

Are You TED Talk Material? Comparing Prosody in Professors and TED Speakers

TJ Tsai

EECS Department, University of California Berkeley, Berkeley, CA USA
International Computer Science Institute, Berkeley, CA USA

tjtsai@eecs.berkeley.edu

Abstract

TED talks are the pinnacle of public speaking. They combine compelling content with flawless delivery, and their popularity is attested by the millions of views they attract. In this work, we compare the prosodic voice characteristics of TED speakers and university professors. Our aim is to identify the characteristics that separate TED speakers from other public speakers. Based on a simple set of features derived from pitch and energy, we train a discriminative classifier to predict whether a 5 minute audio sample is from a TED talk or a university lecture. We are able to achieve $< 10\%$ equal error rate. We then investigate which features are most discriminative, and discuss conflating factors that might contribute to those features.

Index Terms: prosody, public speaking, lectures, TED

1. Introduction

This paper investigates a question that many academics have surely thought about while listening to a talk at a conference: “Why is this talk so incredibly boring?” And, more generally, “Why are so *many* academic talks so boring?” To this end, we will compare two groups of speakers – university professors and TED speakers – and try to identify and understand any consistent prosodic differences between the two groups. We will focus our study solely on *how* a person speaks (prosody), not on *what* they say (content). Given that TED talks are watched by millions of viewers, we can treat them as examples of good public speaking. With proper care of interpretation and analysis, consistent differences between these two groups may provide insight into what sets outstanding speakers apart.

Some works have tried to directly study characteristics of “good” speakers. Strangert [1] studies the prosody of an experienced news reporter and a politician in contrast to non-professional speakers in similar contexts. In a later work, Strangert and Gustafson [2] investigate the correlation between various pitch-related features and subjective ratings of how good a speaker is. They also manipulate the f_0 , fluency, and speech rate of an audio sample to study the perceptual effect of these variables, and they find that f_0 variability has the strongest effect. Since the concept of a “good” speaker is vaguely-defined, however, much more work has been done on understanding the related concept of charisma in speech. Holladay and Coombs [3] demonstrate the importance of vocal delivery in shaping perceptions of leader charisma by exposing respondents to a charismatic message in either a weak or strong delivery condition. Rosenberg and Hirschberg [4] [5] study how lexical and acoustic information affect subjects’ perceptions of charisma in politicians. Signorello et al. [6] likewise study the perception of charisma in an Italian politician and try to iden-

tify latent factors associated with charisma. Several works have studied the effect of culture on perceptions of charisma [7] [8] [9] [10] [11], indicating that many aspects of charisma are common across cultures while some aspects are culture-specific. Other works study how vocal characteristics influence perceptions of a speaker’s personality [12] or leadership effectiveness [13]. A related concept to charisma is the likability of a speaker. Prediction of likability was one of the tasks in the Interspeech Speaker Trait Challenge [14], which spurred additional interest on the topic.

Much of the work on charisma has revolved around political speech, which means that the studies are based on a very few number of speakers. This study shifts away from an emphasis on political speeches and instead considers general talks on a wide range of topics given by a large number of speakers. Additionally, this study does not try to explicitly predict or define any particular concept like charisma or likability (though clearly these concepts are very relevant to being an engaging speaker). Instead, it simply accepts that TED talks are examples of great public speaking (as attested by their millions of views) and simply asks the question, “What is it that great speakers do that other speakers don’t?” We use professors as a contrast to TED speakers.

To study the prosodic differences between TED speakers and university professors, we introduce a binary classification task that predicts whether or not a given audio sample is a TED talk or a university lecture. It’s important to point out that the classification task itself is not our ultimate goal – its purpose is to help us understand the differences between these two groups. The stronger the differences are, the better the classification will be. These differences may be a result of many different factors such as recording environment, speech format (50 minute lecture vs 10 minute talk), speaker personality, and communication ability. Nonetheless, our hope is to study a large number of speakers and to identify the most striking prosodic differences between university lectures and TED talks. Thus, the classification task is merely a means to an end. The end goal is to answer the question, “What makes TED speakers stand out?”

The rest of the paper is organized as follows. Section 2 explains the experimental setup. Section 3 summarizes the results of the classification task. Section 4 investigates which features are most important and how the audio sample length affects prediction accuracy. Section 5 interprets the meaning of the most discriminative features. Section 6 concludes the work.

2. Experimental Setup

We will explain the experimental portion of our setup in 3 parts: the data, the features, and the classifier.

2.1. Data

There are two main sources of data: university lectures and TED talks.

The first set of data came from `webcast.berkeley.edu`, a repository of webcast lectures available from the University of California Berkeley. We wanted a broad sample of as many professors as possible, so we downloaded the first available lecture from 338 different courses. The first lecture is often simply an overview of the course content, so it has the benefit of being representative of the way the instructor speaks in public. This is important for technical courses where lectures can have a lot of writing equations on the board. The 338 courses spanned 54 departments and were taught by 149 unique professors, 135 male and 14 female. Because there was relatively little data from female speakers, we focus only on the male speakers (313 lectures) in this study.

We prepared data samples from the lectures in the following way. After normalizing the volume and downsampling to 16kHz mono, we extracted 10 5-minute audio segments randomly selected from the duration of each lecture. Lectures typically last approximately 50 or 80 minutes, and the design choice of 10 5-minute samples is a tradeoff between having a long segment to aggregate statistics, having a reasonable number of total data samples, avoiding oversampling, and creating a balanced data set. From each 5-minute audio segment, we extract a fixed set of segment-level features. Each 5-minute audio segment thus represents a single data point in our training or testing set. So that our data samples represent the set of professors equally, we randomly chose 10 5-minute samples from each professor. With 135 professors and 10 samples per professor, there are a total of 1350 data points to be used for training and testing.

The second set of data consists of TED talks downloaded using the TED API. We first ordered the talks by the number of total views, and then selected the 391 most popular talks given by male speakers (this number was chosen to yield an approximately balanced data set, as will be seen below). These talks all had over a million views. As before, we normalize the volume and downsample to 16kHz mono. The talks typically range between 10 and 20 minutes, so we randomly select 4 5-minute audio segments from each talk. So that our data samples represent the 334 unique speakers equally, we randomly chose 4 5-minute samples from each speaker. With 334 speakers and 4 samples per speaker, there are a total of 1336 data points to be used for training and testing.

Because of the limited amount of data, we ran all experiments 10 times with random train-test splits on the speakers. For each of the 10 repetitions, all professors and TED speakers were thrown in a bag and 80% were randomly selected for training. All the data samples from the selected training speakers are used for training and the the data samples for the 20% of remaining speakers is used for testing. The reported results are averages across all 10 train-test splits.

2.2. Features

There are two principles that guided our selection of features. The first principle is interpretability. Since our goal is to *understand* the prosodic differences between the two groups of speakers, we restrict our attention to features that are interpretable. In this work, we only consider simple statistics of intuitive quantities. The second principle is simplicity. In this preliminary work, we aim for the lowest hanging fruit. There is no speech activity detection. There is no phrase segmentation. All of the

features we considered are simple statistics based on pitch and energy. We keep things as simple as possible.

We extracted 5 different families of features, each described below. All features are derived from frame-level estimates of pitch and energy, as estimated by the Snack toolkit [15]. The numbers of features are shown in parentheses after their description. For convenience, we define the distribution statistics of a collection to mean the average, standard deviation, quartiles (i.e. the 0, 25, 50, 75, and 100% quantiles), full range, and interquartile range.

f0 (100). The first family of features are various statistics derived from frame-level pitch estimates. This includes the distribution statistics of f0 estimates (18, deciles instead of quartiles), the fraction of frames that are voiced (1), and the inter-segment distribution statistics of intra-segment f0 distribution statistics (81). The latter refers to statistics computed on a segment level, where a group of contiguous voiced frames would constitute a segment. So, for example, this includes the average (across segment) f0 range within a segment and the standard deviation of the maximum f0 within each segment. This family of features describes the speaker’s use of pitch on both a frame and segment level.

Segment length (19). The second family of features are statistics of segment lengths, where segments are blocks of contiguous frames that are all voiced or unvoiced. This includes the distribution statistics on voiced segments (9), the distribution statistics on unvoiced segments (9), and the total number of voiced segments (1). This family of features describes the speaker’s continuity of speech and use of silence.

Voiced energy (198). The third family of features are statistics derived from the frame-level energy of voiced frames. This includes (a) the distribution statistics of energy for windows of voiced frames (36) and (b) the inter-segment distribution statistics of intra-segment energy distribution statistics (162). For both (a) and (b) we consider the statistics of the rms energy over various window lengths. For example, (a) includes the mean rms energy in windows of 10 consecutive voiced frames and the range of rms energy in windows of 50 consecutive voiced frames. (b) includes statistics across segments, such as the average (across segments) of the maximum energy reached by any 50 frame window in a segment of voiced frames. We considered windows of 1, 10, 20, and 50 frames for (a), and windows of 1 and 10 frames for (b). This family of features describes the speaker’s use of volume on both a frame and segment level.

Unvoiced energy (216). This family is identical to the voiced energy features, but applied to unvoiced frames. We consider windows of 1, 10, 20, 50, 100, and 200 frames for (a), and windows of 1 and 10 frames for (b). Unvoiced frames can contain both unvoiced speech and silence, but these statistics can still provide some indication of what is happening when the speaker is not speaking.

Global energy (63). The fifth family of features are statistics derived from the frame-level energy of all frames, irrespective of being voiced or unvoiced. This includes the distribution statistics of energy for windows of 1, 10, 20, 50, 100, 200, and 500 frames (63). This family of features describes the overall sound characteristics throughout the audio segment.

In total, we extracted 596 features from each audio segment.

2.3. Classifier

We used adaboost with tree stumps as our classifier [16]. For each train-test split, we ran cross-validation experiments on the

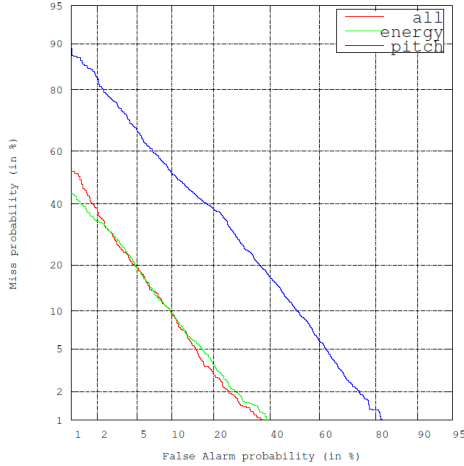


Figure 1: DET curves for the adaboost model with three different features sets: pitch, energy, and all features combined.

training portion to select the number of trees. We will investigate the performance of our adaboost models in the next section.

3. Results

In this section we describe the performance of our adaboost models. Figure 1 shows the DET curves for the adaboost model with 3 different sets of features: pitch-related features (f_0 and segment length), energy-related features (voiced, unvoiced, and global energy), and all features combined. There are two things we can notice about figure 1. First, energy features perform much better than pitch features. In fact, there is little or no benefit in adding pitch features to the energy features – we can see that using all features is no better than using energy features only. Second, all of the models perform much better than random. Note that random guessing corresponds to 50% equal error rate (EER). So, even though the pitch features perform much worse than the energy features, they are still providing useful information in discriminating between the two classes.

4. Analysis

In this section we investigate two questions of interest.

4.1. Feature Importance

The first question we would like to answer is, “Which features are the most discriminative?” One way we can approach this question is to compare the performance of models trained on different subsets of features. Table 1 shows the EERs for the adaboost model trained on various subsets of features, where results are averaged over 10 random train-test splits. Again, we see that energy features perform much better than pitch features, but that all feature subsets show performance much better than random. Interestingly, the unvoiced energy features showed the best performance.

Another way we can approach this question is to look at the relative influence of individual features. Relative influence is the reduction in the loss function attributable to a single feature, normalized by the total reduction in loss due to all features [17]. This measure indicates how much an individual feature influences the adaboost prediction.

| Feature | EER | std |
|------------------------|-------|------|
| Pitch - f_0 | 30.5% | 3.5% |
| Pitch - Segment Length | 35.5% | 4.4% |
| Pitch - All | 27.6% | 3.7% |
| Energy - Voiced | 17.9% | 3.2% |
| Energy - Unvoiced | 9.8% | 2.2% |
| Energy - Global | 17.3% | 3.5% |
| Energy - All | 9.5% | 1.7% |
| All Features | 9.5% | 2.5% |

Table 1: EER of adaboost model with various feature subsets, averaged over 10 random train-test splits.

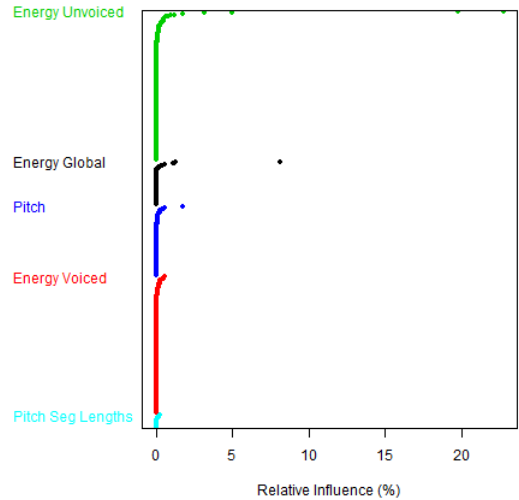


Figure 2: Relative influence of all features in the adaboost model, grouped by feature type and sorted in decreasing order of influence.

Figure 2 shows the relative influence of all 596 features in the full adaboost model, averaged over 10 random train-test splits. The features have been grouped first by family, and then sorted in decreasing order of relative influence within each grouping. There are two things we can notice from this figure. First, a few features dominate. We see that most features have relative influence of approximately 0, and a few features make up the bulk of relative influence. For example, the top 10, 84, and 146 features constitute 66%, 90%, and 95% of the total relative influence, respectively. This suggests that we could substantially reduce the number of features without much degradation in classification performance. We will investigate the top several features more closely in the next section. Second, a few families dominate. The five families of features shown in figure 2 from top to bottom make up approximately 67%, 14%, 9%, 8%, and 2% of the total relative influence, respectively. In particular, the family of unvoiced energy features heavily affects the adaboost predictions.

The answer to our first question is this: the most discriminative features are energy statistics extracted from unvoiced regions.

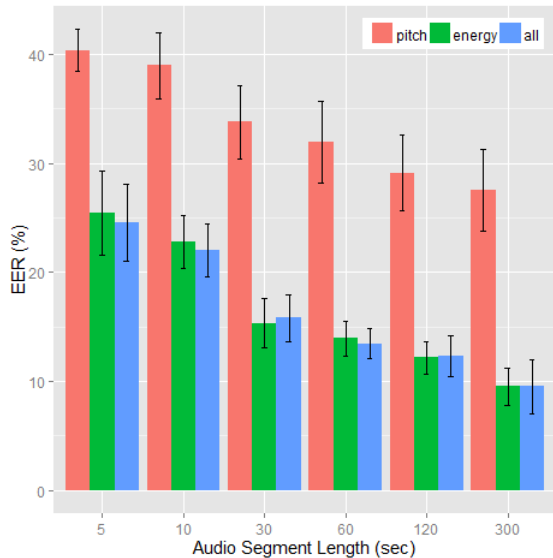


Figure 3: Effect of audio segment length on the EER of adaboost models.

4.2. Segment Length

The second question we would like to answer is, “How does audio segment length affect our results?” For all of the reported results so far, we have used random 5 minute segments from each speaker. We might wonder how well we can discriminate based on shorter samples.

To answer this question, we repeated the original experiments with a range of segment lengths. Figure 3 shows the EER for adaboost models trained on audio segments of length 5, 10, 30, 60, 120, and 300 seconds. Each group of bars shows the EER for 3 different feature combinations (pitch, energy, all) using a fixed audio segment length. Each bar represents the mean EER across 10 random train-test splits, and the error bars show one standard deviation above and below the mean.

There are 3 things to notice about the barplot in figure 3. First, there are consistent and significant improvements as the segment length increases from 5 seconds up to 5 minutes. We do note, however, that the improvements are tapering off (note that the indicated segment lengths do not increase linearly). Second, these experiments all confirm our earlier findings: energy features perform much better than pitch features, and show little or no benefit from combination. This pattern is consistent across all segment lengths. Third, we can do surprisingly well for very short segment lengths. Note that for 5 second segment lengths, we can still achieve about 25% EER. One natural question that follows is whether it would be better to accumulate statistics on one longer audio segment, or to combine the predictions of many smaller adjacent audio segments. This is a question to tackle in future work.

The answer to our second question is this: we can achieve 25% EER with 5 second segments and 10% EER with 5 minute segments.

5. Discussion

In this section, we examine some of the most influential features in our adaboost model, and consider their interpretation.

Features 1, 2, and 4 (when ordered by influence) describe

the spread of the energy distribution in unvoiced regions, where a higher spread is associated with TED talks. This suggests that relatively more “stuff” is happening during unvoiced regions – less of the unvoiced regions are spent in silence (which has low variance) and more time is occupied with some type of acoustic event such as unvoiced speech or audience laughter (which has higher variance). The TED talks tend to be more “dense” in that less time is spent in silence.

Feature 3 describes the 5 sec window within the audio segment which has the least energy, where more silence is associated with lectures. This suggests that TED talks are less likely to have a 5 second window where nothing happens, whereas this may be more common in lectures. Again, this indicates that TED talks spend less time in silence and filler, and more time in high-energy speech.

Feature 6 describes the single .2 second window within any unvoiced region which has the least energy, where lower energy is associated with TED talks. Given how short the window length is, this feature is probably capturing either the acoustic environment or the quality of the audio recording equipment – the silences in TED talks are more silent, either because the environment is more quiet or the recording equipment has a lower noise floor. Feature 9 has a similar interpretation.

Feature 7 is the 10% quantile of f0 and can be thought of as a conservative estimate of the speaker’s lower pitch range, where deeper voices are associated with TED talks. It could be that viewers prefer listening to deeper voices, that speakers with deeper voices are more likely to be giving TED talks, or that the TED speakers have had more training in how to speak in chest (rather than head) voice.

Feature 8 describes the distribution spread of energy in 5 second windows, where less variability is associated with TED talks. This suggests that TED speakers have a more consistent delivery over longer windows of time, whereas the lectures tend to have more varied chunks of silence and speech.

6. Conclusion

This paper examines prosodic differences between university professors giving classroom lectures and speakers giving TED talks. We train a classifier to predict whether a given audio segment is from a TED talk or a classroom lecture. The classifier achieves < 10% EER with 5-minute audio segments, and about 25% EER for 5-sec segments. By studying the relative influence of features in the classifier, we can discern the most striking prosodic differences: TED speakers give talks that are more dense (i.e. less space and silence), speak with a deeper voice, and have a more consistent flow of delivery (energy over 5 second windows). These differences may simply be a result of the difference between a long lecture and a short talk. But it is worthwhile to point out that fulfilling the above characteristics is no easy task. A speaker who spends all of his time in high-energy speech while maintaining a consistent delivery is almost certainly a speaker who is very well prepared – he has something to say and knows how to say it. Perhaps that is one key part of what makes TED speakers stand out.

7. Acknowledgements

Thanks to Elizabeth Shriberg, Dan Jurafsky, Julia Hirschberg, Dilek Hakkani-Tur, and Nikki Mirghafori for useful feedback and discussions.

8. References

- [1] E. Strangert, "Prosody in public speech: Analyses of a news announcement and a political interview," in *Interspeech*, 2005.
- [2] E. Strangert and J. Gustafson, "What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations," in *Interspeech*, 2008.
- [3] S. J. Holladay and W. T. Coombs, "Communicating visions: An exploration of the role of delivery in the creation of leader charisma," *Management Communication Quarterly*, vol. 6, no. 4, pp. 405–427, 1993.
- [4] A. Rosenberg and J. Hirschberg, "Charisma perception from text and speech," *Speech Communication*, vol. 51, no. 7, pp. 640–655, 2009.
- [5] J. Hirschberg and A. Rosenberg, "Acoustic/prosodic and lexical correlates of charismatic speech," in *Interspeech*, 2005.
- [6] R. Signorello, F. D'errico, I. Poggi, and D. Demolin, "How charisma is perceived from speech: A multidimensional approach," in *International Conference on Privacy, Security, Risk and Trust (PASSAT)*, 2012, pp. 435–440.
- [7] F. Biadys, A. Rosenberg, R. Carlson, J. Hirschberg, and E. Strangert, "A cross-cultural comparison of american, palestinian, and swedish perception of charismatic speech," in *Proc. of Speech Prosody*, 2008.
- [8] J. Hirschberg, F. Biadys, A. Rosenberg, and W. Dakka, "Comparing american and palestinian perceptions of charisma using acoustic-prosodic and lexical analysis," in *Interspeech*, 2007.
- [9] A. Cullen, A. Hines, and N. Harte, "Building a database of political speech: Does culture matter in charisma annotations?" in *Proc. of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 27–31.
- [10] D. N. D. Hartog, R. J. House, P. J. Hanges, S. A. Ruiz-Quintanilla, and P. W. Dorfman, "Culture specific and cross-culturally generalizable implicit leadership theories: Are attributes of charismatic/transformational leadership universally endorsed?" *The Leadership Quarterly*, vol. 10, no. 2, pp. 219–256, 1999.
- [11] F. D'Errico, R. Signorello, D. Demolin, and I. Poggi, "The perception of charisma from voice: A cross-cultural study," in *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 552–557.
- [12] M. Zuckerman and R. E. Driver, "What sounds beautiful is good: The vocal attractiveness stereotype," *Journal of Nonverbal Behavior*, vol. 13, no. 2, pp. 67–82, 1989.
- [13] T. DeGroot, F. Aime, S. G. Johnson, and D. Klumper, "Does talking the talk help walking the walk? an examination of the effect of vocal attractiveness in leader effectiveness," *The Leadership Quarterly*, vol. 22, no. 4, pp. 680–689, 2011.
- [14] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. V. Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The interspeech 2012 speaker trait challenge," in *Interspeech*, 2012.
- [15] K. Sjölander *et al.*, "The snack sound toolkit," *KTH Stockholm, Sweden*, 2004.
- [16] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
- [17] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, pp. 1189–1232, 2001.