INFERRING STANCE FROM PROSODY

Nigel G. Ward*,†, Jason C. Carlson*, Olac Fuentes*, Diego Castan‡, Elizabeth E. Shriberg†, A. Tsiartas‡

*University of Texas at El Paso, Computer Science, El Paso, Texas, 79968
† Kyoto University, Media Center, Kyoto, 606-8501, Japan
‡ SRI International, Speech Technology and Research Lab, Menlo Park, California 94035

ABSTRACT

Speech conveys many things beyond content, including aspects of stance and attitude that have not been much studied. We looked at 14 aspects of stance as they occur in radio news stories in English and Mandarin, and investigated whether they could be inferred from prosody. By using time-spread prosodic features and by aggregating local estimates, many aspects of stance were at least somewhat predictable, with results significantly better than chance across many items, even when the categorization of any specific item is not highly reliable.

Table 1 shows the 14 aspects of stance considered in this work. This list reflects several considerations: opinions by some Lorelei program participants regarding likely utility, non-redundancy to what might be accomplished by topic-based retrieval, commonality of occurrence in disaster-related news stories, and ability to be reliably annotated.

This paper explores the 14 novel aspects of audio documents as they occur in English and Mandarin radio news, and reports experiments on automatically detecting them from prosody.

1. MOTIVATION

While most work on speech for information retrieval and filtering has focused on topic and content, with some attention paid to a few other facets — notably including emotion, sentiment, and dialog acts — these do not exhaust the aspects that one might use for information retrieval [1, 2, 3, 4, 5]. In this paper we consider stance. In the social sciences, “stance” refers to a very broad set of feelings and behaviors [6, 7, 8], including all the nuances and subtleties of attitudes and related functions that people display in the course of pursuing various communicative goals. Previous work on modeling stance has examined only a few aspects, such as polarity and strength of opinion [9, 10]; in this work we examine a larger set.

The practical motivation for this work is to support filtering and finding patterns in news broadcasts. In the Lorelei scenario [11], a mission planner needs to find information relevant to organizing a humanitarian intervention after a natural or anthropomorphic disaster. Given the large volume of news broadcasts and social media communications that may potentially be relevant, analysts and planners need tools to organize these to gain situational awareness and support planning. Relevant aspects of these items include attitudes towards situations and facts, evaluations of different actors, the novelty of the information, the magnitude of the disaster, whether an input is well-informed or speculative, and its immediacy in time and place to the disaster and relief needs. Crucially, planners often work with data in languages where ASR and MT tools are rudimentary. They also frequently need big-picture information, such as in which valley the chatter about flooding has a more here-and-now stance. They may use information obtained from statistics and tendencies across many items, even when the categorization of any specific item is not highly reliable.

Table 1 shows the 14 aspects of stance considered in this work. This list reflects several considerations: opinions by some Lorelei program participants regarding likely utility, non-redundancy to what might be accomplished by topic-based retrieval, commonality of occurrence in disaster-related news stories, and ability to be reliably annotated.

Table 1 shows the 14 aspects of stance considered in this work. This list reflects several considerations: opinions by some Lorelei program participants regarding likely utility, non-redundancy to what might be accomplished by topic-based retrieval, commonality of occurrence in disaster-related news stories, and ability to be reliably annotated.

This paper explores the 14 novel aspects of audio documents as they occur in English and Mandarin radio news, and reports experiments on automatically detecting them from prosody.

2. DATA

To investigate the manifestations of these stance aspects, we assembled two sets of radio news broadcasts. The American English set is 500 minutes found at archive.org, consisting mostly of broadcasts from 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. The Mandarin data is the first 250 minutes of the KAZN subset of the Hub4 collection [13].

Each news broadcast was divided into news 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 varied in length from tens of seconds to a few minutes. Each segment was annotated for the presence of each stance aspect as 0 (absent), 1 (weakly present) or 2 (strongly present). Each stance was labeled independently, thus a given segment could be labeled, for example, both deplorable and praiseworthy, if it mentioned both a deplorable act and a praiseworthy one.

Annotation for each language was done by three people working independently. Their agreement levels were measured with 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 1, interannotator agreement was excellent for some stances and poor for others, depending on the language. Assuming that stances are, ultimately, continuous-valued phenomena, we use the average of the three annotators’ labels as the “true”
1. Bad Implications - information with undesirable consequences, such as a raise in taxes, an approaching storm, or a flood (0.72) [0.67 0.07 0.26]; (0.55) [0.85 0.01 0.14]

2. Good Implications - the opposite, such as a peace agreement, a good harvest, or nice weather (0.46) [0.66 0.06 0.28]; (0.45) [0.71 0.13 0.16]

3. Deplorable Action - something bad done by someone or some organization (0.74) [0.86 0.03 0.12]; (0.37) [0.97 0.01 0.02]

4. Praiseworthy Action - the opposite: something good done by someone or something (0.35) [0.86 0.03 0.12]; (0.46) [0.92 0.02 0.06]

5. Controversial - something people do or could disagree about, such as a bold action by some person or group, or new government policy (0.53) [0.96 0.02 0.02]; (0.56) [0.95 0.00 0.05]

6. Factual Information - information presented as facts (0.25) [0.04 0.03 0.94]; (0.59) [0.04 0.02 0.94]

7. Subjective Information - the opposite, such as opinions, either the presenter’s or someone else’s, or information reported skeptically or speculatively (0.36) [0.89 0.06 0.04]; (0.55) [0.54 0.06 0.40]

8. Unusual or Surprising - something quirky, odd, or unexpected (0.22) [0.94 0.03 0.03]; (0.69) [0.93 0.03 0.04]

9. Typical or Unsurprising - the opposite, something expected (0.81) [0.69 0.01 0.30]; (0.59) [0.91 0.00 0.09]

10. Local - personally relevant to the listening audience, like local weather or close-by rioting (0.39) [0.20 0.03 0.77]; (0.66) [0.54 0.00 0.46]

11. 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 (0.69) [0.80 0.05 0.14]; (0.74) [0.58 0.26 0.16]

12. Background - the opposite, information useful just as background, such as an explanation of the causes of a situation (0.47) [0.54 0.12 0.34]; (0.60) [0.77 0.03 0.20]

13. New Information - new information or description of a recent development (0.30) [0.59 0.15 0.26]; (0.92) [0.52 0.01 0.47]

14. Relevant to a Large Group - something affecting many people (0.47) [0.52 0.07 0.42]; (0.83) [0.79 0.01 0.20]

Table 1. Descriptions of each stance aspect, abbreviated from the annotation guidelines [12], and statistics for English and Mandarin, respectively: (interannotator agreements) [fraction of news segments with each label: 0, 1, and 2]

value for that stance for that segment. Examining correlations, we find that these 14 stances are largely non-redundant, with the most related pair, bad and good in English, correlating at only –0.59.

The lists of broadcasts and segments, and the annotations themselves, are available at http://www.cs.utep.edu/nigel/stance/.

3. MODEL

For the automatic inference of stance, many sources of information might be used. In this study we choose to explore only prosodic features. This is for three reasons. First, the need for humanitarian interventions often arises in areas where the language spoken is “low-resource,” in the sense that tools and resources such as speech recognition, dictionaries, and large corpora may not be available. While vocabularies differ arbitrarily across languages, there are universal tendencies for some prosodic features to express certain things across language [14, 15], so a prosody-based approach may be useful for unknown languages. Second, lexical approaches to information retrieval have generally received much more attention, so we were curious about what might be accomplished using prosody alone. Third, previous research suggests that many aspects of stance might be expressed more by prosody than by words.

Our approach is based on the observation that regions that are prosodically similar are often similar also in the stances they express. We focus on regions because each news segment is 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 the more common classification tasks where something — such as an emotion, state or trait — is assumed to be broadly present across the input [16], either because it is a direct indication of a mental or physical state, or because each input is short.

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 [17, 18] Accordingly we use an estimate-locally-then-aggregate method [19]. Thus, for each stance and each segment, every sample in the segment contributes an independent estimate of the strength of that stance in that segment. Samples are offset every 100ms, both in the test data and in the training (reference) data. Depending on the length of the segment, there may thus be tens or thousands of these estimates. The overall estimate is the average of these sample (patch) estimates.

Of the many possible ways to estimate the stance for each patch, we chose a nearest-neighbors algorithm, for three reasons. First, this makes minimal assumptions about the distributions. Second, this can handle cases where the relevant information involves configurations of features, not just distributions of feature values [20, 21, 22]. Third, as an initial investigation, we wanted an interpretable method, so that we could examine its successes and failures to learn about the nature of the problem. We implement 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 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 inversely proportional to the squared distance. Weights are normalized, so that the estimates are not affected by the local density or sparsity of neighbors.
Our hypothesis is that prosody bears information useful for detecting what stance aspects are present in a news story.

As a model should ideally assign to each segment the same label as the average of the human annotators, our primary metric is the mean squared error. In this exploratory study, we are interested in determining whether prosody provides any information at all; thus for each stance aspect our baseline is the performance of a knowledge-free method: predict-the-average. This of course varied for each stance, and for each reference set.

We used leave-one-out testing, that is, cross-validation at the segment level: for each segment and each stance, we predicted the value based on the annotations of that stance in other segments, across each entire dataset. In addition, for English, we did a known-speaker experiment, using the WMMB subset of the corpus. We also did cross-language experiments, where everything was the same, except that the nearest neighbors for reference were sought in the data from the other language.

We tested our hypothesis by computing how close our prediction results were to the true values. We judged the predictor to be outperforming the baseline if its estimates were closer, \( p < 0.05 \) by a one-tailed matched-pairs t-test. Using \( k = 3 \) nearest neighbors, based on preliminary experiments, the results are as seen in Table 2.

5. DISCUSSION

Some stance aspects were predicted fairly well, indicating that prosody does indeed have value for predicting at least some stances. This was true for both languages, for both feature sets, and for both the large-data-multiple-speakers and modest-data-known-speaker conditions.
Overall, performance was better for Mandarin than for English, both absolutely, as seen in the table, and in terms of closer approximating human performance. This may be because the KAZN data had more variety, including more acoustic variation between segments and more segments that were not simply read news but included spontaneous speech and dialog. In the cross-language experiments the performance was very poor: far below baseline for most stance aspects. Clearly the prosodic reflections of stance can vary greatly among languages. Even in the best conditions, for most stance aspects the performance was well below human performance, indicating much potential for improvement.

One obvious factor related to poor performance was cases where the distribution for a stance was very unbalanced; in such cases, marked with x in Table 2, the algorithm’s task was very hard.

Study of the contributions of the various features is a topic for future work. Nevertheless we note that the second feature set did better, although not enormously better. This suggests that, while prosody alone is informative, spectral features provide additional useful information. We investigated further by examining some cases where the model with first set of features worked poorly. These were, of course, to inappropriate similarity estimates. One common cause was missing features, causing, for example, a spurious neighbor relation between a somber region and an upbeat one, and between a deploring region and a positive one. Both of these problems would likely be solved by including features from the second set: spectral features for the first problem, and voicing fraction to capture the consonant lengthening difference in the second one.

Failure analysis also revealed that some patches were less informative than others. 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. In general, there are times where prosody is being used to convey structure, not stance. This fact might be built into a model by using discriminative methods or by somehow using only patches expected to be informative.

### 6. SIGNIFICANCE AND FUTURE WORK

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 and important properties of news stories. It has extended previous work to show how many of these can be inferred [9, 10]. Further, it has shown that at least some of these are somewhat detectable automatically, from prosodic information alone. This serves as a first proof of concept of the idea of using stance for retrieval of audio.

This paper also identified directions for improving performance. Future work should try more features, not only prosodic [27, 28, 29], but also, for some scenarios, also spectral and lexical; try feature weighting and feature selection; and try discriminative, exemplar-based, and other models and machine-learning algorithms [30, 20]. For the latter, we should consider models where, rather than having inverse document frequency.

Also needing further study are the effects of data size and of speaker differences, and the issue of generality across speech genres, such as read news, interviews, speeches, dialog, and video soundtracks. Although the lack of commonality between the prosody-stance mappings of English and Mandarin suggests that expressions of stance are not universal, future work should examine generality within language families.

### Table 2. Performance. MSE: mean squared error. Percent Reduction: the improvement over the baseline, as a percentage of the baseline. Experiments 1 and 3-6 were done with the first feature set; Experiment 2 with the second. and column 3 over column 1. Bolding indicates statistically better than baseline. x indicates low variance (< 0.10), reflecting highly skewed priors.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Reference language</th>
<th>Test language</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test speaker(s)</td>
<td>new</td>
<td>new</td>
<td>new</td>
<td>new</td>
<td>new</td>
<td>known</td>
<td>new</td>
<td>new</td>
</tr>
<tr>
<td>Test language</td>
<td>English</td>
<td>English</td>
<td>English</td>
<td>English</td>
<td>English</td>
<td>Mandarin</td>
<td>English</td>
<td>Mandarin</td>
</tr>
<tr>
<td>Reference minutes</td>
<td>-</td>
<td>710</td>
<td>-</td>
<td>972</td>
<td>-</td>
<td>70</td>
<td>972</td>
<td>267</td>
</tr>
<tr>
<td>Reference segments</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Metric</td>
<td>Baseline MSE</td>
<td>Human MSE</td>
<td>Predictor MSE</td>
<td>Percent reduction</td>
<td>Percent reduction</td>
<td>Percent reduction</td>
<td>Percent reduction</td>
<td>Percent reduction</td>
</tr>
<tr>
<td>1 Bad</td>
<td>0.67</td>
<td>0.11</td>
<td>0.59</td>
<td>12%</td>
<td>18%</td>
<td>9%</td>
<td>26%</td>
<td>7%</td>
</tr>
<tr>
<td>2 Good</td>
<td>0.52</td>
<td>0.28</td>
<td>0.42</td>
<td>20%</td>
<td>20%</td>
<td>16%</td>
<td>12%</td>
<td>-8%</td>
</tr>
<tr>
<td>3 Deplorable</td>
<td>0.39</td>
<td>0.06</td>
<td>0.36</td>
<td>9%</td>
<td>9%</td>
<td>3%</td>
<td>2%</td>
<td>-2%</td>
</tr>
<tr>
<td>4 Praiseworthy</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>x</td>
<td>-20%</td>
<td>x</td>
<td>-100%</td>
<td>x</td>
</tr>
<tr>
<td>5 Controversial</td>
<td>0.06</td>
<td>0.04</td>
<td>0.07</td>
<td>x</td>
<td>-20%</td>
<td>x</td>
<td>-17%</td>
<td>x</td>
</tr>
<tr>
<td>6 Factual...</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>x</td>
<td>-9%</td>
<td>x</td>
<td>63%</td>
<td>x</td>
</tr>
<tr>
<td>7 Subjective...</td>
<td>0.10</td>
<td>0.08</td>
<td>0.17</td>
<td>-81%</td>
<td>0%</td>
<td>x</td>
<td>-1%</td>
<td>42%</td>
</tr>
<tr>
<td>8 Unusual...</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>x</td>
<td>-5%</td>
<td>x</td>
<td>29%</td>
<td>x</td>
</tr>
<tr>
<td>9 Typical...</td>
<td>0.79</td>
<td>0.10</td>
<td>0.53</td>
<td>32%</td>
<td>47%</td>
<td>37%</td>
<td>73%</td>
<td>-19%</td>
</tr>
<tr>
<td>10 Local</td>
<td>0.33</td>
<td>0.26</td>
<td>0.26</td>
<td>20%</td>
<td>42%</td>
<td>x</td>
<td>13%</td>
<td>70%</td>
</tr>
<tr>
<td>11 Immediate...</td>
<td>0.48</td>
<td>0.09</td>
<td>0.39</td>
<td>19%</td>
<td>33%</td>
<td>3%</td>
<td>65%</td>
<td>-37%</td>
</tr>
<tr>
<td>12 Background</td>
<td>0.65</td>
<td>0.27</td>
<td>0.52</td>
<td>21%</td>
<td>32%</td>
<td>18%</td>
<td>51%</td>
<td>-13%</td>
</tr>
<tr>
<td>13 New...</td>
<td>0.40</td>
<td>0.32</td>
<td>0.32</td>
<td>19%</td>
<td>-28%</td>
<td>19%</td>
<td>75%</td>
<td>-232%</td>
</tr>
<tr>
<td>14 Large-Group...</td>
<td>0.58</td>
<td>0.32</td>
<td>0.46</td>
<td>21%</td>
<td>28%</td>
<td>21%</td>
<td>50%</td>
<td>5%</td>
</tr>
</tbody>
</table>
7. REFERENCES


