Ways to Use Prosody in the Detection of Actionable Information in Radio Broadcasts

Nigel G. Ward, James A. Jodoin, Anindita Nath, Olac Fuentes

University of Texas at El Paso
nigelward@acm.org, jajodoin@miners.utep.edu, anath@miners.utep.edu, ofuentes@utep.edu

Abstract

Finding actionable information in unstructured audio collections is a challenge, especially for low resource languages, which may lack not only speech processing tools, but also enough data to train such tools in the normal way. This paper explores alternative ways to detect segments containing actionable information in radio news, using prosody. Examining the prosody of attributes such as urgency, in news broadcasts in six languages, we found consistently correlating prosodic features. Using these for prediction gave a leave-one-language-out AUC of 0.64, and also good results for previously unseen languages.

Index Terms: low-resource languages, filtering, prioritization, urgency, cross-linguistic, universals, stance

1. Motivation

To effectively provide humanitarian assistance after a natural disaster, responders need to understand the situation and know where the needs are greatest [1]. Audio streams, such as radio news, can be valuable sources of information, but generally include also much that is irrelevant. For example, reporting during a forest fire may include not only reports on the extent of the blaze, current containment efforts, and possible further developments, but also discussions of forestry management policies that may have contributed to the fire, eyewitness accounts of how the fire has changed scenic vistas, and interviews with families whose vacation plans have been disrupted. While on-topic, such information is of little interest to those organizing firefighting and relief efforts. Moreover, even during a disaster, radio broadcasts still include entirely unrelated topics, such as sports, entertainment, and politics. Thus there is a need for ways to automatically identify only segments that contain information that mission planners are likely to find useful.

For many major world languages, there exist systems which can help, including speech recognizers and tools for filtering and search. However, disasters may occur anywhere in the world, without warning, so we need means for effectively filtering news audio in any language, including low-resource languages.

This paper explores the use of prosody for this need. We consider prosody for its unique value proposition. First, some aspects of prosody are universal or exhibit universal tendencies, for example the prosody of anger or sadness, so we can hope that aspects of actionability may also have prosodic universals. Second, for any given language, the prosodic inventory is much smaller than the lexical inventory, so we can hope that with just a small amount of time eliciting information from a native-speaker informant we can identify the prosodic properties used in that language to convey actionability and related meanings.

2. Related Research

Previous research has shown that prosody is informative regarding many linguistic and paralinguistic functions [2, 3], and that the prosody of many of these functions has similarities across languages, including the prosody of focus, turn taking, and topic structure. The universality of prosodic expressions has been most studied for emotion, and it has been well demonstrated that some aspects of emotion can be detected at above-chance levels using models trained on data from other languages [4, 5, 6]. However the potential for good cross-language performance on other tasks using prosodic information has not been explored.

Regarding the identification of actionable information specifically, prosody has been observed to have a role, in English, in conveying such stances as whether a statement is presenting good news or bad news [7], whether it is generally viewed positively [8], and whether it is about something in the here-and-now or something about the past or a speculated future [9]. Prosody has been used for detecting action items in meetings [10]. Regarding the prosodic expression of urgency specifically, studies of warnings and alarms, for example in the context of a semi-autonomous car requesting the driver to resume manual control, have found that higher pitch and faster speaking rate correlate with increased perceptions of urgency [11, 12]. Urgency may also correlate with a break in the ongoing flow of discourse, and studies of participants in realtime gameplay and in multi-task situations suggest that high pitch, pitch rise, and breathy voice may signal this meaning [13, 9]. In earlier work we found that news stories that included a call for “immediate action,” such as preparing for an incoming hurricane or ice storm, or an invitation to attend a tonight-only movie showing in the park, could be detected to some extent from specific prosodic configurations spanning a few seconds [14]. However the utility of prosody for urgency detection has not previously been examined.

Overall, such findings suggest that prosody-based methods may have value for this problem. In this paper we investigate the possibilities.

3. Task and Data

Our investigation used data and task descriptions provided by Darpa’s Low Resource Languages for Emergent Incidents (Lorelei) program [1, 15, 16]. This program addresses the challenge of developing language processing solutions rapidly for any low resource language where a need arises, and has fostered numerous advances [17, 18, 19, 20].

In the context of humanitarian assistance and disaster relief, for information to be actionable, it should 1) be relevant to some specific incident (in Lorelei, relating to one or more of 11 specific types, such as evacuation, infrastructure damage, or flooding), 2) be urgent, 3) be current, that is, not something
purely in the past or future, 4) be unresolved, in the sense of not already having been satisfactorily addressed, and 5) include an explicit mention of a location [15]. Accordingly, in the 2018 Lorelei evaluations, systems were expected to be able to process news segments and characterize each by filling in a “situation frame” with accurate values for all five of these properties, among other things. The data we used consisted of segments from news broadcasts. These were divided such that each segment is mostly on one topic and each is no more than 2 minutes in length. The data included between 804 to 1132 segments per language.

The evaluation scenario was designed to simulate what might happen after an actual disaster. At some time an event (an earthquake, for example) occurs somewhere in the world. Soon thereafter, analysts identify the relevant language, pull from the archives a few hours of radio news data for that language, unannotated, and send it to the tech teams. After receipt of this data, at time t, a tech team has 24 hours to develop a system to make the above-mentioned discriminations for the same language. Upon delivery, the system is immediately put into production. For evaluation purposes, it is judged on its ability to accurately process a new, unseen set of data from this language. In this way the tech teams are judged on their ability to rapidly develop a system to process a surprise language. There is later a second system delivery point, 7 days after t, and this is again evaluated on unseen data.

This is a low resource scenario in that development teams may use, in the constrained condition, only the provided data, of which there is only a few hours for the relevant language. Further this data is entirely unannotated. However, to make the scenario realistic, teams were allowed to consult with a native speaker of the language for a short period of time: 1 hour for the 24-hour checkpoint, and 5 hours for the 7-day checkpoint. This was of course too short to develop a large, annotated training set, so a major challenge was how to effectively use this limited time.

To help teams prepare for the evaluations, six language packs were provided, for Bengali, English, Indonesian, Thai, Tagalog and Zulu. While these are not all low-resource languages, for our investigations we limited ourselves to what was provided, on the order of 10 hours per language. Thus performance on these languages could serve as a initial proxy for the utility of our techniques for a surprise language.

Preliminary examination of the training data led us to focus on two attributes: determining whether a news segment was disaster-relevant or not, and determining whether it was urgent or not. These were the most useful because the vast majority of segments that were disaster-relevant were also current, were not resolved, and included a place mention. Since the official Lorelei metrics were dependent on inference also of additional attributes, which in our case were the responsibility of partner teams, we evaluated performance directly, using the area under the ROC curve, that is the true-positive versus false-positive curve, for each attribute. This area under the curve (AUC) is used in all the tables below, with the baseline being of course 0.5.

4. Four Explorations

We tried various approaches to this problem: one that didn’t work, and three that did.

4.1. Exploration 1: Weighted Counts of Exemplar-Matching Patches

Our first attempt directly addressed the need to develop models rapidly, using only a small amount of native-speaker time. The obvious way to build a model — to have the native speaker listen to each segment, annotate it, and then use standard machine learning techniques — was not possible because the time constraints made it impossible to obtain annotations of even a few hundred segments.

Instead we wanted to present to the native informants just a few dozen short audio clips, optimally selected to be maximally informative regarding the meaningful prosodic patterns of the surprise language. To select such points we considered two unsupervised approaches. Both involved covering the unlabeled data with overlapping 6-second patches, and computed 88 diverse time-spread prosodic features for each patch. The first method was to then use k-Means clustering over all patches, and to then solicit judgments of audio clips near the cluster centers. However, based on previous studies [21, 14], we estimated that several hundred clusters would be needed, exceeding the tight time budget.

The second approach was to use Principal Component Analysis to discover the important dimensions of prosodic variation [22, 9]. Based on previous experience, we know that the top dozen or so dimensions are often meaningful, and that examining some extreme values, both positive and negative, often suffices to characterize the meaning of a dimension.

We tried this method for English, Spanish, and Bengali. The results in each case looked promising, in that each dimension-side’s most extreme (exemplar) timepoints were generally also extreme in some semantic sense, and in that several dimensions related to our task. As each dimension is nothing other than a set of loadings over the per-patch prosodic features, we were thus able to identify meaningful prosodic patterns, that is, meaningful temporal configurations of prosodic features. For example, for English, a high pitch syllable or two followed by a short region with lengthened vowels and then a region of low intensity was often used for positive assessment [23].

The next step was to use this information to build a classifier. As this methods ascribes values to timepoints, we needed a way to aggregate to derive story-level features. We tried counts or sums for timepoints which exceeded various thresholds, for each dimension. We evaluated such features first by measuring correlations, which were often modestly good, and then by using them in simple linear regression models, which however gave poor performance. Overall they were not consistently better than chance, even with various various thresholds and weighting schemes, and even with oracle-based selection of dimensions.

4.2. Exploration 2: Patchwise kNN

Our second approach was to apply a model developed earlier for inferring stances from prosody [14]. This classified each patch in a story by k Nearest Neighbors using distance-weighted information from the labels of training-data patches, and then classified each story using the average of the estimates for its patches. This method gave good performance for stances relevant to the current tasks, such as “bad implications,” “unusual or surprising,” and “prompting immediate action,” so we expected it to work well.

We tested this model using 80:20 training/test splits, building a language-specific model for each of the six languages. The results were mostly above chance, but with high variability. Ut-
Emergency was detected best, with an average AUC of .66 across the six languages. While good, this was lower than the performance we had expected. One possible reason is that the data in our earlier studies was exclusively local news, and thus viscerally urgent both for the newsreader and for the audience, and therefore probably expressed with more consistent prosody. In contrast, the Lorelei data included more national broadcasts, with the situation being described often remote both to the newsreader and to the listening audience. Despite the modest performance, we kept this model in our toolbox and later evaluated it on new languages, as discussed below.

4.3. Exploration 3: Multi-Language Models using Linear Regression over Story-Wide Features

The previous two approaches were based on modeling the details of local prosody over regions of a few seconds. There is, however, a poor match between such models and the Lorelei task metrics, which evaluate the ability to infer attributes at the level of news stories, not utterances or utterance sequences. We accordingly developed models which are much less sophisticated but better match the task, using features computed at the level of stories. Our first such exploration was to try "universal" modeling. Not having data for all languages, in fact this was a family of 6 models, trained and tested in leave-one-out procedure.

We investigated 29 features. There were 3 metadata features: broadcast ID, segment ID (position of the story within the broadcast) and segment duration. For both of these we took the log; this helped slightly. The other features were prosodic: there were 13 base features — an intensity feature, 4 pitch features, 2 rate features, 2 articulatory-precision features, a pitch-energy-peak alignment feature [24], speaking fraction, voicing fraction, creakiness, and the voiced-unvoiced intensity ratio — for each of which we used the mean and the standard deviation. All features were track-normalized to reduce the effects of speaker variation [25].

We first examined correlations of these features with urgency. A little preliminary experimentation on the English data showed that the standard deviations were most informative when the features were computed over 50ms windows, except for voicing fraction, for which a 500ms window was best. Since we are interested in possible universals, we looked for features that correlated consistently, either positively or negatively, with urgency across all 6 languages. We found a few: urgency correlates negatively with lengthening and with the segment ID, and it correlates positively with story duration and with the standard deviations of pitch height, pitch narrowness, enunciation, speaking fraction. These are easy to understand: intuitively stories involving urgent situations may come early in the broadcast and may last longer, people conveying urgent information may speak faster, and urgent news stories often involve multiple speakers with various mental states, and thus exhibit variety in certain aspects of speaking style.

To determine which features were most powerful for urgency classification, we trained random forests and observed the importance of each feature. We used the remaining 17 features for the experiments below.

The results are seen in Table 1. For all 6 languages, performance is generally above chance. Prosody seems less informative for the resolved/unresolved distinction, and less useful for Zulu. The former was likely due to the data imbalance: vanishingly few segments were tagged as resolved, both in the training and test data. Regarding the latter, the Zulu newsreaders in this dataset seemed unusually consistent in tone, with less of the expressive variation apparent in the other datasets.

We then explored ways to do better, refining the approach in two ways: first, by trying more sophisticated classifiers, and second in addressing the unbalanced training data problem. Since each language has many more instances of the negative that the positive class (averaging only 14%), we added synthetic positive-class data. Each synthetic example was created by randomly choosing two examples from the minority class and obtaining their weighted average, with random weights in the [0,1] range. The results of testing with one language when training with the original/augmented datasets from all the others are shown in Table 4.4.

We observe that the synthetic data consistently improves performance, although to varying degrees, and that support vector regression trained on the augmented data does best overall. We also observe that simple linear regression does quite well, even without augmenting the data. This is surprising, but may be due to the small size of the dataset or the robustness and appropriateness of these features.

We also explored whether there might be a golden language. In a previous study of disaster-related discriminations across 11 languages, we found a language that was especially useful for training, perhaps because the annotator was careful and consistent, or because of the audio quality. However for this task no language was especially useful for other-language predictions.
4.4. Exploration 4: Language-Specific Models using Linear Regression over Story-Wide Features

We next trained language-specific models, using the same features. Five-fold cross validation gave the results in Table 3. Comparing with the results in Table we observe that the language-specific models perform better for most attributes, but that the cross-language model is better for urgency.

Seeking to understand the limits of this model, we examined some of the stories for which the performance was worst. We noted sensitivity differences among speakers, as in one report from a quiet suburban town, where a morning’s late garbage collection was described using prosody that other newsreaders would likely reserve for truly urgent problems. We also noted style differences, as in one report where a police dispatcher, apparently lacking public speaking experience, read a disaster-related announcement in a flat voice, without much prosodic marking at all.

<table>
<thead>
<tr>
<th>Language</th>
<th>rel.</th>
<th>urg.</th>
<th>cur.</th>
<th>unr.</th>
<th>loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bengali</td>
<td>.57</td>
<td>.61</td>
<td>.57</td>
<td>.09</td>
<td>.56</td>
</tr>
<tr>
<td>English</td>
<td>.53</td>
<td>.51</td>
<td>.55</td>
<td>.56</td>
<td>.55</td>
</tr>
<tr>
<td>Indonesian</td>
<td>.70</td>
<td>.43</td>
<td>.62</td>
<td>.40</td>
<td>.64</td>
</tr>
<tr>
<td>Tagalog</td>
<td>.65</td>
<td>.60</td>
<td>.67</td>
<td>.43</td>
<td>.63</td>
</tr>
<tr>
<td>Thai</td>
<td>.69</td>
<td>.62</td>
<td>.67</td>
<td>.42</td>
<td>.72</td>
</tr>
<tr>
<td>Zulu</td>
<td>.59</td>
<td>.64</td>
<td>.59</td>
<td>.58</td>
<td>.51</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>.62</td>
<td>.57</td>
<td>.61</td>
<td>.41</td>
<td>.60</td>
</tr>
</tbody>
</table>

Table 3: Language-Specific Linear Regression Model Results, AUC

5. New-Language Results

The 2018 Lorelei program evaluation provided data for two surprise languages, Sinhala and Kinyarwanda. Following the scenario, within 24 hours we submitted results from our multi-language model, trained on the 6 development-set languages.

That day we also spent one hour with a native speaker of Kinyarwanda and one with a native speaker of Sinhala to obtain judgments on some training-data segments, and used these to build language-specific models. We continued to obtain judgments, and after a total of 5 hours working with native informants for each language, obtained annotations for a total of 282 segments of Kinyarwanda and 270 of Sinhala. Later in the week, we submitted results for language-specific models trained on all this data. In addition we submitted results using our kNN model, trained on the same data. Since the official Lorelei metrics did not support component-level evaluation, as noted above, and since the testset labels have not been released, we evaluated performance doing leave-one-broadcast-out experiments, using the annotations we had obtained.

As seen in Table 4, for urgency the performance of all models exceeded baseline. Linear regression generally outperformed k Nearest Neighbors. The cross-language model again outperformed the language-specific models, which is not surprising given the low-resource situation. Seeking to better understand the performance, we examined some of the stories for which the linear regression models’ predictions were worst. A general problem was that some stories were ranked by our models as highly urgent, although the native informant had not annotated them as urgent. Listening, we found that many of these included music. As our features are designed to work for speech input only, and are not robust to music, these models should only be applied after diarization has been done to detect and discard music. No other consistent patterns of failure were found.

6. Discussion and Research Directions

We conclude that prosodic information is indeed informative for detecting urgency and other attributes of news broadcasts, even without language-specific training data. While the performance is not outstanding, there is currently no way to produce systems capable of helping prioritize segments for low-resource languages so quickly. We expect these techniques also to be useful for applications beyond disaster response, in other situations needing filtering, characterization or prioritizing of audio segments. While in many workflows prosodic information alone may not be adequate for practical purposes, there is good potential for use in combination with other sources of information, such as lexical, since prosody is a partly independent source of information in many situations. Future work should examine per-speaker variation, enabling normalization with respect to typical sensitivity and style, and thus likely much better performance.

In terms of furthering our knowledge of how much of prosody is universal, we have found that there are prosodic tendencies in the expression of urgency and other attributes that are consistent across 6 languages. Future work should further examine the generality of this finding.

Future work should also investigate the ways in which the languages of the world assign prosodic feature combinations to various functions: emotional, turn-taking, focus, stance-related, and so on. If we can identify both the universal tendencies and the possible range of variations, we may in future be able to efficiently pin down all the prosody-meaning mappings for any new given language, given only tiny amounts of data and a small amount of native-informant time.

7. Acknowledgements

We thank Alonso Granados for help evaluating the kNN model, Chunxi Liu for discussion, and DARPA for support under the LORELEI program, although no official endorsement can be inferred.

8. References


