Inferring Stance in News Broadcasts from Prosodic-Feature Configurations

Nigel G. Ward, Jason C. Carlson, Olac Fuentes

Computer Science, University of Texas at El Paso
500 West University Avenue, El Paso, Texas 79968 USA

Abstract

Speech conveys many things beyond content, including aspects of appraisal, feeling, and attitude that have not been much studied. In this work we identify 14 aspects of stance that occur frequently in radio news stories and that could be useful for information retrieval, including indications of subjectivity, immediacy, local relevance, and newness. We observe that newsreaders often mark their stance with prosody. To model this, we treat each news story as a collection of overlapping 6-second patches, each of which may convey one or more aspects of stance by its prosody. The stance of a story is then estimated from the information in its patches. Experiments with English, Mandarin, and Turkish show that this technique enables automatic identification of many aspects of stance in news broadcasts.

Key words: information retrieval, attitude, broadcast news, prosody, American English, Mandarin, Turkish

1. Introduction

People use language not only to convey factual information but additional information such as attitudes, opinions, feelings, judgments, categorizations, and so on. Aspects of this have been studied under many names, including sentiment, attitude, feelings, appraisal, and stance (Read and Carroll, 2012; Rambow and Wiebe, 2015; Chindamo et al., 2012; Biber and Staples, 2014; Elfardy and Diab, 2016; Mohammad et al., 2016). This article will use "stance" as an umbrella term for such information, including all the nuances and subtleties

[☆]The first author started this work at Kyoto University. We thank Elizabeth Shriberg, Diego Castan, and Andreas Tsiartas for discussion, for tidying the English-data filenames, and for demonstrating the need to consider more features, and Gerardo Cervantes for discussion. This work was supported in part by DARPA under the Lorelei program, by a Fulbright Award, and by the NSF through REU supplements to IIS-1449093. This work does not necessarily reflect the position of the US Government, and no official endorsement should be inferred.

Email addresses: nigelward@acm.org (Nigel G. Ward), jccarlson@miners.utep.edu (Jason C. Carlson), ofuentes@utep.edu (Olac Fuentes)

URL: www.nigelward.com (Nigel G. Ward)

of attitudes and related functions that people display in the course of pursuing various communicative goals.

Stance is richly present in many genres of language use, and especially in spoken language. Such information can be useful for information retrieval and filtering, among other purposes. Previous work has explored how, in speech data, stance can be inferred from the speaker's prosody, but so far only for a handful of aspects of stance. Building on this work, this paper examines 14 aspects of audio documents as they occur in English, Mandarin, and Turkish radio news, showing how they can be automatically detected from prosody. The contributions[1] include:

a. a description of 14 aspects of stance, extending the inventory of computationally-modeled aspects of sentiment and attitude (Section 3)
b. the finding that stance is commonly present in news stories (Section 4)
c. a model of how prosody conveys stance (Sections 6 and 9)
d. a demonstration that this model is able to infer stance, in three unrelated languages (Sections 6 and 11)
e. evidence that the prosodic expressions of stance include configurations of diverse prosodic features, and illustrations of such configurations (Sections 7 and 10)
f. the finding that the significance of a prosodic-feature configuration can depend on where it occurs in a story (Section 11)
g. the finding that the prosodic expressions of stance are extremely language-dependent (Section 12)

2. Motivation and Related Research

Stance can be useful for information retrieval, information filtering, and information extraction (Larson and Jones, 2012; Lee et al., 2015; Purver et al., 2007; Freedman et al., 2011; Wollmer et al., 2013). Although there has been a fair amount of research on stance, so far it has been mostly limited to expressions of sentiment, positive or negative, with also some work on presence of or strength of opinion (Liu and Zhang, 2012; Freese and Maynard, 1998; Pillet-Shore, 2012; Wilson et al., 2005; Freeman et al., 2015). In this paper we examine many more aspects of stance.

Although most computational models of stance so far has been for text, there is also some exemplary work on stance in spoken language. For example, Morency *et al.* (2011) found, in product-review videos, that valenced utterances, in comparison to neutral utterances, have wider pitch range and fewer pauses. Using this information with lexical and video information in a HMM-based model enabled good classification of positive/negative/neutral opinions. Mairesse *et al.* (2012) similarly found and exploited prosody-sentiment mappings, such as a correlation between low pitch variability and negativity. Levow and colleagues studied stance in dialog and found, among other things, that *yeah* when expressing stronger

[1] Compared to our earlier paper on this topic (Ward et al., 2017), the current paper adds full detail on the methods and experiments, additional observations and findings, and new contributions: e, f, and g.

stances tends to have higher intensity and pitch, and when expressing positive stance tends to be longer, quieter, and higher-pitched (Freeman et al., 2015; Levow and Wright, 2017). In this paper we study more specifically how prosody expresses stance.

We were inspired to study stance in spoken language by a practical problem: making humanitarian assistance and disaster relief more effective. After a disaster, whether natural or anthropogenic, the international community often mobilizes to help. However, this involves many challenges, including, as described in the Lorelei scenario (DARPA, 2014), enabling a person planning a relief mission to obtain a clear picture of what is needed and what should be done. Today mission planners rely heavily on news broadcasts and social media communications to obtain relevant information. However, the large volume of such data makes this difficult, and there is the need for better tools to help filter and organize such inputs. A further challenge is that disasters can happen anywhere, and the information sources can be in any language, including low-resource languages, for which there may be neither tools (speech recognizers, machine translation systems, etc.) nor adequate resources for building tools. In such cases even imperfect filtering can be of value. Moreover, given the typically large volumes of raw information, mission planners may use information obtained from statistics and tendencies across many items, and for this they may use tools where the categorization of any specific item is only probabilistically accurate.

Stance is relevant to this domain because mission planners frequently need not only actionable information but also big-picture information relating to situational awareness (Verma et al., 2011) — such as indications of where the flooding situation is worse, what needs are most immediate, which relief organizations are being evaluated positively by the local population, and so on — which often relate to stance. Thus stance is potentially useful, both by itself and as a complement to more traditional methods such as classification by topic. Moreover it may be possible to develop stance detectors rapidly for a suddenly-relevant language, without requiring substantial resources. This is because stance is often conveyed by prosody, and prosody is in some respects simpler to process than lexical information, especially for low-resource languages, where we may lack even fundamental knowledge of the phonology and syntax. In addition, while vocabularies differ arbitrarily across languages, there are universal tendencies for some prosodic features to express certain things across language (Gussenhoven, 2002; Vaissiere, 2008), so a prosody-based approach could be useful for previously-unstudied languages.

To make stance information available for such purposes is one aim of this work; the other is to use this domain to investigate how prosody conveys stance.

3. Fourteen Aspects of Stance

Table 1 shows the 14 aspects of stance considered in this work. Developing this list was a long process, described more fully elsewhere (Ward, 2016). In summary, we considered both what information is available and what is useful, and sought the interesection.

In terms of what is available, we drew from two previous lines of basic work. The first was an investigation of what people find interestingly similar in dialog data (Ward et al., 2015) and what they might be likely to search for, besides of course topic. The

second was an ongoing project to inventory the principal functions expressed by prosody in various languages. We found that, across languages, many of the same functions were being expressed, and many of these related to some kind of stance. We also consulted the literature on stance in various genres of speech.

In terms of what is useful, we started with the the Lorelei operational scenario, and then, regarding various disaster-relief activities learned what aspects of information could be relevant to analysts and mission planners.

These considerations led us to formulate an initial list of 28 stance aspects, organized as 14 binary oppositions. We then winnowed this list considering three things: first, likely utility for the scenario, as judged by conversations with Lorelei stakeholders, second, non-redundancy, both among stance aspects and with respect to what topic-based filtering may provide, and, third, inter-annotator agreement, and thus the ability for predictions to be reliably evaluated. Thus we aimed to find a set that is mostly non-redundant, fairly reliably annotatable, and likely useful.

Over the course of three pilot annotations we refined the descriptions and culled the list, to arrive at the 14 stances in Table 1. In this list there are no explicit pairings of stances; rather each stance is taken to have its own independent existence. Thus an annotator could label a given segment, for example, as both deplorable and praiseworthy, if it mentioned both a deplorable act and a praiseworthy one.

In no sense is this a definitive list of stance items; it is a compromise between these various considerations. Despite the limitations, this list provides a much broader view of stance than any previous work.

4. Stance in Radio News

To investigate the manifestations of these stance aspects, considering relevance to the scenario, we chose to work with radio news broadcasts, local news when possible. The section describes the data and the annotation, and what we learned about the pervasiveness of stance.

To support study of possible universals, we chose to use three unrelated languages: English, Mandarin, and Turkish.

The American English data set is 488 minutes selected from radio broadcasts at archive.org, consisting mostly of local news from three radio stations, WMMB, KBND, and CHEV, but including others chosen to increase the coverage of disaster-related topics, including shootings, protests, earthquakes, floods, power outages, hurricanes, various storms, epidemics, and wildfires. We also experimented with a single-speaker subset, consisting of 125 minutes from WMMB.

The Mandarin data set is the first 279 minutes of the KAZN subset of the Hub4 collection (Huang et al., 1998), comprised of local, national, and international news and a variety of other things. Compared to the English data, the KAZN data had more prosodic variety, including more acoustic variation between segments and more segments that were not simply read news, including spontaneous speech and interactions among announcers.

1. **Bad Implications** - information with undesirable consequences, such as a raise in taxes, an approaching storm, or a flood. Events with more severe implications will rate higher on this scale
2. **Good Implications** - the opposite, such as a peace agreement, a good harvest, or nice weather
3. **Deplorable Action** - something bad done by someone or some organization
4. **Praiseworthy Action** - the opposite: something good done by someone or something
5. **Controversial** something people can or could disagree about, such as a bold action by some person or group, or new government policy
6. **Factual Information** information presented as facts
7. **Subjective Information** - the opposite, such as opinions, either the presenter's or someone else's, or information reported skeptically or speculatively
8. **Unusual or Surprising** - something quirky, odd, or unexpected
9. **Typical or Unsurprising** - something expected
10. **Local** - personally relevant to the listening audience, like local weather or close-by rioting
11. **Something Prompting Immediate Action** - something that may motivate the listening audience to do something, like take shelter from a storm or vote in today's election
12. **Background** - conversely, information useful just as background, such as an explanation of the causes of a situation
13. **New Information** - new information or description of a recent development (rather than a repetition or rehash of something previously reported)
14. **Relevant to a Large Group** - something affecting many people

Table 1: Descriptions of each stance aspect, as given to the annotators

For Turkish we were unable to obtain local news data. Instead we used the first 672 minutes of a Lorelei-program data set, LDC catalog number 2014E115, consisting of 5 broadcasts from China Radio International and 22 from Voice of America.

Each news broadcast was divided into news stories or segments, with topics like: weather, hockey, parenting, bicycle race, jazz festival, hospital donation, erosion, evacuation, highway closing, drug arrest, job fair, burglary, and so on. Segments vary in length from tens of seconds to a few minutes. The Turkish data was further subdivided mechanically so that no segment exceeds 2 minutes. In all, the English data had 981 segments, including 267 segments in the single-speaker subset, the Mandarin 307, and the Turkish 1038.

We then obtained stance annotations for each dataset. We chose to do this at the segment level, rather than at the utterance or entity level, for two reasons: making the annotation task more tractable, and alignment with the Lorelei program, which in the first phase focused on segment-level classifications. While in reality stances are almost certainly

continuous-valued phenomena — for example, a news story can be more or less local, or more or less surprising — for the sake of having a feasible annotation task, we chose to use discrete labels. Thus annotators labeled each segment for each stance aspect as 0 (absent), 1 (weakly present) or 2 (strongly present). Annotation was outsourced. For each language it was done by three native speakers working independently, using the definitions in Table 1 and a small set of sample annotations. More information on the broadcasts and segments, and the annotations themselves, are available at http://www.cs.utep.edu/nigel/stance/ .

We examined the annotations in three ways. First, we computed inter-annotator agreement. For this we used the average pairwise weighted Kappa, giving partial credit, 0.5, for close matches, for example, a rating of 2 by one annotator and a 1 by another. As seen in Table 2, interannotator agreement was excellent for some stances and poor for others, depending also on the language. Second, we computed the correlations among stances, more specifically, among the three-annotator averages for each stance, across all segments. Overall the correlations were low (the most related pair in English, bad and good, correlated at only –0.59), indicating that these 14 stances are largely mutually non-redundant. Third, we examined the frequency of expressions of stances. Table 2 shows, in the rightmost columns, the percent of news stories annotated with a 1 or 2. Overall the stances are fairly common, with some understandable exceptions, such as the lack of local and immediate stances in the Turkish data. While it is sometimes thought that news should be purely objective, in fact, newsreaders strive to contextualize and humanize the news (Cotter, 1993), and the variety of stances frequently expressed confirms this.

	Agreement			Presence		
	Eng.	Man.	Tur.	Eng.	Man.	Tur.
1 Bad	0.72	0.55	0.36	33%	15%	18%
2 Good	0.46	0.45	0.35	34%	19%	5%
3 Deplorable	0.74	0.37	0.60	14%	3%	17%
4 Praiseworthy	0.35	0.46	0.36	14%	8%	6%
5 Controversial	0.53	0.56	0.40	4%	5%	25%
6 Factual Information	0.25	0.59	0.41	96%	94%	46%
7 Subjective Information	0.36	0.55	0.45	11%	46%	39%
8 Unusual or Surprising	0.22	0.69	0.30	6%	7%	18%
9 Typical or Unsurprising	0.81	0.59	0.14	31%	9%	11%
10 Local	0.39	0.66	0.05	80%	46%	8%
11 Prompting Immediate Action	0.69	0.74	0.06	20%	32%	5%
12 Background	0.47	0.60	0.21	36%	23%	29%
13 New Information	0.30	0.92	0.19	41%	48%	8%
14 Relevant to a Large Group	0.47	0.83	0.28	48%	21%	17%

Table 2: Statistics on stance in the three collections: inter-annotator agreement for, and presence of, each stance, in English, Mandarin, and Turkish.

5. Task and Metrics

Below we present models able to automatically predict the stances present in any new news story or segment; this section describes how we evaluate the quality of such predictions.

Because we assume that stances are continuous-valued phenomena, we use the average of the three annotators' labels as the ground truth, target value. This is computed independently for each stance, for each segment. A model is better to the extent that its predictions are closer to these true values, as measured by squared error.

Thus we can test whether one predictor outperforms another by generating predictions and evaluating them, where each news segment is an independent sample. We judge statistical significance using one-tailed matched-pairs t-tests, one test for each stance, with a significance level of $p < .05$.

We measure the overall quality of a model using mean squared eror (MSE). This enables comparison to human performance. As the annotators do not agree perfectly, we can estimate how well a human would perform at this task. Specifically, we estimate this as the mean squared error of his or her judgments, that is, the square of how much one annotator's labels diverge from the average of all three annotators. (This is a slightly optimistic estimate.)

Further, we wanted a summary performance metric, a single number to represent the overall performance. For this we used the global MSE, computed across all stances; this is what we sought to optimize.

Finally, as the MSE depends not only on the method but also on the data set, we also wanted a data-independent metric. For this we use the percent reduction in error, that is, the difference between baseline MSE and model MSE divided by the baseline MSE. This measures how much of the actually-possible prediction power was achieved. In the range from 0 to 100%, this is meaningful for comparisons across datasets. As the baseline we use the performance of a knowledge-free method: predict-the-average. This average of course differs for each stance and for each reference set.

6. Initial Experiments

The first fundamental question is whether prosody bears information useful for inferring these stances. Expecting this to be the case, we hypothesized that using prosodic information will enable prediction of the stance aspects present in news stories better than a baseline predictor. This section describes our test of this hypothesis.

6.1. Approach

We based our models on one observation and two working assumptions.

The observation is that news segments are heterogeneous in terms of what is said and how. A stance, when present, is not necessarily expressed, or even relevant, continuously throughout a news story; rather, it may be indicated mostly in a few specific regions. For example, in a news story containing the sentence *Two SQ constables are being credited with saving three people from a burning house in Rowdon*, the prosodic indications that this was

"praiseworthy" are present more on the subject and predicate than on the village name, let alone on the subsequent descriptions of the fire's origin. Thus this problem is different from classification tasks where something is assumed to be broadly present across an input, either because it is a direct indication of a mental or physical state or trait, or because each input is short: a single utterance or a single word (Schuller, 2011; Mairesse et al., 2012; Freeman et al., 2015).

Thus we need a way to use locally-present stance information to infer story-level stance, as annotated. Ideally we would use a model of the rhetorical and discourse structures of news to locate the most informative regions for any specific type of stance, but no current model is suitable (Cardoso et al., 2013; Liu et al., 2015). Accordingly, we use a simple estimate-locally-then-aggregate method (Shen et al., 2014; Schmitt et al., 2016).

Thus, for every small region we estimate the strength of each stance there. Following computer-vision terminology, we will refer to these regions as "patches." For simplicity, we do this for each patch independently.

The first working assumption is that stance is expressed by configurations of prosodic features within each patch. This implies that patches similar in prosody will be similar in stance.

The second working assumption is that many of these stance-expressing feature configurations are limited in scope, so that patches of about 6 seconds will work well.

6.2. Initial Model

Given these assumptions, it was natural to base our initial model on a k-nearest-neighbors algorithm. Although more sophisticated algorithms would likely perform better, for the initial investigation we preferred simplicity. One advantage of k nearest neighbors is in making minimal assumptions about the distributions. Another advantage is that it is interpretable, enabling us to examine successes and diagnose failures to learn abut the nature of the problem, as will be seen below.

We implemented nearest neighbors straightforwardly. For each patch in the segment to classify, we find the k most similar patches in the reference data set. For each of these k neighbors, we then look up, in the annotations, how that stance was annotated in the segment it came from. For example, in classifying a sports segment, the nearest neighbor of a patch in the middle of *snap their losing streak with a win against* was a patch in the middle of *partly sunny and a warm day*, which was in a segment labeled "local=2, good=2, new=2." This was thus evidence that the sports segment is also conveying something that is locally-relevant, good news, and new information. A reference patch is more relevant to the extent that it is more similar to the new patch, so each neighbor contributes with a weight proportional to the estimated similarity. Weights are normalized such that the estimates are not affected by the local density or sparsity of neighbors.

More formally, let the collection of vectors $\vec{q}_1, \vec{q}_2, \ldots, \vec{q}_n$ represent the feature vectors for each patch i, and the collection of vectors V represent the feature vectors of all the patches from all the segments in the training data. We choose $\vec{v}_1, \vec{v}_2, \vec{v}_3 \in V$ for each $\vec{q}_i$ such that the Euclidean distances $d_j = \|\vec{q}_i - \vec{v}_j\|$ for d_1, d_2, d_3 are minimal, or, more formally:

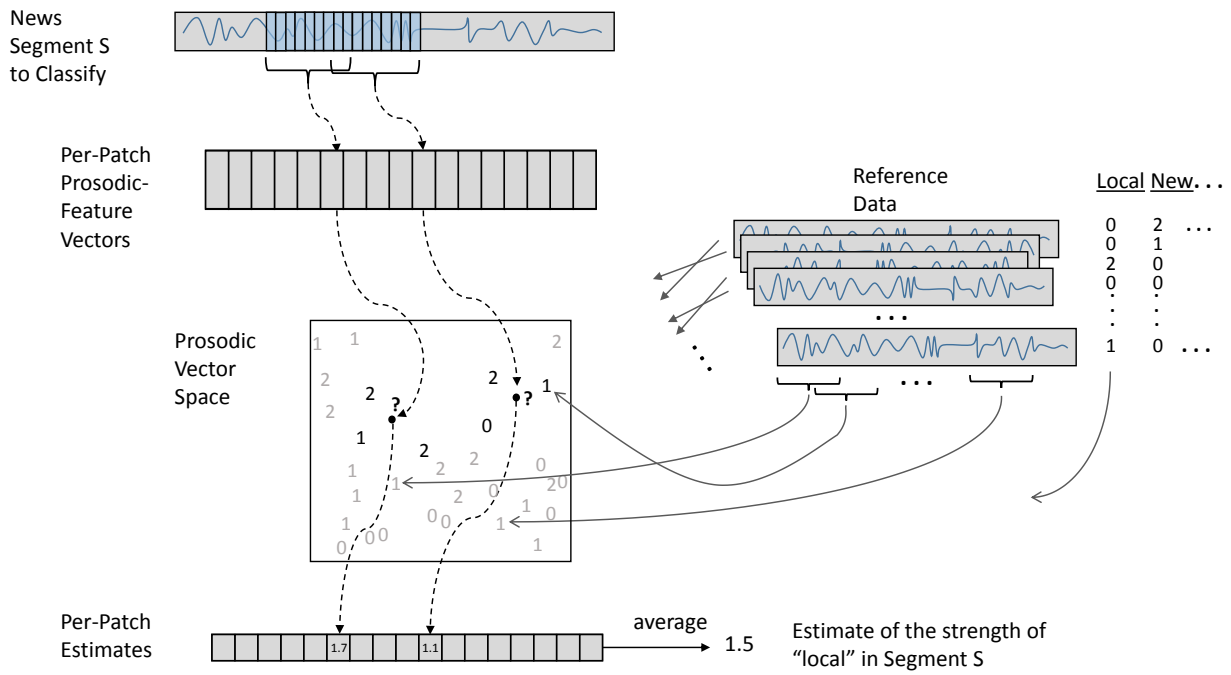

Figure 1: Overview of the initial model. Given a news segment S to classify, we take overlapping samples (patches), and represent each as an n-dimensional prosodic-feature vector. For each patch we find the nearest neighbors, each representing a patch from the reference data, and each inheriting the stance annotations of the news segment which contained it. The stance of each patch is estimated using the values of its neighbors, and the overall stance of the news segment is estimated as the average of the estimates for each patch. From (Ward et al., 2017).

$$d_1, d_2, d_3 \quad \leq \quad \|\vec{q}_i - \vec{v}\| \quad \forall\ \vec{v} \in V \setminus \{\vec{v}_1, \vec{v}_2, \vec{v}_3\}$$

Let s_1, s_2, s_3 be the three annotator's average rating for a particular stance for the reference data segments containing $\vec{v}_1, \vec{v}_2, \vec{v}_3$, respectively. We then find the prediction p_i for that particular stance for each patch i by taking a weighted average of s_1, s_2, s_3, using the inverse squared distance from $\vec{q}_i$ to $\vec{v}_1, \vec{v}_2, \vec{v}_3$ as the weights:

$$p_i = \frac{w_1 s_1 + w_2 s_2 + w_3 s_3}{w_1 + w_2 + w_3}$$

where

$$w_1 = \frac{1}{{d_1}^2}, \ w_2 = \frac{1}{{d_2}^2}, \ w_3 = \frac{1}{{d_3}^2}$$

Finally, assuming for now that all patches are equally informative, we estimate the segment-level stance P as the average of the individual patch predictions:

$$P = \frac{1}{n} \sum_{i=1}^{n} p_i$$

Figure 1 depicts the method. We use $k = 3$ nearest neighbors, based on preliminary experiments. Currently patches are offset every 100 ms, both in the story to classify and in the reference data, thus, depending on the length of the segment, there may be tens or thousands. The 100 ms offset was chosen because it is unlikely that stance often varies faster than this.

6.3. Initial Feature Sets

To find the nearest neighbors we need a way to judge prosodic similarity. Although aspects of the perception of prosody and of prosodic similarity have been studied (Reichel et al., 2009; Rilliard et al., 2011), no existing models are suitable for our needs. Accordingly we adopt a simple model: we compute distances in a vector space, where each dimension is given by the values of one prosodic feature, and estimate similarity between two neighbors as the inverse of their squared distance.

To implement this we need of course a set of prosodic features. Not knowing at first which prosodic features are important for stance, we started with two feature sets developed for other tasks. Both broadly characterize the prosody across a region, using a large number of diverse features.

Our first set, utep-m1, was from a suite originally developed for language modeling and later extended for other purposes (Ward et al., 2011; Ward and Vega, 2012). The specific set used was 88 features, taken unmodified from a recent study of prosody of non-native speakers in dialog (Ward and Gallardo, 2017). This set includes measures of intensity, of pitch height (high or low), of pitch range (narrow or wide), of speaking rate (using energy flux as a proxy), and of creakiness. This is a time-spread set: each feature is computed over various windows, together spanning a patch about 6 seconds long. There is a "foveal region" in the middle of the patch, where the windows are shorter; that is, the features are more fine-grained in that region, as illustrated in Figure 3. The complete list is seen in Figure 2.

This set does not include features that are turn-, utterance-, word-, or syllable-aligned, so that all features can be everywhere-computable and robust. Most features are normalized to reduce dependence on speaker and recording conditions. In addition, each feature is z-normalized, across all audio tracks in a dataset, so that each feature contributes equally to the distance computations.

The implementation details are at (Ward, 2017), from which the code can also be downloaded.

The second set of features was the "extended Geneva Minimalistic Acoustic Parameter Set" (eGeMAPS) (Eyben et al., 2016), designed for emotion recognition. This is an subset selected from among the thousands in Opensmile (Eyben et al., 2010); coincidentally this subset also has 88 features. In addition to prosodic features, generally covering the same phenomena as those in utep-m1, the Geneva set also includes spectral features (mel-frequency cepstral coefficients) and temporal features relating to pause duration and syllable duration.

The Geneva features were extracted using the OpenSmile command-line tool. Since we need features for a window (patch) spanning approximately 6 seconds, we modified the eGeMAPS configuration to generate a feature vector for equivalent-sized windows. Again all features were z-normalized.

intensity (16)	low pitch, high pitch, creakiness (14 each)	narrow pitch, wide pitch (10 each)	speaking rate (10)
-3200 – -1600			
-1600 – -800	-1600 – -800	-1600 – -800	-1600 – -800
-800 – -400	-800 – -400	-800 – -400	-800 – -400
-400 – -300	-400 – -300	-400 – -300	-400 – -200
-300 – -200	-300 – -200	-300 – -200	
-200 – -100	-200 – -100	-200 – 0	-200 – -100
-100 – -50	-100 – -50		-100 – 0
-50 – 0	-50 – 0		
0 – 50	0 – 50		
50 – 100	50 – 100		0 – 100
100 – 200	100 – 200	0 – 200	100 – 200
200 – 300	200 – 300	200 – 300	
300 – 400	300 – 400	300 – 400	200 – 400
400 – 800	400 – 800	400 – 800	400 – 800
800 – 1600	800 – 1600	800 – 1600	800 – 1600
1600 – 3200			

Figure 2: Initial set of prosodic features used, "utep-m1." Start and end times for each window are in milliseconds relative to the patch center.

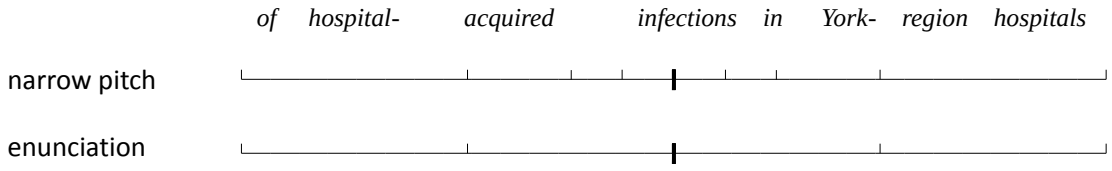

Figure 3: Illustration of how features tile a patch. The vertical lines represent window boundaries.

Among the numerous differences in implementation, many related to robustness considerations, two deserve mention. First, the Geneva set was originally designed to work for inputs that were individual utterances, pre-segmented by hand, whereas the utep-m1 set was designed to be robust for multi-utterance input including regions of silence. Second, the Geneva set was designed to work for emotions, which probably vary little over a patch, and so includes many long-term features. In contrast, the utep-m1 set, thanks to the time-spread windows, may be more sensitive to specific turns of phrase and rhetorical flourishes that happen over just a few words or syllables.

6.4. Initial Results

Using this model we tested our hypothesis, that prosody is informative for stance. We did this using a leave-one-out regime, that is, using cross-validation at the segment level.

	English			Mandarin			Turkish		
	basln.	gen	m1.	basln.	gen	m1.	basln.	gen	m1.
1 Bad	.68	.64	.61*	.35	.27*	.24*	.35	.30*	.31*
2 Good	.54	.49*	.46*	.38	.40	.32*	.15	.14	.14*
3 Deplorable	.40	.41	.36*	x .07	.08	.06	.49	.43*	.42*
4 Praiseworthy	x .06	.07	.06	.15	.16	.13*	.11	.09*	.10*
5 Controversial	x .07	.08	.07	.14	.15	.13*	.39	.35*	.35*
6 Factual ...	x .09	.05*	.07*	.13	.08*	.09*	.58	.38*	.39*
7 Subjective ...	.15	.11*	.13*	.66	.45*	.40*	.48	.40*	.41*
8 Unusual ...	x .08	.08	.07*	.15	.15	.14*	.29	.24*	.26*
9 Typical ...	.64	.49*	.48*	.26	.15*	.11*	.23	.19*	.19*
10 Local	.38	.21*	.25*	.77	.21*	.23*	.15	.12*	.12*
11 Immediate ...	.43	.38*	.36*	.50	.17*	.15*	x .05	.05	.05*
12 Background	.56	.50*	.50*	.49	.31*	.24*	.28	.26	.25*
13 New ...	.41	.31*	.33*	.95	.25*	.23*	.34	.25*	.26*
14 Large-Group ...	.58	.46*	.46*	.59	.37*	.31*	.37	.30*	.30*
average	.36	.31*	.30*	.40	.23*	.20*	.30	.25*	.25*

Table 3: Performance in MSE. The scale of values is 0 to 2. "Basln" is the baseline, "gen" is the Geneva eGeMAPs features, and "m1" is utep-m1. * indicates statistically better than baseline. x indicates low variance (< 0.10), reflecting highly skewed priors.

Specifically, for each segment and each stance, we predicted the value based on the annotations of that stance in other segments, across each entire dataset. We chose this regime because our data sets are modest in size.

One slight complication arose from the existence of very short segments, such as brief weather or stock reports. Because we had difficulty configuring the Geneva features to produce features consistent in meaning across both segments less than 6 seconds in length and longer segments, in this initial experiment we evaluated on only the longer segments. Thus we used subsets of the datasets for these comparisons: for English 877 segments, for Mandarin 306, and for Turkish 1022.

As seen in Table 3, some stance aspects were predicted fairly well. (The results for English here differ from those earlier reported (Ward et al., 2017) because at that time several of the audio files had been incorrectly-downloaded and thus were not identical to the ones on which the annotators had based their judgments.) For all stances and all languages, the model using the utep-m1 features had lower error than the baseline. Regarding our hypothesis, the model often outperformed the baseline by a statistically significant difference, so there is evidence that prosody does have value for predicting stance. This also suggests that our working assumptions are appropriate. Regarding features, the overall success of both sets suggests that this approach is robust and not limited to some specific choices.

Nevertheless, the performance was mixed and sometimes quite marginal. We therefore began to seek to identify the causes of poor performance in some cases. The first, most obvious, cause was skewed distributions: for stances for which the distribution is unbalanced,

performance is generally poor. Indeed, for such cases the annotators also often had difficulty, as seen in Table 2 by the generally low agreement scores for such stances.

The rest of these paper describes how we exploited this initial model to seek a better understanding of how prosody conveys stance, and attempts to use this understanding to build a better model.

7. The Non-Independence of Feature Streams

There are many fundamental open questions regarding how prosody expresses meanings. Applications-oriented work generally seeks to sidestep such questions, and this can be a wise choice when adequate training data is available. However, in order to obtain good performance with little data, as in the current situation, it can be even wiser to seek a better understanding of the phenomena.

One important question is whether individual prosodic cues independently convey meaning, additively, or whether configurations of cues have supra-additive significance. This is a classic question in linguistics, both theoretical and applied (Ladd, 2008; Gravano and Hirschberg, 2011). Most linguistic treatments of prosody have assumed independence. For example, a fundamental assumption of most work in intonation modeling is that the pitch is informative by itself, and there has been much work attempting to identify intonation contours (or pitch-target sequences) that are informative, regardless of how they relate to other prosodic properties. (The only common exception relates to syllable stress, where the interactions between intonation contours and stressed syllable locations, marked in part by intensity and duration, are a topic of recurring interest in phonology.) However, not all models of prosody incorporate such an independence assumption (Couper-Kuhlen, 1986). In particular, the recent "prosodic constructions" approach suggests that prosody actually conveys by means of specific temporal configurations of prosodic features (Ogden, 2010, 2012; Niebuhr, 2014; Ward, 2014).

If configurations are important, then we should use models that can handle them (Kim and Provost, 2013; Ward, 2014; Ferrer et al., 2007). If they are not, that is, if there is independence, then we should simplify our models, to potentially obtain good performance with less training data. (Parenthetically we note that the question of feature-stream independence is less important when we have adequate data: in such cases we can feed any and all features to a powerful model, such as an SVM, random forest, or deep neural network, that can perform well whether or not the features are independent.)

In practical terms, we want to determine whether it is appropriate to model the individual prosodic "streams" as independent, where by streams we mean different types of prosodic features: pitch-related, intensity-related, duration-related, and so on. In other words, we wish to determine the value of late *vs.* early fusion. In general, obtaining independent estimates of some quantity and then combining them, an "ensemble" approach, can improve robustness and thus performance. This is because multiple estimates, whose errors will be uncorrelated if they are truly independent, can enable them to "correct each other" when averaged, giving better performance. However, we hypothesized that here, because specific

prosodic configurations are probably informative, early fusion would outperform late fusion. The stance-inference task provides an opportunity to investigate this question.

The model described above embodies an early-fusion approach, in that all prosodic features are considered at the same time. We therefore built for comparison a late-fusion model, which first uses each stream to estimate the stance, then combines the stream-based estimates. For this we split our feature set into four streams: intensity, pitch, creakiness, and speaking rate. We then perform our usual nearest-neighbors computation on each stream, and then combine them (late fusion) by using the information in the four estimates.

In detail, the late-fusion model first predicts, for each test data segment and each stance, the value using the 3 nearest neighbors from the training data, for each stream, that is, using only the features of that stream. Thus we were, overall, considering 12 nearest neighbors instead of 3, though each one was found using only part of the evidence. We combined the four estimates, one from each stream, using two different methods. In the first we used no weights, for each patch simply averaging the per-stream estimates. In the second we tried to account for the fact that for some inputs one stream may be more informative than another, for example, when in one stream the training data includes very close neighbors for a testset patch. Specifically, for each patch, we estimated the informativeness of a stream as proportional to the inverse of the average distance from the test patch to the 3 nearest neighboring patches, as a fraction of the overall inverse average distance for that stream, as estimated from the 3 nearest neighbors for 1000 random patch samples from the training data. This normalization was important since different streams have different numbers of features, and thus the typical distance varies from stream to stream.

	English	Mandarin	Turkish
early fusion	.292	.198	.250
late fusion, unweighted	.344	.276	.270
late fusion with stream weights	.374	.285	.269

Table 4: Performance in MSE, averaged over all stances.

The results, averaged across all stances, are seen in Table 4. (This and subsequent results are computed over the entire corpora, including short segments. This is why the early-fusion results differ slightly from those in Table 3.)

We see that early fusion outperformed late fusion, both weighted and unweighted. In fact, early fusion performed best for all 42 language-stance pairings. This suggests that specific multi-stream configurations of prosodic features are indeed meaningful. As far as we know, this represents the first corpus-based finding regarding the relative merits of independent-stream and multistream modeling of prosody.

8. Feature Set Improvements

Our initial experiments used two off-the-shelf feature sets. This section describes how we improved on these, both for the sake of better understanding how prosody expresses stance

and for the sake of improving prediction accuracy.

8.1. Feature Selection

Studies with the different streams, each tested in isolation, suggested that the intensity and pitch features were the most valuable, with creakiness contributing little or nothing. The low value of creakiness here, despite its importance in dialog for interpersonal and turn-taking functions, probably reflects the fact that creaky voice negatively affects intelligibility, and so is avoided in radio.

In order to more thoroughly investigate the usefulness of the various features, we constructed decision-tree models of the data and examined the frequency with which each feature was used in the splits. (Although it can be more informative to consider also how high in the trees each feature tended to appear, here we only considered the frequency of appearance.)

Specifically we trained 42 decision trees, one for each of the 14 stances in each of the 3 languages. Despite a good deal of variation, some tendencies were broadly present across all trees. Intensity features were the most frequently used, and the creakiness features consistently ranked at or near the bottom. Further, across all feature types, shorter windows generally ranked lower than wider ones, and in particular the 50 ms windows for pitch highness and pitch lowness were very infrequently used.

We also investigated the value of reducing the feature space using Principal Component Analysis (PCA). PCA is known to be useful for discovering meaningful underlying factors underlying complex prosodic variation (Ward, 2014). It is also, in many cases, effective at separating useful dimensions of variation from lower-ranked dimensions that mostly represent unhelpful noise. We accordingly transformed the raw 88 features to 88 PCA dimensions, and experimented with dropping the lower-ranked dimensions, as a possible way to add noise robustness. However this did not help; we believe this suggest that all dimensions of prosodic variation are sometimes useful for some stances. We also experimented with dropping some of the *highest*-ranked dimensions, as a possible way to overcome irrelevant contextual variation (Belhumeur et al., 1997). This also gave no benefit, probably because our features are designed to be well-normalized already.

Based on the decision-tree observations we dropped one feature type and for others modified the window sizes used. Specifically, we dropped all of the creakiness features and we replaced the 50 ms windows with wider windows, for all feature types except intensity.

8.2. Augmenting the Feature Set

To gain further insight regarding the adequacy of the features we examined some segments where the model's predictions were most incorrect. Poor predictions at the segment level are due to poor predictions for some component patches, which in turn can be due to reference-data patches which are close neighbors, by our metric, but which are different in some stance.

We found that our utep-m1 feature set was sometimes inadequate to distinguish patches which were, perceptually, clearly different. There was, for example, a somber-sounding patch that our features considered to be prosodically similar to a clearly upbeat one, and

we attributed this to the absence of features able to catch the difference in emotional tone. We also found a patch clearly expressing a deploring tone that matched one with a positive tone, and attributed this to the lack of a feature able to represent that only one of these had onset lengthening. Together with the fact that the Geneva features outperformed the utep-m1 set in some cases, these observations motivated us to try additional features.

As our initial set, utep-m1, provided decent coverage of the big-three aspects of prosody — pitch, intensity, and timing — we chose to focus on features of other types. Because previous work has said little about what other features might convey stance, and because we lack intuitions about this, we choose features for non-redundancy and variety. However, we chose to exclude purely spectral features, as those could lead to a model that was sensitive to the presence of specific lexical items, which would likely be less robust across languages.

First we added a feature for syllable lengthening, to complement our energy-flux estimate of speaking rate. Lengthening is estimated as high in windows over which the frame-to-frame spectral change is low, as for example occurs in lengthened vowels. Specifically, it is the inverse of the squared sum of the cepstral differences, multiplied by the energy in that window, to reduce misdetections due to silent regions. Lengthening is known to be important in turn-taking and in marking high-entropy words (Bell et al., 2009).

We added a late-pitch-peak feature. Late pitch peak is the phenomenon in which the pitch peak occurs after the energy peak of a syllable; that is, later than its normal time (Barnes et al., 2012). While various measures of late peak exist, we developed a new, automatically-computable proxy, disalignment, that measures the extent to which there are salient pitch peaks and energy peaks which are close but not aligned. (The vast majority of non-aligned peaks are in fact late peaks.) Late pitch peaks are known to be involved in expressing politeness, attitudes like incredulity, and topic starts (Wichmann et al., 1997; Zellers et al., 2009).

We added the voiced/unvoiced intensity ratio, a measure of the difference between the average intensity over voiced frames and the average intensity over speech-containing but unvoiced frames. This is known to be important in emotion detection.

We added enunciation and reduction, to represent the degree to which the speech is carefully articulated versus slurred. While various measures exist, we developed new, fully-automatic features. These take the average cepstum across all voiced regions as an estimate of the neutral sound — presumably nearly a schwa or other central vowel — and then in each window compute the average distance of the frames' cepstrums from this neutral cepstrum. When high this indicates enunciation, and when low reduction. Articulatory precision is known to vary with information predictability, and hyperarticulation to mark opinions and urgency (Freeman, 2014; Hodoshima, 2016).

Figure 4 shows the improved set, utep-m4. After the various deletions, modifications, and additions, this coincidentally also has 88 features.

As seen in the second and third lines of Table 5, this new set outperformed utep-m1.

Curious about the relative contributions of the new features, we did an additional post-hoc experiment. Most discussions of prosody focus on big-three features, but we thought that the additional features would bring a sizeable benefit. To explore this we built a "big-three" featureset by including all intensity, pitch, and lengthening features from utep-m4.

intensity (16)	high pitch, low pitch narrow pitch, wide pitch energy flux, lengthening late pitch peak, v/uv intensity (8 each)	enunciation reduction (4 each)
-3200 – -1600		
-1600 – -800	-1600 – -800	-1600 – -800
-800 – -400	-800 – -400	
-400 – -300		
-300 – -200	-400 – -200	
-200 – -100		
-100 – -50		
-50 – 0	-200 – 0	-800 – 0
0 – 50	0 – 200	0 – 800
50 – 100		
100 – 200	200 – 400	
200 – 300		
300 – 400		
400 – 800	400 – 800	
800 – 1600	800 – 1600	800 – 1600
1600 – 3200		

Figure 4: Final set of prosodic features used, "utep-m4." Start and end times for each window are in milliseconds relative to the patch center.

Then we added features, one-by-one. As seen in the bottom half of Table 5, the contributions of the various features were small and variable, contrary to our expectation. The only new feature consistently giving a benefit was the reduction feature. We conclude that most of the advantage of our new utep-m4 set comes not from the additional features but from the other improvements: adjusted window sizes, deletion of the creaky feature, and the improved rate (lengthening) feature.

While utep-m4 improved on utep-m1, we must add a caveat: because of our numerous exploratory experiments, some knowledge of the specific properties of these datasets leaked in to our process of developing utep-m4. Thus the increases in performance seen here may overestimate what we would see on new data, although we lack pristine held-out data available to confirm this.

Nevertheless utep-m4 does seem to be well-suited for stance inference. While we expect that further improvements are possible, at this point we stopped experimenting with features, to instead work on improving the model.

	English	Mandarin	Turkish
baseline	.362	.395	.305
utep-m1	.292	.198	.250
utep-m4	.268	.193	.244
big three	.277	.190	.245
+ creak	.283	.193	.249
+ vvir	.274	.198	.246
+ late peak	.280	.190	.244
+ enunciation	.275	.191	.243
+ reduction	.270	.187	.243

Table 5: Performance of various feature sets, MSE

9. Model Improvements

This section briefly describes three ideas, one of which substantially improved performance.

Our first idea was to more effectively use the information provided by the nearest neighbors. Thus, to estimate the stance at each testset patch, we built a model based on some similar training data patches. Specifically we tried locally-weighted linear regression, using up to 100 neighbors for each testset patch. As this gave about the same level of performance, we did not pursue this further.

Our second idea was based on observations that some stories seemed to have distinctive prosodic indications of the overall stance on the very first words. This led us to suspect that the same prosodic configuration occurring early in a news story might have a different significance from when occurring a little later or in the middle or near the end. We implemented this simply, within the nearest-neighbors framework, by simply adding temporal features: thus obtaining a model that tended to find neighbors that were not only prosodically similar but also in roughly the same temporal zone in the story. Thinking that this would be most important for patches very early or very late in stories, we chose to use log time, and accordingly added two new features: log of the time since the start of the segment and log of the time until the end of the segment. We also tried other temporal features, such percent of time into a segment, but these gave little or no benefit. The two log features, however gave a large improvement, so to better exploit this we tried giving them larger weights, obtaining best performance for English when each had a weight 4 times that of the other features. Table 6 shows the resulting performance. The last, best-performing combination we will refer to as utep-m4t.

Our third investigation was motivated by the observation that speakers differed in their prosodic behavior. We therefore experimented to see whether performance would improve with matched data, where all the reference data was from same speaker as the to-be-classified segment. For this we used the single-speaker subset of the English data. Table 7 shows the results. Comparing to the general (multi-speaker) case, matched data gave better performance, even with only a quarter of the data. Comparing to a same-size subset, selected

	English	Mandarin	Turkish
utep-m4	.268	.193	.244
+ log(time-since-start) and log(time-to-end)	.246	.179	.232
+ log(time-since-start) and log(time-to-end), x 4	.202	.162	.230

Table 6: Performance with and without temporal features, MSE

	baseline MSE	model MSE	percent reduction
all speakers (488 minutes)	.36	.20	44%
mixed speakers (126 minutes)	.37	.24	33%
single speaker (125 minutes)	.35	.17	50%

Table 7: Performance as a function of training-data size and similarity, English.

at random, matched data boosted performance (reduced error) by a factor of 1.5. However in practice single-speaker models would be useful only if speaker identification were reliable and if adequate samples of the speaker were present in the reference data, conditions that rarely hold, so we did not pursue this further.

10. How Prosody Conveys Stance

This section illustrates specific prosodic configurations that do or do not convey stance. It further discusses attempts, so far unsuccessful, to use these observations to build better models.

While examining poor predictions we discovered some prosodic configurations unrelated to stance. For example, the prosody at one appositive-comma pause strongly resembled the prosody at an appositive-comma pause in a different segment, regardless of the very different stances in the two segments overall (Wichmann, 2000). As another example, the prosodic construction indicating contrast (which is the monolog form of the same contradiction contour often noted in dialog (Liberman and Sag, 1974; Hedberg et al., 2003)), occurred in rhetorical structures involving various stances. The existence of such cases is not surprising; it simply reflects the fact that there are times when prosody is being used to convey things other than stance. For our pruposes, this means that some patches are not informative, and should somehow be excluded from consideration.

While patch-based failure analysis can lead to insight, it may lead to outlier phenomena rather than to more common sources of good or bad performance. Accordingly we next set out to examine the meanings of prosody using clusters. To do this we first used k-means over all patches in the English set to find 100 clusters. For this we used the 88 utep-m4 prosodic features. The stance labels were not used in the clustering. Inspired by earlier work on bag-of-audio-words models (Schmitt et al., 2016), one might think of each cluster as an appoximation to a "prosodic word," to the extent that the variation within a cluster

is less important for meaning than the differences between clusters.

To investigate which stances were being conveyed by which clusters, we built a very simple predictive model. This model represents each story as a vector of cluster-counts, based on how many of the patches in that story fall within each cluster, that is, for each cluster j the count of patches in the to-be-classified segment that are closest to the centroid of cluster j. For example, if a story contained 7 patches that fall within cluster 1, 3 that fall within cluster 2, and so on, that story was represented by the vector (7, 2 ...). Thus we represent each story as a "bag of prosodic words." We then trained a model to predict stance. For interpretability this model used linear regression over the cluster counts.

We then examined the coefficients of this model to find clusters that were informative regarding one stance or another. To give three examples:

- Stories with many patches falling within Cluster 1 tended to be annotated highly on the "new" stance, that is, to contain new information. To understand what Cluster 1 involves, we plotted the feature values at its centroid, as seen in Figure 5. We then listened to some patches close to this centroid. While the cluster involves many prosodic features, a few were clearly salient: there is a brief pause near the center (the dip in intensity), the sentence before the pause ends with a slightly raised and somewhat narrow pitch, and the sentence after the pause starts with a high pitch, a clearly enunciated word or two, and a very salient late pitch peak.

  One patch near the centroid was in *... the plant ... has had its operations halted ... [due to] permit issues. The plant expects to resume operations again in around six months ...* and another patch was in *... Markham is testing out some new designs ... [for] wheelchair-accessible picnic tables. Ten accessible tables are being placed in various parks ....* In both cases the newsreader was transitioning from giving background information to providing new information. Thus it makes sense that a segment with many patches like this is likely to be conveying new information. As noted above, late pitch peaks have previously been implicated with new topic start, a related function, but here, in concert with other features in the local neighborhood, it seems to convey something more specific. This configuration has two other interesting properties. First, it spans an utterance boundary, meaning that approaches which try to classify stance utterance-by-utterance would likely miss it (Perez-Rosas et al., 2013). Second, it can convey newness even when, as in these examples, there are no lexical indications of newness.

- Stories rich in patches from Cluster 35 tended to be "local" in stance. At the centroid this involved about a second of slightly increased articulatory precision and slightly narrowed pitch range, followed by about 400 milliseconds of moderately high intensity, quite narrow pitch range, and extreme lengthening. For examples, this was present on a patch on *under partly sunny skies this afternoon*, a patch in *MS, formerly of Sweet Home [Oregon], was shot in the leg ...*, and a patch in *for local news ... head over to our website ...*, where underlining marks the lengthened word in each case. Perceptually, in each case it seems as if the announcer is more directly speaking to the

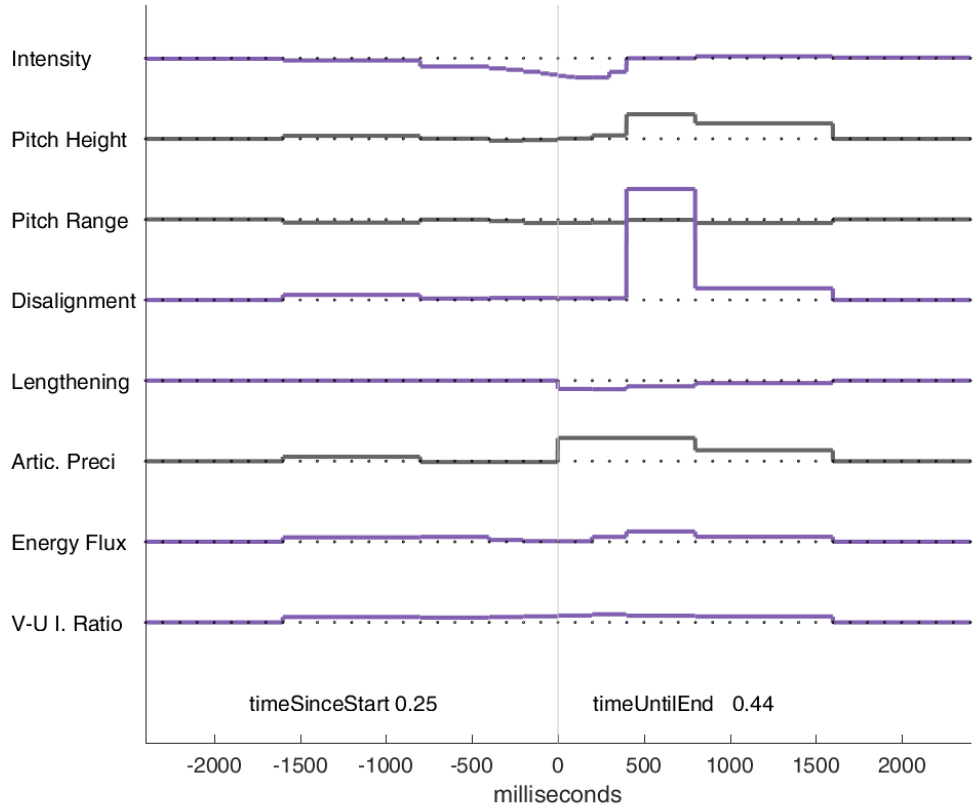

Figure 5: Feature values at the centroid of Cluster 1. Each feature's values are plotted over that feature's time range, thus for example over the window –1600 ms to –800 ms the intensity is slightly below average. The dotted lines in each case are at zero. All features are z-normalized. For conciseness three pairs of feature types are combined: the difference between the values of the high-pitch and low-pitch features is shown as "pitch height," the difference between wide pitch and narrow pitch as "pitch range", and the difference between enunciation and rediction as "articulatory precision.

audience about something of shared interest. Unlike the previous example, this configuration does not appear to be linked to any specific structure – discourse, syntactic or otherwise — so it may be a relatively free-floating prosodic morpheme, adding a nuance of "relevant to you, the local listening audience" wherever it occurs.

- Our final example is cluster 36, which is an indicator for stories that are "relevant to a large group." Patches in this cluster are fairly low in pitch over about 3 seconds with a narrow pitch range, and a slight pause in the middle of that region. Examples near the centroid included a patch at *feared . . . a release of radioactivity. The government announcement may signal the danger . . .*, a patch at *. . . police officers . . . throughout the county, as a deterrent . . .*, and a patch at *. . . fundraising goal is one hundred fifty nine thousand dollars. . .* Thus this configuration can appear when the speaker is conveying that the situation is not limited or merely local, but covering an large region. However this configuration is not specific to the "large group" stance, as seen in the

third example, in which it seemes to convey merely a large amount. This configuration is easy to understand as a special case of the general tendency for lower pitch to relate to increased physical size in many ways (Ohala, 1984).

Thus there are clearly configurations of prosodic features that convey stance, supporting again the validity of our two working assumptions. The existence of such configurations raises questions for linguistic models of prosody (Ladd, 2008); examining the theoretical implications would be an interesting topic for future work.

Theoretical issues aside, we tried to use these observations to build a better model. Because we learned that patches are in fact not all equally informative regarding stance, we attempted to estimate the informativeness of each patch and use this informativeness to weight their estimates. We estimated informativeness based on whether the close neighbors were consistent in stance. However this gave only a modest benefit.

Our next attempt to build a better model used the prosodic bag of words representation. As described above, the basic idea is to estimate stance via the properties of clusters, where each cluster center represents the information in hundreds or thousands of patches. Because some clusters have insignificant correlates with some stances, this representation can handle the fact that some patches are less informative that others with respect to some stance. To improve performance, we made two changes to the simple model described earlier. First, based on the expectation that there would be diminishing returns in informativeness with increasing counts; we replaced the raw counts c_j with $\log(1+c_j)$ for each cluster. Second, because in the simple model the predictions could be less than 0 or more than 2, we added postprocessing to clip such values, to force them into the known range of target values. Preliminary experiments indicated that these changes improved performance, and that the temporal features were useful in this model also, but performed best without extra weighting. Other preliminary experiments suggested that performance would improve with more clusters, but for the sake of quick experimentation we chose to use only 200 clusters. Although creating the clusters is time-consuming, after the clusters are fixed and the model is built, it is very quick to compute predictions for a new story. Preliminary experiments also showed that the bag-of-prosodic-words representations needed to be language-specific: a model for Mandarin using English-derived clusters performed only slightly above baseline. As an alternative to the linear regression over the prosodic bags of words, we also tried nearest neighbors over prosodic bags of words, where for each segment the estimate was derived from the stances of the three nearest neighbors in bag-of-prosodic-words space, distance-weighted as described in Section 6.2.

To evaluate these new models we used five-fold cross-validation. The clustering was done for all the data, training and test, but the regression model was trained separately for each fold, reflecting a real-world scenario where there is a lot of data but only a fraction has been annotated. As seen in Table 8 the prosodic-words models did not outperform the k nearest neighbors model. We think this indicates any benefit from modeling informativeness differences was overwhelmed by the loss of information in the clustering step.

	English	Mandarin	Turkish
baseline	.36	.39	.31
patches→kNN per patch→average	.20	.16	.23
patches→cluster counts (BOPW)→regression	.27	.46	.29
patches→cluster counts (BOPW)→kNN	.25	.22	.29

Table 8: Performance of the two models. BOPW is a bag of prosodic words

11. Final Results

Figure 9 shows the results for each stance and each language for our best model, utep-m4t. We observe that the performance is good overall: across the three languages the models do well for most stances, subject again to the caveat regarding the lack of pristine test data.

	English			Mandarin			Turkish		
	basln	m4t	hum.	basln	m4t	hum.	basln	m4t	hum.
1 Bad	.65	.41	.11	.35	.24	.14	.36	.30	.24
2 Good	.53	.35	.28	.38	.28	.19	.15	.13	.09
3 Deplorable	.37	.24	.06	.06	.06	.04	.48	.39	.14
4 Praiseworthy	.05	.05	.04	.15	.13	.08	.11	.10	.07
5 Controversial	.07	.05	.03	.14	.13	.05	.39	.31	.23
6 Factual ...	.08	.06	.08	.12	.08	.05	.60	.32	.26
7 Subjective ...	.14	.08	.08	.66	.28	.26	.48	.36	.25
8 Unusual ...	.07	.06	.07	.15	.14	.03	.29	.25	.22
9 Typical ...	.74	.22	.10	.25	.09	.08	.23	.18	.30
10 Local	.37	.22	.26	.77	.19	.23	.15	.10	.38
11 Immediate ...	.41	.28	.09	.49	.10	.07	.05	.05	.08
12 Background	.57	.21	.27	.49	.18	.16	.28	.21	.29
13 New ...	.41	.25	.32	.95	.11	.05	.35	.23	.48
14 Large-Group ...	.60	.33	.32	.58	.26	.07	.38	.28	.35
average	.362	.202	.151	.395	.162	.107	.305	.230	.244

Table 9: Performance of the final models, MSE. m4t is the model averaging patchwise kNN-baed estimates using the utep-m4t feature set; hum. is human performance.

For some of the stances the model did better than the human annotators. This is possible because the target in each case is the average of the three annotators' judgments, and any individual annotator may diverge from that target, meaning that the average error for the humans may be higher than the error of the model. This was seen for locally-relevant in all languages. This was also seen for 6 of the 14 stances for Turkish, although this result largely reflects the low interannotator agreement for this language.

Nevertheless there is clearly much room to improve. Table 10 puts these results in perspective, showing what fraction of the stance information present was inferred. Overall

it is possible to infer on average between 25 and 59% of the stance information present in news stories. Sampling a few segments which were badly misclassified, most had a disconnect between the stance as conveyed by prosody and the stance that would be suggested by the facts. Thus, we think that the major reason why our model performs less well than humans is, unsurprisingly, that the human annotators were able to consider information beyond prosody, including the words said, the full context, and their knowledge of the world and how news readers talk about it.

	English	Mandarin	Turkish
kNN utep-m4t	44%	59%	25%
human	58%	73%	20%

Table 10: Obtained improvement as a percent of the possible improvement, averaged across all stances.

The table also supports an additional observation: performance varies across languages. However, given the peculiarities of the data sets, as described above, this is likely due more to genre differences than to language differences.

12. Stance Expressions across Languages

Disasters can occur anywhere, without warning, so a stance-inference system would ideally be able to detect stance in news from any language, even if not previously modeled. Thus it would be convenient if expressions of stance were universal. To explore this we did cross-language experiments, in which everything was the same, except that the nearest neighbors for reference were sought in the data from a different language. Unfortunately, as seen in Table 11, the cross-language performance was very poor: at or below baseline. Clearly the prosodic reflections of stance can vary greatly among languages.

13. Open Questions

This section considers some open questions and avenues for future work.

The feature set could doubtless be further improved. One might add aligned features, aligned with boundaries or with words or syllables. One might try more features (Ferrer et al., 2010; Slaney et al., 2013; Arsikere et al., 2016; Levow and Wright, 2017), including

	Performance when trained on		
	English	Mandarin	Turkish
English	44%	-11%	0%
Mandarin	-1%	59%	1%
Turkish	-4%	-20%	25%

Table 11: Performance across languages, as above.

also spectral and lexical features. One could do exhaustive feature selection, and refine the model to have different weights for all features. Since we know, from preliminary studies, that different feature sets are advantageous for for different stances, one could also micro-tune the features independently for predicting each stance.

One might also try new models, including exemplar-based models more sophisticated than clustering, perhaps with a bag-of-words model based on soft instead of hard clustering (Muscariello et al., 2009; Kim and Provost, 2013). One might also try to classify patches using a model more sophisticated than nearest neighbors. One might also abandon the assumption of independent patches to instead model the temporal relations across wider ranges (Poria et al., 2017). One might use some kind of attention model to overcome the simplicity of assuming that all patches are equally informative (Yu et al., 2017).

One might explore the generality of the configurations and models, including the effects of data size, of speaker differences, of domains, of topics, and of genres, such as read news, interviews, speeches, dialog, and video soundtracks.

Another way to extend this work would be to model stance in a more fine-grained way (Yang and Cardie, 2013; Socher et al., 2013). Rather than infering the stance of stories, one could seek to infer the stance of the speaker towards the various entities, situations, and activities discussed.

Although the lack of commonality between the prosody-stance mappings of these three languages suggests that expressions of stance are not universal, future work should examine generality within language families. It is also possible that there are universal constraints in how prosody maps to stance; for example, certain types of feature configurations may be universally irrelevant to stance, and if these can be discovered, we should be able to obtain good performance on new languages with less training data.

To enable others to also explore these issues, we have made available our code, at `https://github.com/nigelgward/stance/` and `https://github.com/nigelgward/midlevel/`.

14. Summary and Significance

This paper has explored the potential for using stance in information retrieval of spoken language. It presented a list of 14 aspects of stance that are often relevant properties of news stories. It showed that specific configurations of prosodic features over 6 seconds or less are important for conveying stance. It showed how these can be used to infer stance automatically, from prosodic information alone, performing sometimes as well as human annotators.

References

Arsikere, H., Sen, A., Prathosh, A. P., Tyagi, V., 2016. Novel acoustic features for automatic dialog-act tagging. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6105–6109.

Barnes, J., Veilleux, N., Brugos, A., Shattuck-Hufnagel, S., 2012. Tonal center of gravity: A global approach to tonal implementation in a level-based intonational phonology. Laboratory Phonology 3, 337–382.

Belhumeur, P. N., Hespanha, J. P., Kriegman, D. J., 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 711–720.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., Jurafsky, D., 2009. Predictability effects on durations of content and function words in conversational English. Journal of Memory and Language 60, 92–111.

Biber, D., Staples, S., 2014. Exploring the prosody of stance: Variation in the realization of stance adverbials. In: Raso, T., Mello, H. (Eds.), Spoken corpora and linguistic studies. John Benjamins, pp. 271–294.

Cardoso, P. C. F., Taboada, M., Pardo, T. A. S., 2013. On the contribution of discourse structure to topic segmentation. Proceedings of the Special Interest Group on Discourse and Dialogue (SIGDIAL), 92–96.

Chindamo, M., Allwood, J., Ahlsen, E., 2012. Some suggestions for the study of stance in communication. In: IEEE International Conference on Social Computing (SocialCom). pp. 617–622.

Cotter, C., 1993. Prosodic aspects of broadcast news register. In: 19th Annual Meeting of the Berkeley Linguistics Society. pp. 90–100.

Couper-Kuhlen, E., 1986. An introduction to English prosody. Edward Arnold.

DARPA, 2014. Low resource languages for emergent incidents (LORELEI), Solicitation Number DARPA-BAA-15-04.

Elfardy, H., Diab, M., 2016. Addressing annotation complexity: The case of annotating ideological perspective in Egyptian social media. In: 10th Linguistic Annotation Workshop. pp. 79–88.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al., 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing 7, 190–202.

Eyben, F., Wöllmer, M., Schuller, B., 2010. OpenSmile: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the International Conference on Multimedia. pp. 1459–1462.

Ferrer, L., Scheffer, N., Shriberg, E., 2010. A comparison of approaches for modeling prosodic features in speaker recognition. In: IEEE Acoustics Speech and Signal Processing (ICASSP). pp. 4414–4417.

Ferrer, L., Shriberg, E., Kajarekar, S., Sonmez, K., 2007. Parameterization of prosodic feature distributions for SVM modeling in speaker recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing. Vol. 4. pp. IV:233–236.

Freedman, M., Baron, A., Punyakanok, V., Weischedel, R., 2011. Language use: what can it tell us? In: 49th Association for Computational Linguistics, Volume 2. pp. 341–345.

Freeman, V., 2014. Hyperarticulation as a signal of stance. Journal of Phonetics 45, 1–11.

Freeman, V., Levow, G.-A., Wright, R., Ostendorf, M., 2015. Investigating the role of yeah in stance-dense conversation. In: Interspeech. pp. 3076–3080.

Freese, J., Maynard, D. W., 1998. Prosodic features of bad news and good news in conversation. Language in Society 27, 195–219.

Gravano, A., Hirschberg, J., 2011. Turn-taking cues in task-oriented dialogue. Computer Speech and Language 25, 601–634.

Gussenhoven, C., 2002. Intonation and interpretation: phonetics and phonology. In: Speech Prosody. pp. 47–57.

Hedberg, N., Sosa, J. M., Fadden, L., 2003. The intonation of contradictions in American English. In: Prosody and Pragmatics Conference.

Hodoshima, N., 2016. Effects of urgent speech and preceding sounds on intelligibility in noisy and reverberant environments. In: Interspeech. pp. 1696–1700.

Huang, S., et al., 1998. Mandarin Broadcast News Speech (HUB4-NE). Linguistic Data Consortium, catalog No. LDC98S73, ISBN: 1-58563-125-6.

Kim, Y., Provost, E. M., 2013. Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3677–3681.

Ladd, D. R., 2008. Intonational Phonology, second edition. Cambridge University Press.

Larson, M., Jones, G. J. F., 2012. Spoken content retrieval: A survey of techniques and technologies. Foundations and Trends in Information Retrieval 5 (4-5), 235–422.

Lee, L.-S., Glass, J., Lee, H.-Y., Chan, C.-A., 2015. Spoken content retrieval: Beyond cascading speech recognition with text retrieval. IEEE/ACM Transactions on Audio, Speech, and Language Processing 23, 1389–1420.

Levow, G.-A., Wright, R., 2017. Exploring dynamic measures of stance in spoken interaction. In: Interspeech. pp. 1452–1456.

Liberman, M., Sag, I., 1974. Prosodic form and discourse function. In: Papers from Tenth Regional Meeting, Chicago Linguistic Society. pp. 402–427.

Liu, B., Zhang, L., 2012. A survey of opinion mining and sentiment analysis. In: Mining Text Data. Springer, pp. 415–463.

Liu, S.-H., Chen, K.-Y., Chen, B., Wang, H.-M., Yen, H.-C., Hsu, W.-L., 2015. Positional language modeling for extractive broadcast news speech summarization. In: Interspeech. pp. 2729–2733.

Mairesse, F., Poifroni, J., Di Fabbrizio, G., 2012. Can prosody inform sentiment analysis? Experiments on short spoken reviews. In: IEEE ICASSP.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C., 2016. SemEval-2016 task 6: Detecting stance in tweets. Proceedings of SemEval 16.

Morency, L.-P., Mihalcea, R., Doshi, P., 2011. Towards multimodal sentiment analysis: harvesting opinions from the web. In: ICMI: 13th International Conference on Multimodal Interfaces. pp. 169–176.

Muscariello, A., Gravier, G., Bimbot, F., 2009. Audio keyword extraction by unsupervised word discovery. In: Interspeech.

Niebuhr, O., 2014. Resistance is futile: The intonation between continuation rise and calling contour in German. In: Interspeech. pp. 132–136.

Ogden, R., 2010. Prosodic constructions in making complaints. In: Barth-Weingarten, D., Reber, E., Selting, M. (Eds.), Prosody in Interaction. Benjamins, pp. 81–103.

Ogden, R., 2012. Prosodies in conversation. In: Niebuhr, O. (Ed.), Understanding Prosody: The role of context, function, and communication. De Gruyter, pp. 201–217.

Ohala, J. J., 1984. An ethological perspective on common cross-language utilization of F_0. Phonetica 41, 1–16.

Perez-Rosas, V., Mihalcea, R., Morency, L.-P., 2013. Utterance-level multimodal sentiment analysis. In: ACL. pp. 973–982.

Pillet-Shore, D., 2012. Greeting: Displaying stance through prosodic recipient design. Research on Language & Social Interaction 45, 375–398.

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.-P., 2017. Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. pp. 873–883.

Purver, M., Dowding, J., Niekrasz, J., Ehlen, P., Noorbaloochi, S., Peters, S., 2007. Detecting and summarizing action items in multi-party dialogue. In: Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue. pp. 200–211.

Rambow, O., Wiebe, J., 2015. Sentiment and belief: How to think about, represent, and annotate private states. In: Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL.

Read, J., Carroll, J., 2012. Annotating expressions of appraisal in English. Language Resources and Evaluation 46, 421–447.

Reichel, U. D., Kleber, F., Winkelmann, R., 2009. Modelling similarity perception of intonation. In: Interspeech.

Rilliard, A., Allauzen, A., Boula de Mareuil, P., 2011. Using dynamic time warping to compute prosodic similarity measures. In: Interspeech.

Schmitt, M., Ringeval, F., Schuller, B., 2016. At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. In: Interspeech, San Francsico, USA. pp. 495–499.

Schuller, B., 2011. Voice and speech analysis in search of states and traits. In: Salah, A. A., Gevers, T. (Eds.), Computer Analysis of Human Behavior. Springer, pp. 227–253.

Shen, Y., He, X., Gao, J., Deng, L., Mesnil, G., 2014. A latent semantic model with convolutional-pooling structure for information retrieval. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. pp. 101–110.

Slaney, M., Shriberg, E., Huang, J.-T., 2013. Pitch-gesture modeling using subband autocorrelation change detection. In: Interspeech. pp. 1911–1915.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., Potts, C., 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In: Empirical Methods in Natural Language Processing. pp. 1631–1642.

Vaissiere, J., 2008. Perception of intonation. In: Pisoni, D., Remez, R. (Eds.), The handbook of speech perception. John Wiley & Sons, pp. 236–263.

Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., Schram, A., Anderson, K. M., 2011. Natural language processing to the rescue? Extracting "situational awareness" tweets during mass emergency. In: International Conference on Web and Social Media.

Ward, N. G., 2014. Automatic discovery of simply-composable prosodic elements. In: Speech Prosody. pp. 915–919.

Ward, N. G., 2016. Preliminaries to a study of stance in news broadcasts. Tech. Rep. UTEP-CS-16-66, University of Texas at El Paso, Department of Computer Science.

Ward, N. G., 2017. Midlevel prosodic features toolkit, https://github.com/nigelgward/midlevel.

Ward, N. G., Carlson, J. C., Fuentes, O., Castan, D., Shriberg, E., Tsiartas, A., 2017. Inferring stance from prosody. In: Interspeech, submitted.

Ward, N. G., Gallardo, P., 2017. Non-native differences in prosodic construction use. Dialogue and Discourse 8, 1–31.

Ward, N. G., Vega, A., 2012. A bottom-up exploration of the dimensions of dialog state in spoken interaction. In: 13th Annual SIGdial Meeting on Discourse and Dialogue.

Ward, N. G., Vega, A., Baumann, T., 2011. Prosodic and temporal features for language modeling for dialog. Speech Communication 54, 161–174.

Ward, N. G., Werner, S. D., Garcia, F., Sanchis, E., 2015. A prosody-based vector-space model of dialog activity for information retrieval. Speech Communication 68, 86–96.

Wichmann, A., 2000. Intonation in text and discourse: Beginnings, middles and ends. Longman.

Wichmann, A., House, J., Rietveld, T., 1997. Peak displacement and topic structure. In: Intonation: Theory, Models, and Applications. ISCA, pp. 329–332.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., Patwardhan, S., 2005. Opinionfinder: A system for subjectivity analysis. In: Proceedings of HLT/EMNLP on interactive demonstrations. pp. 34–35.

Wollmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., Morency, L.-P., 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. Intelligent Systems, IEEE 28, 46–53.

Yang, B., Cardie, C., 2013. Joint inference for fine-grained opinion extraction. In: ACL (1). pp. 1640–1649.

Yu, H., Gui, L., Madaio, M., Ogan, A., Cassell, J., Morency, L.-P., 2017. Temporally selective attention model for social and affective state recognition in multimedia content. In: ACM Multimedia. pp. 1743–1751.

Zellers, M., Post, B., D'Imperio, M., 2009. Modeling the intonation of topic structure: two approaches. In: Interspeech. pp. 2463–2466.