Please Clap: Modeling Applause in Campaign Speeches

Jon Gillick
School of Information
University of California, Berkeley
jongillick@berkeley.edu

David Bamman
School of Information
University of California, Berkeley
dbamman@berkeley.edu

Abstract

This work examines the rhetorical techniques that speakers employ during political campaigns. We introduce a new corpus of speeches from campaign events in the months leading up to the 2016 U.S. presidential election and develop new models for predicting moments of audience applause. In contrast to existing datasets, we tackle the challenge of working with transcripts that derive from uncorrected closed captioning, using associated audio recordings to automatically extract and align labels for instances of audience applause. In prediction experiments, we find that lexical features carry the most information, but that a variety of features are predictive, including prosody, long-term contextual dependencies, and theoretically motivated features designed to capture rhetorical techniques.

1 Introduction

Every public speech involving a large audience can be seen as a game of coordination (Asch, 1951): at each moment, each individual member of the audience must decide in a split second whether to applaud at what has just been said. Applause is a potentially risky action: if an individual spontaneously claps but no one joins in, they suffer some negative social cost; the game is to judge from their own private information and content of the speech whether the rest of the audience will applaud at the same time they do.

Because of this cost, audiences respond to several interacting factors in a speaker's behavior: a.) the content of the message; b.) their delivery (so that changes in pitch, duration and gaze signal salient moments for which applause may be licensed); and c.) the verbal *design* of the message—those rhetorical strategies that speakers use to signal that applause is welcome (Atkinson, 1984; Heritage and Greatbatch, 1986).

In this work, we attempt to model all three of these dimensions in developing a computational model for applause. While past work has focused on these elements in isolation (Guerini et al., 2015; Liu et al., 2017) or for related problems such as laughter detection (Purandare and Litman, 2006; Chen and Lee, 2017; Bertero and Fung, 2016), we find that developing a holistic model encompassing all three aspects yields the most robust predictor of applause.

We focus on political speeches, and in particular those at campaign rallies, which lend themselves well to analysis of rhetorical strategies for several reasons. First, the speakers at these events prioritize maintaining the crowd's attention (Strangert, 2005). Motivated to drum up excitement and fervor among their supporters that they hope will carry beyond the event and into the voting booth, speakers pull out their strongest rhetorical tactics. Second, campaign speeches usually consist of a series of self-contained messages that can be fully expressed within a few utterances (Heritage and Greatbatch, 1986), yielding a well-defined observation of a complete rhetorical strategy. Lastly, these speeches are delivered by a single speaker to a partisan crowd, and clapping, cheering, and other responses are invited and expected.

We focus in particular in this work on operationalizating the verbal design of the speech; in so doing, one contribution we make is operationalizing the concepts of *tension* and *release*. Writers and performers often communicate with their audience on a fundamental level by building up tension, and then, at the proper time, delivering a satisfying release. These simple but pervasive concepts structure our experience of different modes of communication used throughout everyday life, including music (Madsen and Fredrickson, 1993), literature (Rabkin, 1973) and film (Carroll, 1996).

Tension in music can be built up by harmonic

Proceedings of NAACL-HLT 2018, pages 92–102
New Orleans, Louisiana, June 1 - 6, 2018.

movement away from a tonal center; release then comes with a return to that established tonic (Hindemith, 1937). One form of tension in literature is realized as suspense (Barthes and Duisit, 1975; Vorderer et al., 1996; Algee-Hewitt, 2016), in which a reader's knowledge of events is uncertain (either because those events take place in the narrative future or are withheld from narration), and released when that knowledge is revealed. In film, sudden changes in camera perspective create graphic tension, which is then released as the shot returns to a stable position (Bordwell, 2013). Often, it is the confluence of multiple sources of tension that mark the climax of a narrative (Hume, 2017). We draw on each of these strands of work in operationalizing tension and release as a rhetorical strategy.

In this work, we make the following contributions:

- We collect a new dataset of text and audio from 310 speeches from campaign events leading up to the 2016 U.S presidential election with associated tags for over 19,000 instances of audience applause.
- We introduce new textual and acoustic features inspired by tension and release, combine and compare them with features used in previous work, and deploy those features in a logistic regression model and in an LSTM to predict when applause is likely to occur. Code, data, and trained models are openly available to the public at `https://github.com/jrgillick/Applause/`.

2 Background and Previous Work

2.1 Rhetoric and Response

Heritage and Greatbatch (1986) conduct an extensive analysis of nearly 500 speeches from British political party conferences, manually associating each of over 2000 instances of applause with coded message types (e.g. External Attacks or Statements of Approval), rhetorical devices (e.g. Contrast/Antithesis or Headline-Punchline), and performance factors (e.g. speech stress or body language). They find most of these factors to be positively correlated with applause; one especially striking result is over two thirds of observed instances of applause can be explained through a set of seven rhetorical devices (including contrast, pursuit, position taking, and "the 3-part list"). Though each device is different, a common feature of most of these techniques is that they are not always carried out within a single sentence or utterance; they often depend on the relationship between a series of utterances or phrases. We argue in this work that some of these relationships can be characterized and subsequently operationalized within models as tension and release.

2.2 Predicting Applause

Recent work from Guerini et al. (2015) and Liu et al. (2017) approaches the task of applause prediction by looking at textual features of the individual sentences that immediately precede audience applause. Both follow the methodology proposed by Danescu-Niculescu-Mizil et al. (2012) in constructing a data set for binary classification, which is composed of sentences that generated applause, each paired with a single nearby sentence from the same document that did not lead to applause.

Guerini et al. (2015) examine a set of features designed to capture aspects of euphony, or "the inherent pleasantness of the sounds of words" that might make an utterance memorable or persuasive—such as rhyme, alliteration, homogeneity, and plosives. On the CORPS dataset (Guerini et al., 2013), which consists of the text of several thousand political speeches dating from 1917 to 2011, they define persuasive sentences as those that preceded annotations of either applause or laughter.

Liu et al. (2017), working with a corpus of TED talks, use logistic regression to predict applause from sentences using a combination of features: euphony (again from Guerini et al. (2015)), linguistic style markers derived from membership in LIWC categories, markers of emotional expression derived from membership in the NRC Emotion Lexicon, mentions of names, rhetorical questions (string matching for "?"), expressions of gratitude (matching a handcrafted list of word stems including "thank*" and "grateful*"), and expressions seeking applause (matching the pattern "applau*"). Liu et al. (2017) also report that adding the same features for earlier sentences beyond the final sentence that preceded the applause caused the prediction accuracy to go down. Chen and Lee (2017) and Bertero and Fung (2016) run similar binary classification experiments but pre-

dict laughter as opposed to applause. Bertero and Fung (2016) analyze punchlines from the TV sitcom "The Big Bang Theory" and report 70% accuracy using an LSTM. They touch briefly on the notion of tension and release in humor, as punchlines typically depend on a previous line as a setup in order to be funny.

3 Data

3.1 Corpus Acquisition

In this work, we focus on a new data set of campaign speeches from the 2016 U.S. presidential race, which we obtain from the public domain broadcasts of C-SPAN. We downloaded about 500 speeches from presidential candidates, vice presidential candidates, or former presidents, collecting audio files and transcripts that were tagged in the categories "Campaign 2016" and "Speech" and which took place between 12/01/2015 and 12/01/2016. We then excluded events that took place outside of a traditional campaign speech setting (e.g. town hall events) or events that contained multiple speakers without a speaker identification tied to the transcript, which yielded a final set of 310 speeches from 16 speakers. Because different types of events have different social norms around when and whether applause is appropriate (Atkinson, 1984; Heritage and Greatbatch, 1986), we control for these factors to some degree by restricting our dataset to events in similar settings and within a single year. As a point of comparison, the C-SPAN dataset contains 62 instances of applause per speech on average, whereas the CORPS data (Guerini et al., 2013) contains 13.

3.2 Applause Detection in Audio

Since our C-SPAN data originates in video, we have access to the audio information of a speech event, which we employ both for feature extraction and for automatically identifying when applause occurs. Following Clement and McLaughlin (2016), we train an acoustic model using a set of poetry readings from the PennSound archive to distinguish applause from speech. We used logistic regression on the standard set of MFCC features and found similar results on the PennSound data to the reported classification accuracy of 99.4%. In a manual inspection of 100 applause segments from 5 different speeches in the C-SPAN corpus, our applause detector achieved 92% precision, 90% recall, and 91% F1 score. Due to variation in the nature of applause in a crowd (sometimes we observe examples of isolated clapping and cheering, mixed laughter and applause, or applause interrupting the speaker), some ambiguity is inherent among the labels.

We also measure the applause by first running the speeches through the audio source separation algorithm from Chandna et al. (2017), which was trained to separate voice from music, and then measuring the RMSE loudness of the separated non-vocal track. We found that the separation worked well, qualitatively matching with the results from the applause detection classifier.

3.3 Forced Alignment

To match the identified segments of applause in the audio files with the relevant text from the transcriptions, we ran forced alignment using the Kaldi Toolkit (Povey et al., 2011). Since the C-SPAN transcripts are sourced from uncorrected closed captioning, the text contains a number of misspellings and paraphrases, which we handled by discarding the 12% of words for which forced alignment failed. Though these transcriptions are not as accurate as what we would find in professionally transcribed datasets, previous work has shown that it is possible to achieve good accuracy in downstream tasks even with high error rates in transcription (Peskin et al., 1993; Novotney and Callison-Burch, 2010). Moreover, the caliber of transcripts derived from closed captioning is representative of the data that would be available in real time for practical use at future speech events.

To estimate the accuracy of the closed captions, we manually transcribed selections from 5 speeches in the C-SPAN data totaling about 25 minutes and 2250 words, finding 30.9% WER relative to the reference transcriptions in our sample. Many of the errors are due to omitted words and phrases in the closed captions, which may occur as a result of transcribers' inability to keep up with the pace of fast speeches; in this sample, the closed caption texts contained 17% fewer words than our gold standard transcriptions.

After finding the alignments, we segmented out a list of utterances by defining a minimum period of silence between words. Since many of the transcripts do not have punctuation, we find that dividing the text into utterances yielded qualitatively more coherent units than sentence boundary detec-

Speaker	Number of Speeches	Number of Utterances	Number Applauded	Percentage
Donald Trump	86	27493	7357	0.27
Hilary Clinton	72	12825	3933	0.31
Bernie Sanders	40	10994	3529	0.32
Ted Cruz	23	5873	1041	0.18
Marco Rubio	20	4407	797	0.18
John Kasich	17	4023	319	0.08
Barack Obama	10	3888	920	0.24
Bill Clinton	8	2087	292	0.14
Joe Biden	7	1847	270	0.15
Mike Pence	6	1302	246	0.19
Carly Fiorina	5	1222	129	0.11
Jeb Bush	5	1482	191	0.13
Rand Paul	4	939	134	0.14
Gary Johnson	3	354	56	0.16
Chris Christie	3	1868	42	0.022
Rick Santorum	1	245	17	0.07
Total	**310**	**80849**	**19273**	**0.24**

Table 1: Speakers and applause in C-SPAN corpus

tion. Dividing into utterances is also conducive to building a dataset for binary classification, since every pause by the speaker yields an opportunity for applause. We chose a pause length of 0.7 seconds, but in future work we might be able to improve our models by adapting this threshold to the rate of speech in order to maintain consistent phrase sizes across different speakers. Given this set of utterances, we paired each utterance with a "positive" or "negative" label, determined by whether applause occurred within 1.5 seconds of the end of the utterance. All of these preprocessing choices were made during the corpus preparation phase, prior to any experimental evaluation.

Table 1 provides summary statistics for the number of speakers, speeches, utterances, and acts of applause in our data.

4 Models

In our models, we draw features from previous work on applause or humor prediction and then supplement them with a new set of features inspired by the ideas of tension and release and by the rhetorical strategies of Heritage and Greatbatch (1986).

4.1 Features adapted from existing work

LIWC. Features for membership in 73 LIWC categories proved to be the most effective for applause prediction in TED talks (Liu et al., 2017).

Euphony. We adopt the 4 features for "euphony" defined by Guerini et al. (2015): Rhyme, Alliteration, Homogeneity, and Plosives.

Lexical. Guerini et al. (2015) find n-grams to be highly predictive of both applause and laughter. We operationalize these features with bigrams, including in our model all bigrams that appear at least 5 times in the corpus.

Embeddings. Bertero and Fung (2016) use sentence embeddings learned from a CNN encoder as input to an LSTM. We adopt this feature for use in our neural models, encoding phrases using the Skip-Thought model of Kiros et al. (2015).

Acoustic. Purandare and Litman (2006) use a set of features intended to capture elements of prosody in a model for humor prediction in television dialogue. These features include the mean, max, min, range, and standard deviation values in an utterance's pitch (F0) and energy (RMS), along with features for internal silence and for tempo. We compute the F0 statistics with Reaper (Talkin, 2015) and the energy statistics with Librosa (McFee et al., 2015).

4.2 New Features

4.2.1 Repetition

Repeated Words. Rhetorical strategies such as "The 3-part List" and "Contrast" rely on repetition to drive home important points. We capture this phenomenon by computing the proportion of words in each utterance that also appear in the immediately preceding phrase.

Longest Common Subsequence. Repeating an entire phrase, especially one with a politically charged topic, serves to build tension through the notion of "theme and variation" as is often realized

in music (Cope, 2005); an example of this phenomenon in our data can be found in the following passage:

> *We will not allow the* party of Lincoln and Reagan to fall into the hands of a con artist. *We will not allow the* next president of the United States to be a socialist like Bernie Sanders. And *we will not allow the* next president of the United States to be someone under FBI investigation like Hillary Clinton.
>
> [Marco Rubio, Mar. 1, 2016]

We calculate this theme and variation by measuring the longest common subsequence between adjacent phrases.

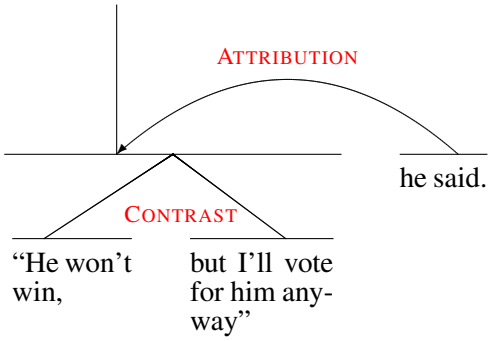
4.2.2 Deltas

Delta features (local approximations to derivatives) are commonly used in speech recognition and audio classification systems (Povey et al., 2011). In a discourse, either highly similar or drastically different neighboring pairs of utterances may indicate dramatic moments. We operationalize these features by explicitly adding a delta measurement for every feature in our model, which captures the difference between every feature at time t and the same feature at time $t-1$. For K-dimensional vector embeddings, we calculate deltas as their cosine distance.

4.2.3 RST

Rhetorical Structure Theory (RST) provides a foundation for describing the ways in which functional components of a text combine to form a coherent whole (Thompson and Mann, 1987). At the core of RST is a categorization system consisting of relations between elementary discourse units (EDUs). Relations between units are typically hierarchical (a nucleus and a satellite), but can also be defined between equally significant units (two nuclei).

A typical RST tree can be seen below, where the sentence *"He won't win, but I'll vote for him anyway", he said* is decomposed into three elementary discourse units (EDUs); those discourse units form the leaves of a tree with intermediate structure between subphrases and labeled edges along each branch.



Some of the rhetorical strategies defined by Heritage and Greatbatch (1986), such as "Contrast," map directly to RST relations, while others do not have a clear one-to-one mapping but are qualitatively similar in their descriptions. While RST has been used with success for classification problems in the past (Ji and Smith, 2017; Bhatia et al., 2015), it has not yet been employed in existing models for applause prediction. In our work, we parse the rhetorical structure of the extracted sequence of phrases using the RST parser of Ji and Eisenstein (2014). From the structure of this RST tree, we extract two classes of features.

RST label. First, we operationalize the rhetorical category for an individual elementary discourse unit. While the span of text within a single EDU is implicated in several rhetorical relations throughout the tree (as *He won't win* bears a CONTRAST relationship with *but I'll vote for him anyway* and is part of the ATTRIBUTION relationship with *he said*), each EDU bears exactly one leaf relationship with the rest of the tree—here, *He won't win* is a nucleus of a CONTRAST relationship, *but I'll vote for him anyway* is also a nucleus of a CONTRAST relationship, and *he said* is the satellite of an ATTRIBUTION relationship.

We featurize a sentence as the set of all such typed relationships that EDUs within it hold; each typed relationship is the conjunction of the label (e.g., CONTRAST, ATTRIBUTION) and directionality (Nucleus, Satellite).

Rhetorical phrase closures. In order to further operationalize the notion of predictability of applause, we measure the number of rhetorical phrases that a given discourse segment brings to *closure*. We can illustrate this with figure 1, which presents a sample RST tree with only the spans annotated (i.e., without RST labels or nucleus/satellite directed edges). This tree spans 10 elementary discourse units; each non-terminal node is annotated with the span of the subtree

rooted at that node (so the root spans all ten EDUs, while its left child spans only the first five). The final discourse unit (EDU 10) is the final EDU in three rhetorical phrases (those spanning EDUs 9-10, 6-10 and the entire discourse 1-10). We might hypothesize that the greater number of discourse phrases that a given discourse unit closes, the stronger the signal it provides that applause is licensed (and hence the greater likelihood to be followed by applause empirically). For a sentence with multiple discourse units, we featurize this value as the maximum number of rhetorical phrases closed by any unit it contains.

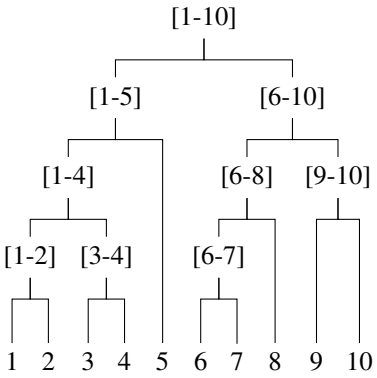


Figure 1: Unlabeled RST phrase tree; non-terminal nodes list the ranges of the elementary discourse units they span.

5 Experiments

We present two experiments to uncover the degree to which we are able to predict applause from different operationalizations of a politician's campaign speech: one in which have access to a politician's previous speeches, and can learn their specific nuances and stock phrases used to solicit applause; and another in which we seek to uncover the broader rhetorical strategies common to multiple speakers.

We refer to the following sets of features when we summarize results:

- **Guerini.** Euphony features from Guerini et al. (2015).
- **Liu.** LIWC features and additional matchers for handcrafted regular expressions from Liu et al. (2017)
- **Audio.** All acoustic features described in §4.1 above.
- **Combined.** Combination of features from Guerini, Liu, and Audio.
- **Tension.** Combination of RST (§4.2.3), repetition (§4.2.1), and delta features (§4.2.2).
- **N-gram.** Bigram features.
- **Skip-Thought.** 4800 dimensional Skip-Thought embeddings.

5.1 Intra-speaker validation

Access to a politician's previous speeches provides a great deal of evidence for understanding their rhetorical strategies for soliciting applause; speakers often give variations of the same speech at different campaign events, and rely on a fixed set of stock phrases (e.g., "Yes, We Can," "Make America Great Again") and general strategies to solicit reactions (Lu, 1999; Miller, 1939; Petrow and Sullivan, 2007). To model this, we attempt to predict a speaker's likelihood of applause using only information from their own speeches.

We use logistic regression with ℓ_2 regularization for this experiment, with hyperparameters chosen through cross-validation on the training data. We run 10-fold cross validation for each speaker, and leave-one-out cross validation for those speakers with fewer than 10 speeches (we exclude Rick Santorum from this experiment because we have only one speech from him), with whole speeches divided across folds so that no utterances from the same speech ever appear in both training and test sets. Reported results aggregate the predictions across all speakers to calculate the final accuracies. We choose utterances (or sequences of utterances) that directly precede applause as positive examples, pairing each one with a negative example randomly chosen from the same speech. Since we use different amounts of data for each speaker, we are not able to compare accuracies across all speakers, but we can see that some speakers are significantly easier to model: for example, our best model reaches 0.719 accuracy on Bernie Sanders but only 0.660 on Donald Trump.

Table 2 summarizes the results, comparing across different combinations of features as well as across a scope of a single phrase or multiple phrases. All feature combinations are scoped over a single utterance unless otherwise noted.

5.2 Inter-speaker validation

At the same time, many of the strategies identified by Heritage and Greatbatch (1986) are gener-

Model	Mean Accuracy	Mean F1	Max F1	Min F1
Guerini	0.566	0.533	0.659 (Bernie Sanders)	0.422 (Donald Trump)
Liu	0.601	0.594	0.649 (Bernie Sanders)	0.499 (Jeb Bush)
Audio	0.598	0.574	0.634 (Hilary Clinton)	0.516 (Donald Trump)
Combined	0.646	0.640	0.685 (Bernie Sanders)	0.598 (Marco Rubio)
N-gram	0.637	0.578	0.672 (Bernie Sanders)	0.478 (Barack Obama)
Combined+Tension	0.639	0.635	0.682 (Bernie Sanders)	0.585 (Jeb Bush)
Combined (3-Phrase)	0.645	0.640	0.671 (Bernie Sanders)	0.587 (Bill Clinton)
Combined+Tension (3-Phrase)	0.626	0.624	0.665 (Bernie Sanders)	0.602 (Marco Rubio)
Combined+N-gram	0.673	0.661	0.711 (Bernie Sanders)	0.600 (Marco Rubio)
Combined+Tension+N-gram	0.671	0.658	0.711 (Bernie Sanders)	0.599 (Marco Rubio)

Table 2: Intra-speaker predictive accuracy (logistic regression). The 95% confidence interval for Mean Accuracy and Mean F1 is within $\pm$ 0.005, and the 95% confidence interval for Max F1 and Min F1 (1 speaker at a time) is within $\pm$ 0.05.

alized rhetorical devices used to solicit applause; we should expect then that a model trained on a fixed set of speakers should be able to generalize to speakers not in the training data. To test this more realistic scenario, we performed K-fold cross-validation on all of the speakers in our dataset, holding out one speaker in turn for each fold (so that the same speaker did not appear in the training and test partitions).

In this experiment, we use both logistic regression and neural models (sharing training data between speakers has the added benefit of allowing us enough data to reasonably train a neural model). All logistic regression models were trained in the same way is in the intra-speaker case. Our feed-forward and LSTM models use a hidden state size of 100 for models including phrase embeddings (4800 dimensions) and a hidden state of size 25 for models without phrase embeddings. All LSTM models use a standard formulation of attention (Bahdanau et al., 2014), and all neural models are trained with dropout (Srivastava et al., 2014) and the ADAM optimizer (Kingma and Ba, 2014). We implemented the models using Keras (Chollet et al., 2015) and Tensorflow (Abadi et al., 2016).

Table 3 summarizes these results, and table 4 shows the coefficients for the most significant features.

Logistic Regression Models	Acc.	F1
Guerini	0.557	0.534
Liu	0.577	0.541
Audio	0.573	0.548
Combined	0.615	0.601
N-gram	0.594	0.578
Combined+Tension	0.617	0.605
Combined (3-Phrase)	0.614	0.601
Combined+Tension (3-Phrase)	0.615	0.600
Combined+N-gram	0.633	0.598
Combined+Tension+N-gram	0.630	0.594
Neural Models	**Acc.**	**F1**
Feed-Forward:Skip-Thought	0.577	0.562
Feed-Forward:Combined+Tension	0.620	0.620
LSTM:Skip-Thought(3-Phrase)	0.585	0.583
LSTM:Combined+Tension(3-Phrase)	0.626	0.616
LSTM:Combined+Tension(5-Phrase)	0.628	0.625
LSTM:Combined+Tension(8-Phrase)	0.629	0.621

Table 3: Inter-speaker predictive accuracy. The 95% confidence interval for each measurement of accuracy is within $\pm$ 0.005.

6 Analysis

Each of the feature classes we operationalize offers some ability to recognize what Heritage and Greatbatch (1986) term the "projectability" of applause—the ability of an audience to see an applaudable moment on the horizon.

Audio. Perhaps not surprising in retrospect is the ability of acoustic features (only summary statistics of the pitch and energy) to solicit applause: higher pitch and energy, and a broader pitch range are all predictive of applause; while past work has focused on textual indicators of applause, these results suggest that *how* a message is delivered is equally important.

Lexical. The use of explicit n-grams improves performance significantly in the intra-speaker setting, where they are able to capture stock phrases employed by the same speaker at different events. N-grams are also predictive across different speakers, though the performance gains are not as high in the inter-speaker setting.

The strongest bigrams predictive of applause include moral declaratives like *should not* (e.g., "and billionaires **should not** be able to buy elections" [Bernie Sanders]), *right to* ("you have a **right to** be angry" [Marco Rubio]), and *should be* ("They should be ashamed of that kind of behavior" [Hillary Clinton]); call-outs to the audience such as *this room* ("Love the people in **this room**"

Significant Features	Coefficient
Expression of Gratitude	0.472
LIWC FOCUSFUTURE	0.340
Homogeneity (Guerini)	0.301
Mean Energy (Audio)	0.293
LIWC BODY	0.203
Min Energy (Audio)	0.165
Max Pitch (Audio)	0.157
LIWC TENTATIVE	-0.161
LIWC THEY	-0.172
LIWC VERB	-0.216
LIWC FUNCTION	-0.228
Pitch Standard Deviation (Audio)	-0.249
LIWC SHEHE	-0.275
LIWC FOCUSPAST	-0.342

Table 4: Most significant positive and negative features for the Combined+Tension regression model in the inter-speaker setting.

[Donald Trump]) and *listening to* ("our campaign is **listening to** our Latino brothers and sisters" [Bernie Sanders]); and politically charged topics such as *political revolution*, *equal pay*, *immigration reform*, *planned parenthood*, *campaign contributors* and *police officers*.

LIWC. Among broader lexical category features, we see the LIWC FOCUSFUTURE category strongly indicative of applause; this category includes auxilaries like *will, going, gonna* (including conjunctions *I'll*) and future-oriented verbs like *anticipate*; also important are categories of BODY (including *heart, hands, brain*) and REWARD (including *succeed, optimism, great*).

Rhetorical. While RST features were not as predictive for applause as other (likely correlated) features, we still see a strong alignment between the RST features most associated with applause and those rhetorical devices outlined by Heritage and Greatbatch (1986): in particular, a clear relationship between applause and the RST category of ANTITHESIS (a contrastive relation between two discourse units with a clear nucleus and satellite, rather than two equal nuclei) and PURPOSE (a relation between a discourse unit that must take place in order for another to be realized). As expected, phrases that close more discourse units tend to be more predictive of applause.

Contextual. Though lexical features from the final utterance significantly outweigh the effects of previous context in the intra-speaker setting, in the inter-speaker case we leveraged gains from long-term context in the LSTM to reach a similar level of performance attained from the lexical features, but without access to lexical cues provided by the n-grams at all. This result suggests that the improved performance in the intra-speaker setting may be largely due to the presence of specific words and catch-phrases; the other stylistic features are more easily generalized to new speakers.

7 "Please clap"

As a further measure of out-of-sample validity, we can analyze the predictions we make for the single example where a speaker wears his communicative intent on his sleeve. On February 2, 2016, presidential candidate Jeb Bush spoke to a crowd in New Hampshire a week before their state primary. His speech ended with the following:

> So here's my pledge to you. [I] will be a commander-in-chief who will have the back of the military, I won't trash talk, I won't be a divider-in-chief or an agitator-in-chief, I won't be out there blowharding talking a big game without backing it up; I think the next President needs to be a lot quieter but send a signal that we're prepared to act in the national security interests of this country to get back in the business of creating a more peaceful world . . . . . Please clap.
>
> [Jeb Bush, Feb 2, 2016][1]

Bush's admonition to the audience ("please clap") earned criticism in news coverage at the time (Benen, 2016), but also presents us with a rare insight into a speaker's true rhetorical intention; in this case, Bush was soliciting applause and was vocal about not being able to do so.

Does our model recover this true intention? Indeed it does; while the opening *So here's my pledge to you* is predicted to not solicit applause (with applause probability of 24.8%), the segment that ends with *peaceful world* is strongly predicted to have been followed by applause (with an applause probability of 94.5%). The strongest features are again lexical (*this country*, *commander in chief*), a LIWC focus on the future (elicited by *will*), and an RST PURPOSE relation (evoked by *to get back in the business of creating a more peaceful world*).

[1] Video of this speech can be found at: https://www.youtube.com/watch?v=DdCYMvaUcrA

8 Conclusion

We present in this work a new dataset for the analysis of political rhetoric derived from the public campaign speeches of politicians during the 2016 United States presidential election, along with empirical results assessing the performance of different operationalizations of rhetoric derived from the theoretical work of Heritage and Greatbatch (1986) and others in order to measure and predict the occurrence of applause. We introduce several new features designed to capture elements of tension and release in public performance, including rhetorical contrast, closure, repetition and movement across speech segments; while each of these features in isolation is able to predict applause to varying degree and comport with our prior understanding of their utility, we find that lexicalized features are among the strongest source of information in determining applause; while audiences react to many dimensions of a speaker's style, the words they use—as slogan, stock phrases, and indicators of more complex rhetorical functions like moral valuations and imperatives—matter most.

As detailed in previous work (Liu et al., 2017; Haider et al., 2017; Clement and McLaughlin, 2016), understanding and identifying climactic moments in speeches can be useful for a variety of reasons, including learning to give better talks, automatically summarizing videos and transcripts, and analyzing social dynamics within crowds. One additional interesting application of this work is to bring to the surface occasions where a speaker uses typical applause-seeking devices but does not receive applause (the "Please Clap" moments); we leave to future work identifying the reverse, when speakers receive applause without invoking common techniques (for example, to identify instances of *claques* paid to clap).

9 Acknowledgments

Many thanks to the anonymous reviewers for their helpful feedback. The research reported in this article was supported by a UC Berkeley Fellowship for Graduate Study to J.G. and by resources provided by NVIDIA.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* .

Mark Algee-Hewitt. 2016. The machinery of suspense. http://markalgeehewitt.org/index.php/main-page/projects/the-machinery-of-suspense/.

S. E. Asch. 1951. Effects of group pressure on the modification and distortion of judgments. In H. Guetzkow, editor, *Groups, Leadership and Men*. Carnegie Press.

J. Maxwell Atkinson. 1984. Public speaking and audience responses: some techniques for inviting applause. In *Structures of Social Action*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Roland Barthes and Lionel Duisit. 1975. An introduction to the structural analysis of narrative. *New Literary History* 6(2):237–272. http://www.jstor.org/stable/468419.

Steve Benen. 2016. Jeb Bush urges audience, 'Please clap'. http://www.msnbc.com/rachel-maddow-show/jeb-bush-urges-audience-please-clap.

Dario Bertero and Pascale Fung. 2016. A long short-term memory framework for predicting humor in dialogues. In *HLT-NAACL*. pages 130–135.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. *arXiv preprint arXiv:1509.01599* .

David Bordwell. 2013. *Narration in the fiction film*. Routledge.

Noel Carroll. 1996. Toward a theory of film suspense. In *Theorizing the Moving Image*. Cambridge University Press.

Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. 2017. Monoaural audio source separation using deep convolutional neural networks. In *International Conference on Latent Variable Analysis and Signal Separation*. Springer, pages 258–266.

Lei Chen and Chong Min Lee. 2017. Predicting audience's laughter during presentations using convolutional neural network. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. pages 86–90.

François Chollet et al. 2015. Keras.

Tanya Clement and Stephen McLaughlin. 2016. Measured applause: Toward a cultural analysis of audio collections. *Journal of Cultural Analytics* .

David Cope. 2005. *Computer models of musical creativity*. MIT Press Cambridge.

Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012. You had me at hello: How phrasing affects memorability. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pages 892–901.

Marco Guerini, Danilo Giampiccolo, Giovanni Moretti, Rachele Sprugnoli, and Carlo Strapparava. 2013. The new release of CORPS: A corpus of political speeches annotated with audience reactions. In *Multimodal Communication in Political Speech. Shaping Minds and Social Action*, Springer, pages 86–98.

Marco Guerini, Gözde Özbal, and Carlo Strapparava. 2015. Echoes of persuasion: The effect of euphony in persuasive communication. *arXiv preprint arXiv:1508.05817* .

Fasih Haider, Fahim A Salim, Saturnino Luz, Carl Vogel, Owen Conlan, and Nick Campbell. 2017. Visual, laughter, applause and spoken expression features for predicting engagement within ted talks. *Feedback* 10:20.

John Heritage and David Greatbatch. 1986. Generating applause: A study of rhetoric and response at party political conferences. *American journal of sociology* 92(1):110–157.

Paul Hindemith. 1937. *The craft of musical composition*. Associated Music Publishers.

A Hume. 2017. Hook, line and sinker: How songwriters get into your head. *PORESO 2015: Redefining the boundaries of the Event* .

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *ACL (1)*. pages 13–24.

Yangfeng Ji and Noah A. Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 996–1005. http://aclweb.org/anthology/P17-1092.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*. pages 3294–3302.

Zhe Liu, Anbang Xu, Mengdi Zhang, Jalal Mahmud, and Vibha Sinha. 2017. Fostering user engagement: Rhetorical devices for applause generation learnt from ted talks. *arXiv preprint arXiv:1704.02362* .

Xing Lu. 1999. An ideological/cultural analysis of political slogans in Communist China. *Discourse Society, 10 (4), 487-508* .

Clifford K. Madsen and William E. Fredrickson. 1993. The experience of musical tension: A replication of Nielsen's research using the continuous response digital interface. *Journal of Music Therapy* 30(1):46–63. https://doi.org/10.1093/jmt/30.1.46.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*. pages 18–25.

C.R. Miller. 1939. *How to detect and analyze propaganda*. Town Hall, Inc.

Scott Novotney and Chris Callison-Burch. 2010. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 207–215.

Barbara Peskin, Larry Gillick, Yoshiko Ito, Stephen Lowe, Robert Roth, Francesco Scattone, James Baker, Janet Baker, John Bridle, Melvyn Hunt, et al. 1993. Topic and speaker identification via large vocabulary continuous speech recognition. In *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, pages 119–124.

Gregory A. Petrow and Terry Sullivan. 2007. Presidential persuasive advantage: Strategy, compliance-gaining and sequencing. *Congress and the Presidency* .

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, EPFL-CONF-192584.

Amruta Purandare and Diane Litman. 2006. Humor: Prosody analysis and automatic recognition for F*R*I*E*N*D*S. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 208–215.

Eric S. Rabkin. 1973. *Narrative suspense: When Slim turned sideways*. University of Michigan Press.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research* 15(1):1929–1958.

Eva Strangert. 2005. Prosody in public speech: analyses of a news announcement and a political interview. In *Ninth European Conference on Speech Communication and Technology*.

D. Talkin. 2015. Reaper: Robust epoch and pitch estimator. https://github.com/google/REAPER.

Sandra A Thompson and William C Mann. 1987. Rhetorical structure theory. *IPRA Papers in Pragmatics* 1(1):79–105.

Peter Vorderer, Hans Jurgen Wulff, and Mike Friedrichsen, editors. 1996. *Suspense: Conceptualizations, Theoretical Analyses, and Empirical Explorations*. Routledge.