) or from raw audio (60% F-score). 3) A classifier that uses this dialog structure to detect acts at the stop level (e.g., "Does this stop contain a Reason?") (81% F-score from raw audio). 4) An analysis of salient dialog structure patterns in traffic stops; demonstrating its potential as a tool for police departments to assess and improve police community interactions.

2 Background

Computational work on human-human conversation has long focused on *dialog structure*, beginning with the influential work of Grosz showing the homology between dialog and task structure (Grosz, 1977). Recent work has integrated speech act theory (Austin, 1975) and conversational analysis (Schegloff and Sacks, 1973; Sacks et al., 1974; Schegloff, 1979) into models of dialog acts for domains like meetings (Ang et al., 2005), telephone calls (Stolcke et al., 2006), emails (Cohen et al., 2004), chats (Kim et al., 2010), and Twitter (Ritter et al., 2010).

Our models extend this work by drawing on the notion of institutional talk (Atkinson and Drew, 1979), an application of conversational analysis to environments in which the goals of participants are institution-specific. Actions, their sequences, and interpretations during institutional talk depend not only on the speaker (as speech act theory suggests) or the dialog (as conversational analysts argue), but they are inherently tied to the institutional context.

Institutional talk has been used as a tool to understand the work of social institutions. For example, Whalen and Zimmerman (1987) studied dialog structure in transcripts of citizen calls for help. They observed that the "regular, repetitive and reproducible features of calls for police, fire or paramedic services [...] arise from situated practices responsive to the sequential and institutional contexts of this type of call". Such recurring patterns in language and conversation exist across different institutional contexts such as doctor-patient interactions, psychological counseling, sales calls, court room conversations, as well as traffic stops (Heritage, 2005).

Deviations from these sequential configurations are consequential. A police officer failing to explain the reason for the traffic stop can lead to aggravation in the driver (Giles et al., 2007), and an officer's perceived communication skills (e.g. do they listen, take civilian views into account) predict civilian's attitudes towards the police (Giles et al., 2006).

These findings demonstrate the importance of understanding the role of institutional context in shaping conversation structure. In doing so, our paper also draws on recent research on automatically extracting structure from human-human dialog. Drawing on Grosz's original insights, Bangalore et al. (2006) show how to extract a hierarchical task structure for catalog ordering dialogs with subtasks like opening, contact-information, order-item, related-offers, and summary. Prabhakaran et al. (2012) and Prabhakaran et al. (2014) employ dialog act analysis to study correlates of gender and power in work emails, while Althoff et al. (2016) studied structural aspects of successful counseling conversations, and Yang et al. (2013) and Chandrasekaran et al. (2017) investigated structures in online classroom conversations that predict success or need for intervention. Our work also draws on an important line of unsupervised work that models topical structure of conversations (Blei and Moreno, 2001; Eisenstein and Barzilay, 2008; Paul, 2012; Nguyen et al., 2012).

Our work is closely related to the active line of research in NLP on dialog act classification. Recently, recurrent neural network-based dialog act taggers, e.g., Khanpour et al. (2016), Li and Wu (2016) and

Liu et al. (2017), have posted state-of-the-art performance on benchmark datasets such as the Switchboard corpus (Jurafsky et al., 1997) and MRDA (Ang et al., 2005). Since these corpora come from significantly different domains (telephone conversations and meeting transcripts, respectively) than ours, and since we are interested specifically in the institutional acts (e.g., *did the officer request documentation from the driver?*) rather than the general dialog acts (*did the officer issue a request?*), these taggers do not directly serve our purpose. Furthermore, our data is an order of magnitude smaller (around 7K sentences) than these corpora; making it infeasible to train in-domain recurrent networks.

Prior to neural network approaches, support vector machines and conditional random fields (Cohen et al., 2004; Kim et al., 2010; Kim et al., 2012; Omuya et al., 2013) were the state-of-the-art algorithms on this task. These approaches also incorporated contextual and structural information into the classifier. For instance, Kim et al. (2012) used lexical information from previous utterances in predicting the dialog act of a current utterance; and Omuya et al. (2013) uses features such as the relative position of an utterance w.r.t the whole dialog. We draw from this line of work; we also experiment with positional and contextual features in addition to lexical features. Furthermore, we use features that capture the institutional context of the conversation.

3 Institutional Dialog Acts of Traffic Stops

We begin with a framework for analyzing the structure of interactions in this important but understudied domain of traffic stop conversations, developed by applying a data-oriented approach to body camera footage. Our goal is to create a framework that can be a tool for police departments, policy makers, and the general public to understand, assess and improve policing practices.

3.1 Data

We use the Voigt et al. (2017) dataset of body camera audio from 981 vehicle stops conducted by the Oakland Police Department during the month of April 2014. This amounts to 35 hours of speech, hand-transcribed to 94K speaker turns and 757K words.

Officer.:	Sir, hello, my name's Officer [NAME] of the Oakland Police Department. [Greeting]
Driver:	Hi.
Officer.:	The reason why I pulled you over is when you passed me back there you were texting or talking on your cell phone. [Reason]
Driver:	I was looking at a text, yes.
Officer.:	Okay. Do you have um, what year is the car you're driving? [Details]
Driver:	It's a 2010.
Officer.:	2010. Do you still live in [ADDRESS]? [Details]
Driver:	Yes.
[...]	
Officer.:	All right, sir. This is a citation for having your cell phone in your hand while you're driving. [ ]You actually have two months on or before June 7th to take care of the citation, okay? Please drive carefully. [Sanction; PositiveClosing]
Driver:	Okay.
Officer.:	Thank you.

Table 1: Excerpt from a traffic stop conversation with institutional acts in [blue] (names/addresses redacted).

3.2 Traffic Stops as Institutional Talk

Traffic stops possess all three characteristics of institutional talk (Heritage, 2005): i) participants' goals are tied to their institution-relevant identity (e.g. officer & driver); ii) there are special constraints on what is allowable within the interaction; iii) there are special inferences that are particular to the context. Table 1 presents an excerpt from a traffic stop conversation from our corpus: The officer greets the community member, gives the reason for the stop, asks about personal details, issues the sanction, and closes by encouraging safe driving. We are interested in such recurring sequences of institution-specific dialog acts, or *institutional acts*, which combine aspects of dialog acts and those of topical segments, all conditioned by the institutional context.

3.3 Developing the Typology

To develop the taxonomy of institutional dialog acts, we begin with a data-oriented exploration: identifying recurring sequences of topic segments using the (unsupervised) mixed membership Markov model (Paul, 2012).[3] Figure 1 shows the topic segments assigned by a 10-topic model on the traffic stop of Table 1. The model identified different spans of con-

[3] We trained the model on a subset of 541 stop transcripts from our data, exploring different numbers of topics.

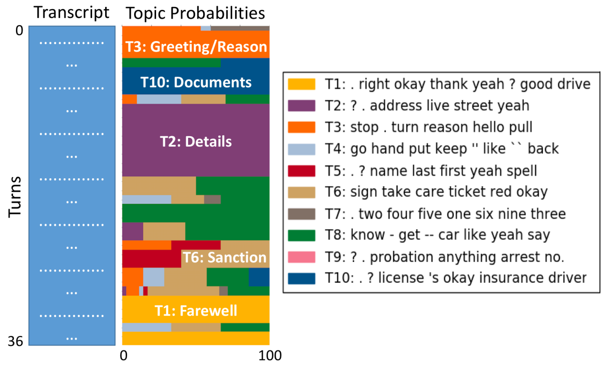

Figure 1: Topic assignments from Mixed Membership Markov Modeling (Paul, 2012) on a sample stop (turns go from top to bottom; x-axis shows probabilities assigned to each topic; right are the top topic words). The model identifies the reason for the stop (orange), driver's documents (blue), driver's address and demographics (purple), the sanction (beige) and closing (yellow).

versation; the officer gives the reason for the stop (orange), asks for documents (blue), collects driver information (purple), then in the end, there are spans of issuing a sanction (beige) and closing (yellow).

While these topical assignments helpfully suggest a high-level notion of the structure of these conversations, they do not capture the specific acts officers do. We next turned to the procedural justice literature, which highlights specific acts. For instance, questioning the driver's legitimacy for being somewhere (*why are you here?*) or driving a car (*whose car is it?*) are acts that trigger negative reactions in drivers (Epp et al., 2014). On the other hand, officers introducing themselves and explaining the reasons for the stop are important procedural justice facets that communicate fairness and respect (Ramsey and Robinson, 2015). Informed by the procedural justice literature, the President's Task Force recommendations, and a review of the unsupervised topic segments, two of the authors manually analyzed twenty stop transcripts to identify institutional dialog acts.

We focused on acts that tend to recur (e.g. citations), and those with procedural justice interest (e.g. reasons, introductions), teasing apart acts with similar goals but different illocutionary force (explicitly stating vs. implying the reason for the stop; or requesting to search the vehicle vs. stating that a search was being conducted). This process resulted in an initial coding scheme of twenty two institutional acts in nine categories. We also observe that the recurring acts by community members were often in response to officers' acts (e.g., responding to demographic questions), as their position of power gives them higher influence in shaping the conversation (Giles et al., 2007). Hence, we focus on officer speech to capture our institutional act annotations.

3.4 Annotating Institutional Acts

From each stop transcript, we selected all officer turns (excluding those directed to the radio dispatcher), and annotated each sentence of each turn.

In the first round, three annotators annotated the same 10 stops using the taxonomy and manual developed above with an average pair-wise inter-annotator agreement of κ=0.79. We discussed the sources of disagreement, ratified the annotations, and updated the annotation manual to clarify act descriptions. During this process, we also updated the annotation manual to include four additional institutional acts, resulting in a set of twenty five acts in eleven categories. Table 2 presents this final typology, along with actual examples from our data.

We then performed two subsequent rounds of three-way parallel annotations obtaining average pair-wise κ values of 0.84 and 0.88, respectively. Once we obtained high agreement, we conducted a fourth round where each annotator annotated a separate set of 30 stops. Stops were chosen at random from the entire corpus for each round; however, seven of the previously annotated stops were incorrectly included in the final round of annotations, resulting in a total of 113 annotated stops (7081 sentences, 4245 turns). Table 1 shows resulting labels.

4 Learning to Detect Institutional Acts

We now investigate whether we can train a model that can automatically detect the institutional acts during the course of a traffic stop. In Sections 5-7, we present an institutional act tagger, and describe three increasingly difficult evaluation settings:

1. **Using manual transcripts**: We train and test an institutional act tagger on the manual transcripts. This task is similar to dialog act tagging (e.g., (Stolcke et al., 2006)), but it has the important distinction that it needs to captures dialog structure at the intersection of the general dialog acts

Event (Coarse-grained)	Event (Fine-grained)	Count	Example Utterances
GREETING	Greeting	98	“Whats up, yall?”, “How you doing, man?”, “Hello.”
	Introduction	16	“Hi. Im officer ____, Oakland PD”
REASON	Question Awareness	12	“You know why Im pulling you over?”
	Explicit	127	“Reason I pulled you over is for a cell phone violation.”
	Implicit	19	“Didnt see the stop sign?”
DOCUMENTS	Requesting Documents	252	“You have your drivers license, registration and insurance?”
DETAILS	Demographics	71	“How old are you?”, “Whats your last name?”
	Address	65	“What’s your address?”,“Where do you live at?”
SANCTION	Issuing Citation	37	“Okay, as I say, the reason I’m citing you is for failure to yield to oncoming traffic.”
	Issuing Fix-it Ticket	31	“I’ll give you a fix-it ticket for the headlight, left front headlight, all right?”
	Issuing Warning	19	“I’ll give you a warning today.”
	Mention Lenience	50	“Im cutting you guys a break”
POSITIVECLOSING	Farewell	86	“All right. Drive safe”, “All right, guys. Take care”, “Have a good day.”
ORDERS	Hands On Wheel	9	“Hey just keep your hands on the steering wheel man”
	Turn Car Off	37	“Hey, turn the car off”
LEGITIMACY	Vehicle Ownership	41	“This your car?”
	Questioning Intent	15	“What are you doing out here?”
HISTORY	Warrants	3	“Do you know you got a little warrant too?”
	Probation/Parole	16	“You know you’re on probation, right?”
	Arrests	4	“Do you, um, have you ever been arrested?”
OFFERHELP	Giving Voice	19	“Do you have any questions?”, “You understand?”
	Offering Help	5	“Need any help getting back on the traffic?”, “You need directions?”
SEARCH	Request for Search	3	“Do you mind if I uh search the car?”
	Statement of Search	7	“Youre on probation so you have a search clause.”
	Weapons	15	“You got nothing on you I need to worry about?”, “No weapons, right?”

Table 2: Typology of institutional acts during traffic stops. Column 1 shows the 11-way coarse-grained groupings. Column 2 shows the 25-way fine-grained institutional act labels used for annotations, and Column 3 shows the number of sentences labeled with each acts.

(e.g., requests, responses) and the topical structure. Section 5 presents the experiments on building the institutional act tagger for this domain.

2. **Using ASR**: We develop an automatic speech recognizer that works in our domain, and uses the text it generates, instead of manual transcripts, to train and test the model. The downstream institutional act tagging framework stays the same. This setting is not fully automatic, as we still rely on the manually identified segments of audio where officers spoke. Section 6 first presents experiments on building the ASR system for this domain, and then presents results on using ASR-generated text for institutional act tagging.
3. **From raw audio**: We build automatic means to detect the segments of officers’ speech, apply the ASR on those segments, and then use the text thus produced to detect institutional acts, building a fully automatic tagger with no human intervention. Section 7 first describes the experiments on detecting the officers’ speech automatically, and then presents results on institutional act tagging in this fully automatic setting.

For all our experiments, we merge labels from all sentences in each turn, making this a multi-label (instead of multi-class) classification task.[4] Only around 7% of the institutional act bearing utterances had multiple acts. Common co-occurrences were GREETING and REASON, and GREETING and ORDERS, e.g., *Hey, turn the car off. How you doing?*

5 Institutional Act Tagging from Manual Transcripts

We adopt a supervised machine learning approach to the task of institutional act tagging. We draw from prior work in the area of dialog act modeling, while also adding features that specifically capture the institutional context of traffic stop conversations.

5.1 Algorithms

We compared three supervised text classification methods: Support Vector Machine (SVM) (Cortes and Vapnik, 1995) and Extremely Randomized

[4] We present turn-level (instead of sentence-level) predictions to facilitate comparisons with experiments presented in Section 6 & 7; sentence-level experiments were performed using manual transcripts and yielded slightly better numbers.

Trees (ERT) (Geurts et al., 2006),[5] which are efficient and tend to work well with smaller datasets like ours, and Convolutional Neural Network (CNN) (Kim, 2014), which captures variable length patterns without feature engineering. For SVM, we use the one-vs-all multi-label algorithm (ERT and CNN inherently deal with multi-label classification) and use the *balanced* mode to address the skewed label distribution (0.5% to 3.5% positive cases). In the balanced mode, positive and negative examples are balanced at training time. For CNN, we use two convolutional layers of filter sizes 3 and 4 and 20 filters with *relu* activation and max-pooling with pool size 2. This is followed by two dense layers, and a final layer with sigmoid activation and binary cross entropy loss to handle multi-label classification.

While some prior work in dialog act tagging (e.g., (Kim et al., 2010; Kim et al., 2012) have shown that sequence tagging algorithms such as conditional random fields (CRF) have some advantage over text classification approaches such as SVMs, preliminary experiments using CRFs revealed this to not be the case in our corpus.

5.2 Features

Lexical Features: We used unigrams and bigrams as indicator features for SVM and ERT. We initialize the input layer of CNN with word embeddings trained using our entire transcribed dataset.[6]

Pattern features: We use indicator features for two types of patterns. 1) For each institutional act, we hand-crafted a list of linguistic patterns; e.g., the pattern feature for GREETING included how are you, hello, and good morning, among others. 2) We use a semi-automatically built dictionary of offenses (e.g., *tail light*) by querying the word embedding model trained on all transcripts with a seed list of offenses, resulting in a large list of offenses and variations of their usage (e.g., break light, rear lite) with high incidence in some acts (e.g., REASON, SANCTION).

Algorithm	P	R	F
Extremely Randomized Trees	**80.9**	63.6	71.2
Conv. Neural Network	77.4	57.3	65.8
SVM	78.9	**76.2**	**77.5**
SVM (- ngrams)	15.4	83.3	26.0
SVM (- patterns)	78.4	74.4	76.4
SVM (- structure)	76.3	74.2	75.3
SVM (- patterns&structure)	76.3	71.9	74.0

Table 3: Micro-averaged precision (P), recall (R) and F-score (F) for experiments using manual transcripts.

Structural features: 1) The number of words in the utterance, since some acts (e.g., GREETING) require fewer words than others (e.g., SANCTION). We binned this feature into four bins: <3, 4-10, 11-20, and >20. 2) The position of the utterance within the conversation (e.g., SANCTION is likely to happen late, and GREETING early), binned to one or more of: first five, first quarter, first third, first half, last half, last third, last quarter, and last five.

Other features: We tried other features such as 1) ngrams from previous utterances, 2) ngrams from driver's responses, 3) dependency parse patterns, 4) word/sentence embeddings, and 5) topic assignments obtained from the mixed membership Markov model (Paul, 2012) discussed in Section 3.3. These features turned out not to be helpful for this task, and we do not include those results here.

5.3 Experiments and Results

Table 3 presents micro-averaged (i.e., weighted average of each class) precision, recall and F-measure obtained on 10-fold cross validation.[7] While ERT posted the highest precision of 80.9% at a low recall of 63.6%, SVM reported the highest recall of 76.2% without a huge dent in precision. Overall, we obtain the best micro-averaged F-score of 77.5% using SVM. CNN performed worse than both ERT and SVM.[8] We also performed an ablation study to see the relative importance of features in the SVM

[5] ERT is a variant of the random forest algorithm, with the difference that the splits at each step are selected at random rather than using a preset criteria.

[6] In preliminary experiments, we found that SVMs using these word embeddings (or GloVe embeddings) performed worse than using ngram features directly.

[7] CNN: batch size of 10, dropout of 0.3, adam, 10 epochs. SVM: C=1, linear kernel. ERT: 100 estimators, max tree depth 75, # of features capped at 20% of all features. Parameter values obtained using grid-search within the training set for each fold.

[8] Since CNN performed much worse than SVM with lexical features alone (last row), presumably because of the small amount of data, we did not perform more CNN experiments.

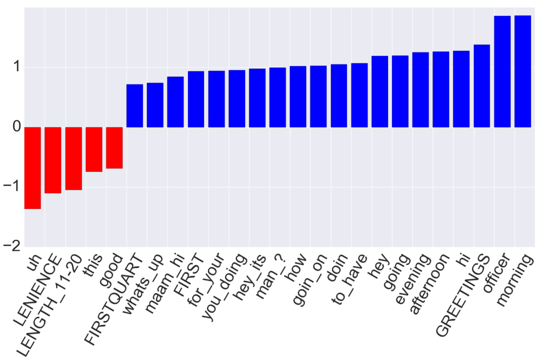

Figure 2: Top 25 most (by absolute value) weighted features in the GREETING model.

model. As expected, the ngram features contribute the most; removing them drastically lowered performance. Patterns and structural features had a smaller impact on performance.

We inspected the weights assigned to the features by a model trained on the entire dataset. The models created for each institutional act had at least one pattern or structure feature in the top twenty five features. Figure 2 shows the feature weights assigned to the model detecting GREETING. The model up-weighted utterances with greeting patterns (*GREETINGS*), first utterances (*FIRST*), and utterances in the first quarter (*FIRSTQUART*), while down-weighting longer utterances (*LENGTH_11-20*) and those that mention lenience (*LENIENCE*).

6 Institutional Act Tagging using ASR

The institutional act tagger of Section 5 relies on manual transcriptions, making it not scalable to the thousands of traffic stops conducted every month. We now investigate using automatic speech recognition, while assuming manual segmentation, i.e., we know the time segments where an officer spoke to the driver; in the next section we explore the additional task of automatic officer turn detection.

6.1 Data Augmentation

Traffic stops have considerable noise (wind, traffic, horns), overlap, and difficult vocabulary (names, addresses, jargon), making it a challenging domain for off-the-shelf automatic speech recognizers (ASR). However, our 35 hours of transcribed speech is insufficient to train a domain-specific recognizer. We therefore employ two data augmentation techniques.

Data	Recordings	Utterances	Hours
Train	603 + 2435	407,408	494
Dev	66	3,241	3.6
Test	113	4,248	4.6

Table 4: Data used to build the ASR models.

First, we perturb our data by frame-shifting and filterbank adjustment following the procedure described in (Ko et al., 2015). In frame-shifting, we change the starting point of each frame, making features generated from these frames slightly different from the original ones. For filterbank adjustment, we move the locations of the center frequencies of filterbank triangular frequency bins during feature extraction. This method increases our training data 5-fold to 180 hours. Second, we make use of the 300-hour Switchboard telephone speech dataset (Godfrey and Holliman, 1997) to create additional data. We first upsample Switchboard speech to the 16 KHz of our data, and then mix them with noise samples randomly picked from our data where speech is not identified, using a random speech-to-noise-ratio between 0 and 10. This method contributes another 300 hours of speech for training.

6.2 Acoustic Modeling

We implemented two acoustic models, a Bidirectional Long Short-Term Memory network (BLSTM) (Graves et al., 2013) and a Deep Neural Net Hidden Markov Model (DNN-HMM) tri-phone baseline. While LSTM based approaches generally work better, they are much slower to train, so we wanted to know if their word error improvements indeed translated to act tagger improvements.

DNN-HMM system training follows the standard pipeline in the Kaldi toolkit (Povey et al., 2011; Veselý et al., 2013). Frame alignments generated from a traditional Gaussian mixture model based system are used as targets and 40-dimension fMLLR features (Gales, 1998) are used as inputs to the DNN to aid speaker adaptation. The network was trained using Restricted Boltzmann Machine (RBM) based pretraining (Salakhutdinov et al., 2007) and then discriminatively trained using stochastic gradient descent with cross-entropy as loss function. (Veselý

Data	Perplexity
Traffic stops	79.4
+Switchboard	75.9
+Fisher	74.3

Table 5: Language model perplexity on Dev set.

et al., 2013) describes more training details.

We trained the BLSTM using the recipe proposed by Mohamed et al. (2015). The BLSTM is used to model short segments of speech (with a sliding window of 40 frames), and predict frame-level HMM states at each time frame[9]. We use 6 hidden layers and 512 LSTM cells in each direction. Dropout (Srivastava et al., 2014), peephole connections (Gers et al., 2002) and gradient clipping are adopted to stabilize training (Sak et al., 2014). As in DNN-HMM training, fMLLR features and frame alignments are used as inputs and targets respectively.

For decoding, frame posteriors from the acoustic model are fed into a weighted finite state transducer with HMMs, context-dependent tri-phone models, a lexicon,[10] and a 3-gram language model with Kneser-Ney smoothing (Kneser and Ney, 1995).

6.3 Language Model Data Augmentation

To mitigate language model data scarcity, we use transcriptions from the Switchboard and Fisher (Cieri et al., 2004) corpora, adding about 3.12M and 21.1M words, respectively. Separate language models are trained on these datasets, and then interpolated with the traffic stop language model; interpolation weights were chosen by minimizing perplexity on a separate Dev set. Table 5 shows the perplexities of different language models on this Dev set.

6.4 Evaluating ASR Models

Table 4 shows statistics of the data used to build the ASR system. We kept aside the 113 institutional act annotated stops from Section 3 as test set. The remaining 669 stops were divided 9:1 into Train and Dev sets. The Train set also includes the 2435 recordings from the Switchboard corpora.

Model	Dev	Test
DNN	57.0	48.5
BLSTM	**49.7**	**45.0**
BLSTM (- data augmentation)	56.9	51.4
BLSTM (- LM interpolation)	50.2	45.7

Table 6: Word error rate for different ASR models.

ASR Source	1Best	10Best
DNN	57.2	63.6
BLSTM	**65.0**	**65.3**

Table 7: Micro-averaged F-scores on institutional act prediction using different ASR sources.

Table 6 shows word error rates under different settings. Overall, we obtain relatively high error rates, largely due to the noisy environment of the audio in this domain. BLSTM performs better than DNN-HMM, consistent with prior research (Mohamed et al., 2015; Sak et al., 2014).[11] Interpolating Switchboard and Fisher language models provides a further boost of 0.7 percentage points.

6.5 Institutional Act Tagging Experiments

We now use text generated by ASR to train and test the institutional act tagger of Section 4. To increase recall, we also made use of N-best list output from the ASR systems, collecting ngram and pattern features from the top 10 candidate transcriptions. The L1 penalty in the SVM limits the impact of the resulting noisier ngrams on precision.

Table 7 presents micro-averaged F-scores. BLSTM with 10Best obtained the best F-score of 65.3. While using 10Best lists only helped marginally for BLSTM, it helped the DNN enough to eliminate most of the gap in performance with BLSTMs. Our results suggest that downstream tasks with efficiency constraints could employ DNNs without a huge dent in performance by making use of NBest or lattice output.

[9] Note that this recipe is different from the end-to-end approach where LSTM model takes in the whole utterance and predict phone / word outputs directly (Graves and Jaitly, 2014)

[10] CMU dictionary (CMUdict v0.7a) is used.

[11] Note that our Test set, designed for measuring institutional act detection, consists of only police officers talking close to the camera; hence the word error rate can be lower than the Dev, which is designed to measure overall ASR performance and includes community member speech as well.

ASR Source	1Best	10Best
DNN	43.7	56.0
BLSTM	53.8	59.8

Table 8: Micro-averaged F-scores on institutional act prediction from raw audio using different ASR sources.

7 Institutional Act Tagging from Raw Audio

We now turn to the task of detecting institutional acts directly from raw body camera audio. This requires detecting spans with speech activity and distinguishing them from noise— voice activity detection—and identifying segments spoken by the police officers.

7.1 Finding Officer Speech Segments

Our goal is to find regions of the audio with a high probability of being officer speech. We could not build a standard supervised officer-versus-other classifier, because the stops contain large untranscribed regions of officer speech (we did not transcribe segments where the officer was, for example, talking to the dispatcher in the car). We therefore instead built a two-output classifier to discriminate between the officer and community member speech, and used a tuned threshold (0.55) on the posterior probability of officer as our voice activity detector, drawing on the intuitions of (Williams and Ellis, 1999; Verma et al., 2015) who found that posterior features on speech tasks also improved speech/non-speech performance. Our model is a 3-layer fully connected neural network with 1024 neurons trained with cross entropy loss.[12] Figure 3 sketches the architecture. We run the classifier on each .5 second span; (recall=.97 and precision = .90 on the Dev set of Table 4), and then merge classifications to a single turn if adjacent spans are classified as officer speech, with a 500 ms lenience for pauses.

7.2 Institutional Act Tagging Experiments

We now present experiments using the automatically identified officer speech segments. At training time, we use the ASR generated text using gold segments; at test time, we use the same ASR model to generate text for the predicted segments. Since the predicted segments do not exactly match gold segments, we use a fuzzy-matching approach for evaluation. If a gold segment contains an act and an overlapping predicted segment has the same act, we consider it a true positive. If a gold segment contains an act, but none of the overlapping predicted segments have that act, it is counted as a false negative. If an act is identified in one of the predicted segments, without any of the overlapping gold segments having it, then we consider it a false positive.

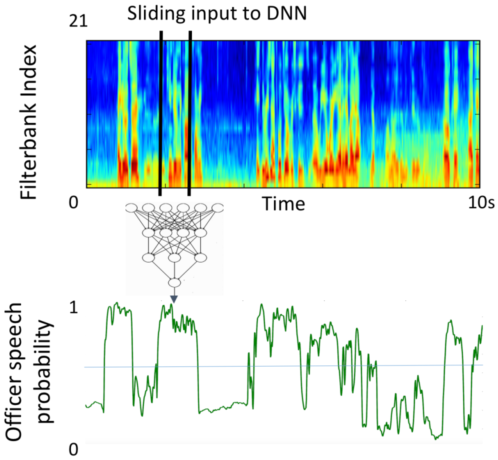

Figure 3: Detecting Officer Speech segments.

Table 8 presents results using this evaluation scheme. Again, BLSTM using the 10Best strategy obtained the best F-score of 59.8%. Both BLSTM and DNN benefited significantly from using the 10Best likely predictions. As in the ASR experiments, the DNN substantially closes the gap in performance by using the 10Best strategy.

8 Stop Level Act Detection

Our three previous sets of models focused on labeling each officer *turn* with one or more institutional acts. For many purposes, it suffices to ask a far simpler question: does an act occur *somewhere* in the traffic stop? From a procedural justice standpoint, for example, we want to know whether the officer explained the reason for the stop; we may not care about the turn in which the reason occurred.

We call this task *stop-level* act detection, in which each stop is labeled as a positive instance of an act if that particular act occurred in it in the gold labels.

[12] Patch of 210ms with a stride of 50ms. Audio was downsampled to 16kHz, and converted to 21-dimensional magnitude mel-filterbank representation covering frequencies from 0-8 kHz. FFT size was 512 with 10ms hop and 30ms frame size.

		Using Manual Transcripts			Using ASR Transcripts			Using Raw Audio		
Event	Count	Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
GREETING	80	92.3	90.0	91.1	84.5	88.8	86.6	70.2	91.3	79.4
REASON	96	94.7	93.8	94.2	94.3	86.5	90.2	96.4	84.4	90.0
DOCUMENTS	100	97.0	97.0	97.0	95.9	93.0	94.4	96.8	92.0	94.4
DETAILS	56	86.2	89.3	87.7	68.8	78.6	73.3	66.1	66.1	66.1
SANCTION	79	94.1	81.0	87.1	84.2	81.0	82.6	90.3	82.3	86.1
POSITIVECLOSING	71	91.2	87.3	89.2	84.4	76.1	80.0	90.6	67.6	77.4
ORDERS	32	87.1	84.4	85.7	90.3	87.5	88.9	96.6	87.5	91.8
LEGITIMACY	41	78.4	70.7	74.4	89.7	63.4	74.3	85.7	29.3	43.6
HISTORY	11	77.8	63.6	70.0	75.0	54.6	63.2	71.4	45.5	55.6
OFFERHELP	18	71.4	83.3	76.9	82.4	77.8	80.0	82.4	77.8	80.0
SEARCH	10	70.0	70.0	70.0	66.7	20.0	30.8	60.0	30.0	40.0
Micro Average (Weighted)		90.4	87.5	89.0	86.5	81.7	84.0	85.5	77.1	81.1
MacroAverage (Unweighted)		85.5	82.8	83.9	83.3	73.4	76.8	82.4	68.5	73.1

Table 9: Stop level institutional act presence detection results (for each label).

Our algorithm is simple: run our best turn-based act tagger, and if the tagger labels an institutional act anywhere in the conversation, tag the conversation as having that class.[13] We explore all three settings: manual segments and transcripts, manual segments with ASR, and automatic segments with ASR.

We compare our results with a dialog-structure-ignorant lexical baseline: simply merge all text features (ngrams and patterns) from all the officer turns in a stop and use them to classify whether the stop did or didn't contain an act. Our goal here is to see whether dialog structure is useful for this task; if so, the tagger based on dialog turns should outperform the global text classifier.

Table 10 shows that using the output of the turn-based classifier to do stop classification offers a huge advantage over the structure-ignorant baseline, reducing F-score error by 49% while using manual transcripts, and by 22% while applied to raw audio.

	P	R	F
Manual (Lexical baseline)	79.6	77.6	78.6
Manual (Our Tagger)	**90.4**	**87.5**	**89.0**
ASR (Lexical baseline)	78.0	75.6	76.8
ASR (Our Tagger)	**86.5**	**81.7**	**84.0**
Raw Audio (Lexical baseline)	79.6	71.4	75.2
Raw Audio (Our Tagger)	**85.5**	**77.1**	**81.1**

Table 10: Stop level institutional act detection using our tagger, compared to a lexical baseline model trained on all the words spoken by the officer, without accounting for the dialog structure.

Text source	Manual	ASR	ASR
Segmentation source	Manual	Manual	Auto
Turn level	77.5	65.3	59.8
Stop level	89.0	84.0	81.1

Table 11: Summary: Micro-averaged F-scores across different text/segmentation sources.

Table 9 and Table 11 summarize the different experiments presented in Sections 4-8. Table 9 breaks down performance for each of the 11 acts, while Table 11 compares turn-level to stop-level results.

Despite our relatively small training resources (113 stops with dialog act labels, ASR and segmentation training data from one month), performance at the stop level directly from raw audio is surprisingly high. For instance, detecting whether or not the community member was explained the reason they were stopped—an important question for procedural justice—we obtained around 96% precision with an 84% recall from raw camera audio.

9 Conversation Trajectories

The institutional acts that happen during a traffic stop, when they occur, and in what order are all of importance to police departments. For instance, the President's Task Force on 21st Century Policing recommends (and some departments require) that officers identify themselves and state the reason for the stop as an important aspect of fairness. However,

[13] We use the best system from each set of experiments: SVM model using ngrams, patterns, and structure features trained on manual transcripts or from the BLSTM ASR model.

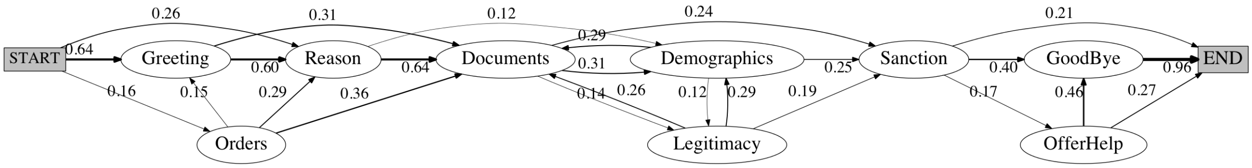

Figure 4: Prototypical conversation structure of traffic stops; transition probabilities based on 900 stops from Apr ’14.

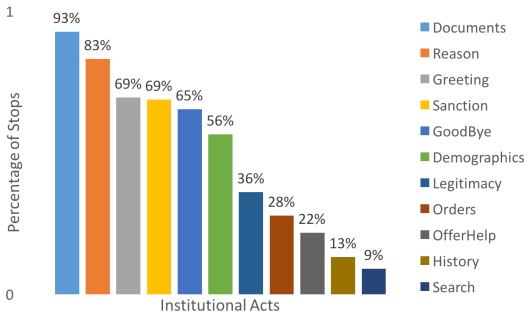

Figure 5: Presence of institutional acts in the 900 stops of black or white drivers from the month of April 2014.

police departments currently have no way of easily measuring how consistently such policies are carried out during traffic stops. They also have no way to test the effectiveness of any training programs or policy updates that are meant to affect these conversations.

In this section, we demonstrate that our institutional act tagger provides an efficient and reliable tool for departments to detect and monitor conversational patterns during traffic stops. Specifically, we focus on conversational openings, a fundamental aspect of conversations (Schegloff and Sacks, 1973) that is also important for procedural justice (Whalen and Zimmerman, 1987; Ramsey and Robinson, 2015). For instance, do officers start the conversations with a greeting? Are the drivers told the reason why they were stopped? Was the reason given before or after asking for their documentation?

We first apply our high performance (78% F-score at turn level; 89% at stop level) tagging model on manual transcripts. Figure 5 shows the percentage of stops made in which each of the eleven institutional acts was present. Around 17% of stops did not provide a reason at all. Only 69% of the stops started with a greeting, and an even smaller percentage of stops ended with a positive closing. While these high level statistics provide a window into these conversations, our institutional event tagger allows us to gain deeper perspectives.

Using the turn-level tags assigned by our system, we calculate the transition probabilities between dialog acts. Figure 4 shows a traffic stop ‘narrative schema’ or script, extracted from the high probability transitions. Variations from this prototypical script can be a useful tool for police departments to study how police community interactions differ across different squads, city locations, or driver characteristics like race.

Figure 6, for example, shows different conversational paths that officers take before explaining the reason for the stop. In over a quarter of the stops, either the reason is not given, or it is given after issuing orders or requesting documents. These violations of policing recommendations or requirements can impact the drivers’ attitude and perception of the legitimacy of the institution.

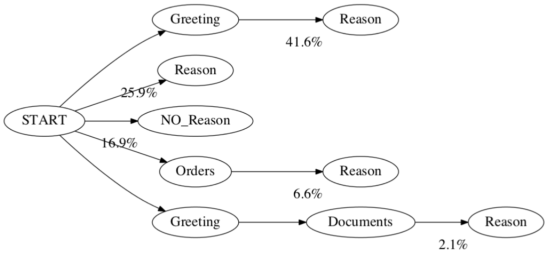

Figure 6: Conversational Paths to Giving Reason.

10 Discussion

In this section, we outline some of the limitations of our work and discuss future directions of research.

First, our work is based on data from a single police department (the Oakland Police Department in the State of California) in the U.S. The schema we developed may need to be updated for it to be applicable to other police departments; especially those in other countries, where the laws, policies and culture around policing may be significantly different.

Due to the sensitive nature of the data, we will not be able to publicly release the raw annotations described in Section 3.4. However, we will release the labeling scheme for institutional acts in traffic stops, along with the annotation manual. We believe that it will serve as a starting point for future researchers working in this domain.

Like any data-oriented approach, our machine learning models may have captured the idiosyncrasies of the particular department represented in our dataset. Since we are not aware of any other police departments' body-worn camera footage that is available for research, we have no way to guarantee that our models are directly applicable to other police departments' data.

Our institutional act tagger enables us to perform large scale social science analyses controlling for various confounds, which is infeasible to perform using hand-labeled data. However, although our models obtain high performance in detecting individual institutional acts, it may also capture biases that exist in the data (Hopkins and King, 2010). Hence, our models should be corrected for biases before they may be used to estimate proportions in any category of stops.

In this paper, we focus on officers' speech alone, since the conversational initiative with respect to the institutional acts lies mostly with the officer. However, drivers' speech may also need to be taken into account sometimes; e.g., if an officer says *yes* to a driver's question *did you stop me for running the red light?*, the officer has in fact given the reason for the stop even though their words alone don't convey that fact. Moreover, drivers' speech may also contribute to how the conversations are shaped. However, since the camera is further away from the driver than the officer, and since the environment is noisy, the audio quality of drivers' speech is poor, and further work is required to extract useful information from driver's speech. This is an important line of future work.

The video information from the body-camera footage may potentially help in the diarization and segmentation tasks, and in analyzing the effects the institutional acts have on the driver. However, since many of the stops occur at night when the video is often dark, it is not straightforward to extract useful information from them. This is another direction of future work.

11 Conclusion

In this paper, we developed a typology of institutional dialog acts to model the structure of police officer interactions with drivers in traffic stops. It enables a fine-grained and contextualized analysis of dialog structure that generic dialog acts fail to provide. We built supervised taggers for detecting these institutional dialog acts from interactions captured on police body-worn cameras, achieving around 78% F-score at the turn level and 89% F-score at the stop level. Our tagger detects institutional acts at the stop level directly from raw body-camera audio with 81% F-score, with even higher accuracy on important acts like giving the reason for a stop. Finally, we use our institutional act tagger on one month's worth of stops to extract insights about the frequency and order in which these acts occur.

The strains on police-community relations in the U.S. make it ever more important to develop insights into how conversations between police and community members are shaped. Until now, we have not had a reliable way to understand the dynamics of these stops. In this paper, we present a novel way to look at these conversations and gain actionable insights into their structure. Being able to automatically extract this information directly from raw body-worn camera footage holds immense potential not only for police departments, but also for policy makers and the general public alike to understand and improve this ubiquitous institutional practice.

The core contribution of this paper is a technical one of detecting institutional acts in the domain of traffic stops, from text and from unstructured audio files extracted from raw body-worn camera footage. Current work aims to improve the performance of the segmentation and diarization components, with the hope of reducing some of the performance gap with our system run on gold transcripts. We also plan to extend the preliminary analyses we describe in Section 9, for instance, studying how the different conversational paths and the presence or absence of certain acts (such as greetings or reason) shapes the rest of the conversation, including how it changes the community member's language use. Finally, our model allows us to study whether police training has an effect on the kinds of conversations that police officers have with the communities they serve.

Acknowledgments

We thank the anonymous reviewers as well as the action editor, Jordan Boyd-Graber, for helpful feedback on an earlier draft of this paper. This research was supported by a John D. and Catherine T. MacArthur Foundation award granted to J.L. Eberhardt and D. Jurafsky, as well as NSF grants IIS-1514268 and IIS-1159679. We also thank the City of Oakland and the Oakland Police Department for their support and cooperation in this project.

References

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.

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 1061–1064. IEEE.

J. Maxwell Atkinson and Paul Drew. 1979. *Order in Court*. Springer.

John Langshaw Austin. 1975. *How To Do Things With Words*. Oxford University Press.

Srinivas Bangalore, Giuseppe Di Fabbrizio, and Amanda Stent. 2006. Learning the structure of task-driven human-human dialogs. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 201–208. Association for Computational Linguistics.

David M. Blei and Pedro J. Moreno. 2001. Topic segmentation with an aspect hidden Markov model. In *Proceedings of the 24th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 343–348. ACM.

Rod K. Brunson. 2007. "Police Don't Like Black people": African-American Young Men's Accumulated Police Experiences. *Criminology & Public Policy*, 6(1):71–101.

Muthu Kumar Chandrasekaran, Carrie Epp, Min-Yen Kan, and Diane Litman. 2017. Using discourse signals for robust instructor intervention prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher corpus: A resource for the next generations of speech-to-text. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA).

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 4, pages 309–316. Association for Computational Linguistics.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Justine Coupland, Nikolas Coupland, and Howard Giles. 1991. Accommodation theory, communication, context and consequences. *Contexts of Accommodation*, pages 1–68.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 334–343. Association for Computational Linguistics.

Robin S. Engel. 2005. Citizens' perceptions of distributive and procedural injustice during traffic stops with police. *Journal of Research in Crime and Delinquency*, 42(4):445–481.

Charles R. Epp, Steven Maynard-Moody, and Donald P. Haider-Markel. 2014. *Pulled Over: How Police Stops Define Race and Citizenship*. University of Chicago Press.

Mark J. F. Gales. 1998. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, 12(2):75–98.

Felix A. Gers, Nicol N. Schraudolph, and Jürgen Schmidhuber. 2002. Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research*, 3(Aug):115–143.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Howard Giles, Jennifer Fortman, René Dailey, Valerie Barker, Christopher Hajek, Michelle Chernikoff Anderson, and Nicholas O. Rule. 2006. Communication accommodation: Law enforcement and the public. *Applied Interpersonal Communication Matters: Family, Health, and Community Relations*, 5:241–269.

Howard Giles, Christopher Hajek, Valerie Barker, Mei-Chen Lin, Yan Bing Zhang, Mary Lee Hummert, and Michelle C. Anderson. 2007. Accommodation and institutional talk: Communicative dimensions of policecivilian interactions. In *Language, Discourse and Social Psychology*, pages 131–159. Springer.

Augusto Gnisci. 2005. Sequential strategies of accommodation: A new method in courtroom. *British Journal of Social Psychology*, 44(4):621–643.

John J. Godfrey and Edward Holliman. 1997. Switchboard-1 release 2. *Linguistic Data Consortium, Philadelphia*, 926:927.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning*, pages 1764–1772.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6645–6649. IEEE.

Barbara J. Grosz. 1977. The representation and use of focus in dialogue understanding. Technical report, SRI International Menlo Park United States.

John Heritage. 2005. Conversation analysis and institutional talk. *Handbook of Language and Social Interaction*, pages 103–147.

Daniel J. Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 88–95. IEEE.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2012–2021.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2010. Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 862–871. Association for Computational Linguistics.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2012. Classifying dialogue acts in multi-party live chats. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 463–472.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for M-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184. IEEE.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proceedings of Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3586–3589.

Lynn Langton and Matthew R. Durose. 2013. *Police behavior during traffic and street stops, 2011*. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics Washington, DC.

Wei Li and Yunfang Wu. 2016. Multi-level gated recurrent neural network for dialog act classification. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical papers*, pages 1970–1979.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in DNN framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178.

Richard J. Lundman and Robert L. Kaufman. 2003. Driving while black: Effects of race, ethnicity, and gender on citizen self-reports of traffic stops and police actions. *Criminology*, 41(1):195–220.

Abdel-rahman Mohamed, Frank Seide, Dong Yu, Jasha Droppo, Andreas Stoicke, Geoffrey Zweig, and Gerald Penn. 2015. Deep bi-directional recurrent networks over spectral windows. In *Proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 78–83. IEEE.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2012. SITS: A hierarchical nonparametric model using speaker identity for topic segmentation in multiparty conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 78–87. Association for Computational Linguistics.

Adinoyi Omuya, Vinodkumar Prabhakaran, and Owen Rambow. 2013. Improving the quality of minority class identification in dialog act tagging. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 802–807.

Michael J. Paul. 2012. Mixed membership Markov models for unsupervised conversation modeling. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 94–104. Association for Computational Linguistics.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012. Predicting overt display of power in written dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–522. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Emily E. Reid, and Owen Rambow. 2014. Gender and power: How gender and gender environment affect manifestations of power. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1965–1976. Association for Computational Linguistics.

Charles H. Ramsey and Laurie O. Robinson. 2015. Final report of the President's task force on 21st century policing. *Washington, DC: Office of Community Oriented Policing Services*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180. Association for Computational Linguistics.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735.

Haşim Sak, Andrew Senior, and Françoise Beaufays. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. 2007. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning*, pages 791–798. ACM.

Emanuel A. Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8(4):289–327.

Emanuel A. Schegloff. 1979. Identification and recognition in telephone conversation openings. *Everyday-Language: Studies in Ethnomethodology, New York, Irvington*, pages 23–78.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2006. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Dialogue*, 26(3).

Prateek Verma, T. P. Vinutha, Parthe Pandit, and Preeti Rao. 2015. Structural segmentation of Hindustani concert audio with posterior features. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 136–140. IEEE.

Karel Veselý, Mirko Hannemann, and Lukas Burget. 2013. Semi-supervised training of deep neural networks. In *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE.

Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526.

Marilyn R. Whalen and Don H. Zimmerman. 1987. Sequential and institutional contexts in calls for help. *Social Psychology Quarterly*, pages 172–185.

Gethin Williams and Daniel P.W. Ellis. 1999. Speech/music discrimination based on posterior probability features. In *Proceedings of the Sixth European Conference on Speech Communication and Technology*.

Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn P. Rose. 2013. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-Driven Education Workshop*, volume 10, pages 13–20.