Using Prosody to Spot Location Mentions

Gerardo Cervantes, Nigel G. Ward

University of Texas at El Paso
gervantes8@miners.utep.edu, nigelward@acm.org

Abstract
Identifying location mentions in speech is important for many information retrieval and information extraction tasks. Most commonly this is done with speech recognition and a gazetteer, but here we explore the value of prosody for this task. While previous work has explored the use of prosody for spotting named entities, including locations, the specific value of prosody for finding locations in spontaneous speech is not known. Using the Switchboard corpus and LSTM modeling we obtain modest performance. Further, we identify specific prosodic features and configurations that tend to mark locations in American English.

Index Terms: named entities, LOC, information retrieval, spontaneous speech, prosodic patterns, LSTM

1. Introduction
Location spotting is important for many tasks, including information retrieval, information extraction, question answering, summarization and translation. For speech, location spotting is most commonly done with speech recognition and a gazetteer, but this strategy is not always applicable and effective. First, for many low-resource languages there are no good recognizers, or no recognizers at all [1]. Second, even when good recognizers exist, many locations will be out-of-vocabulary. We may nevertheless need to identify likely location mentions, either to send them to a human for transcription and lookup, or for special processing. For example, since location names tend to be pronounced similarly across languages — for example Graz in English and Guraatsu in Japanese — cross-language ASR, using acoustic models trained on other languages [2], and gazetteers in other languages may be effective.

For these reasons we are interested in ways to find locations without use of speech recognition. We hypothesize that prosody can be useful for this. Casual observation suggests that, across languages, introductory mentions of new entities, including locations, may share common prosodic features, such as late pitch peak. To the extent that locations are mentioned in certain specific contexts and associated with certain specific pragmatic functions, for example, introducing new topics or grounding, it makes sense that certain specific prosodic patterns may co-occur. Thus we thought that it may be possible to identify such general patterns, and then leverage information across languages, that is, to train models on languages with plentiful data and use them to help spot locations in low resource languages.

2. Related Work
In general, words of different classes and with different functions may have different typical prosodic forms. This has been exploited, for example, in spotting important words to include in summaries [3]. This general property of lexical prosody has also been exploited in language models which incorporate prosodic information [4, 5, 6]. More specifically, the value of prosodic information for named entity recognition has been shown, especially for person names.

Hakkani-Tur and colleagues did the first study of this [7], motivated by the idea that name mentions would generally have "prominent" prosody. For broadcast news, comparing with an entity tagger that used lexical information alone, they reported only a modest performance benefit, and found that the benefit came largely from distinguishing content words from function words, rather than from distinguishing entity mentions from other content words.

Work by Katerenchuk and Rosenberg [8], on the Wall Street Journal corpus, showed that acoustic (prosodic) cues can help detect named entities, when used in combination with recognizer output, in cases where the recognition error rate was high.

Rangarajan and Narayanan [9] obtained good results using prosody for detecting person names, although their task was relatively easy because they used read inputs, word boundaries were given, all input sentences contained exactly one person mention, and all person names were from a non-English language, but embedded in an English sentence.

Thus previous work has not shown whether prosody is useful for discriminating location mentions rather than just named entities in general, or even whether prosody is serving only to enable a general discrimination between content and function words. Previous work has also been limited to read speech; here we also examine the prosody of locations in spontaneous speech.

3. Task
Classical formulations of the task of named entity recognition assume that transcripts are available and exact word boundaries are important [10], which is not realistic in general. Instead, we formulate the task as one of identifying speech frames that are parts of location mentions. Specifically, we aim to classify each frame (50 ms) in the audio as part of a location mention (1) or not (0). In applications, such labels would probably be smoothed or otherwise post-processed, however this task formulation is adequate for our aim here, namely to evaluate the pure ability of prosody to discriminate location mentions from all other speech regions.

4. Data
We used the Switchboard corpus of American English telephone conversations, as this is large and fully transcribed with exact word boundaries [11, 12]. The locations are however not labeled, so we created a small script to automatically label locations based on the transcripts. This script relies on the fact that location names in English are capitalized. To exclude person and organization names, it retains only capitalized words that are on our list of likely-relevant locations. This list includes the names of all countries, all states in the United States, all
cities with population over one million, and all cities in Texas, the state in which most of the Switchboard speakers were located. This script was of course not exact. For example, the word Dallas in the Dallas Cowboys was tagged as a location mention, although technically this word is here not a location but part of the team name. However, depending on the intended purpose [13], spotting the word Dallas in this context as a location could still be useful, so the script did not attempt to exclude such cases.

To evaluate the quality of this script, we hand-labeled the first 100 location mentions in 16 Switchboard conversations. Of these, 76 were tagged as locations by the script; thus the recall was 76%. In a sample of 92 words tagged as locations by our script, there were 16 false positives (all of one of three kinds: a car model name, a person name, or a football team name), and thus the precision was 83%. Thus our labels were only slightly noisy, so we chose to use them uncorrected, both for purposes of training and evaluation.

5. Prosodic Features

Our hypothesis is that, in addition to lexical and phonetic indications of locations, the prosody is also informative. Departing from previous work, we accordingly use a wide set of prosodic and associated features, including not only track-normalized pitch, intensity, and duration, but also energy flux, and measures of the degree of creaky voice, lengthening, disalignment between intensity and pitch peaks, and the voiced/unvoiced intensity ratio. These were designed to be robust, as is necessary for spontaneous speech in general, and especially for Switchboard, given its varied audio quality [14]. This feature set has been shown in previous work to be informative regarding many semantic and pragmatic functions [15, 16, 17, 14].

Thinking that indications of location mention may be found not only on the word itself or its immediate neighbors, we used prosodic features spanning a wide context, extending 3200 milliseconds before and after the frame to be classified. Thinking that the behavior of the interlocutor may also be informative, we used prosodic features for both speakers. In another departure from previous work, we computed features over fixed-length windows, without concern for alignment to word, utterance, or syllable boundaries, as we cannot in general assume that these are available.

Further, for the LSTM models, described below, we used a reduced feature set, since LSTMs are in general able to learn temporal patterns, such as the dynamics of and relations among pitch and intensity. For the LSTM we accordingly used only 5 features per speaker, each computed over only one frame, namely absolute pitch, z-normalized pitch, voicing, energy, and cepstral flux (as an indicator for both rate and phonetic reduction). Each frame in the audio thus had 10 (5 + 5) prosodic features.

The code for computing these prosodic features is available open-source in the Mid-Level Prosodic Feature Toolkit [18].

6. Models and Training

We experimented with two models: linear regression, because it is easy to analyze what it learns, and a Long Short Term Memory (LSTM) model, because it can learn temporal patterns and has demonstrated good performance in numerous speech processing tasks.

For both models, 15% of the data was used for testing, 15% for dev, and the rest was used as training data. Since the predictions given by the models are continuous-valued, they were converted to binary by using a threshold. The threshold was set to the value giving the highest performance on the dev dataset by the F1-measure. This threshold was then used for the test set for evaluation.

6.1. Linear Regression Model

Location mentions are not that common: only 1 in 256 frames in this data. To enable learning, we accordingly downsam-pled to have equal numbers of positive and negative examples. Specifically, all frames that had a location mention are used, and the negative frames were selected randomly from places where there is speech but with no location mention.

Linear regression is trained with the computed prosodic features and the binary labels as targets. The predictions are converted to binary by thresholding.

6.2. LSTM Model

Because LSTM models require sequence data, we prepared the training data differently. Still wanting to reduce the preponderance of negative frames, we selected for training only sequences with at least one location mention. To minimize the imbalance, these should be short, but to give the LSTM adequate context, they should be long. We chose as a compromise a sequence length of 10 seconds. These training sequences were selected to be non-overlapping. Sequences of 10 seconds without any location mentions were excluded from training. This gave a positive:negative ratio of 1:14, which we felt was acceptable for training.

In training the sequence of prosodic features was fed to the model together with the sequence of labels. The neural network was bidirectional, so the output can depend on right context information. Based on informal experimentation on the training and dev sets, we chose an architecture with 2 hidden layers, each a bidirectional LSTM layer with 25 units. After the LSTM layers, there was a simple dense feedforward layer of 25 units. The input layer was the prosodic features and the output was the location likelihood estimate. Cross-entropy was used as the loss function. A dropout of 0.6 (keep probability) and L2 regularization of 0.001 were used.

7. Evaluation

Models were trained with 1290 conversations, each about 5 to 10 minutes long, in total about 124 hours of data, and tested with about 26 hours of data. Across all the data our script found 9673 location mentions.

7.1. Comparison of Models

Table 1 compares performance of the linear regression and LSTM models. Both were evaluated on evenly balanced data, and non-speech frames were excluded. However the data was not exactly the same. For the linear regression model, we down-

\[
\begin{array}{|c|c|c|}
\hline
& \text{Linear Regression} & \text{LSTM} \\
\hline
\text{Threshold} & 0.332 & 0.034 \\
\text{Precision} & 0.535 & 0.541 \\
\text{Recall} & 0.920 & 0.922 \\
\text{F1-measure} & 0.676 & 0.681 \\
\hline
\end{array}
\]

Table 1: Model comparison on balanced datasets
sampled the negative frames, as described above. The LSTM Model had to be tested on 10-second segments, and it made a prediction for each frame in each segment, but before computing precision and recall we downsampled the negative-class frames so that the data was balanced in this case also.

We see that both models have higher precision than baseline (0.50) and that the LSTM performs slightly better than the linear regression model.

### 7.2. Comparison to Baselines

To better understand the level of performance for the LSTM, Table 2 shows results for three baselines: a) choosing location frames randomly, b) choosing location frames randomly but excluding non-speech frames, and c) choosing location frames randomly but excluding frames in function words or non-speech regions. We defined function words to be those on the NLTK stoplist. In each baseline we randomly predicted that 50% of the qualifying frames are in locations. All the results in Table 2 are for data selected so that there is at least 1 location in each 10 second sequence.

Further, to evaluate whether the interlocutor-track features were informative, we built another LSTM model using only one track, excluding features computed from the audio track of the other speaker. As seen in the rightmost column of Table 2, the performance was slightly lower, indicating that it is indeed helpful to use both tracks of a dialog, when available. This may be in part because this enables the LSTM to learn to correct for the cross-track bleeding present in some conversations, and likely also because the interlocutor’s listening behavior and responses are actually informative, as illustrated below.

### 7.3. Locations and Other Entities

We next sought to determine whether our model had indeed learned to spot location frames, rather than entity frames in general. We hypothesized that the prediction values for frames that were locations would be generally higher than the prediction values for other named entities. We wrote a script to detect all capitalized words; these were in general names of people and organizations, and we used this set, uncorrected, as our list of entities. (In the annotations capitalization was only for proper names and titles; sentence-initial words were not capitalized.) We then compared the prediction values at the location frames to those at the non-location entity frames. The means were 0.124 and 0.107, respectively, which were significantly different by a t-test (p < 0.0001).

### 7.4. Generality Across Languages and Genres

As a preliminary investigation of whether the model was specific to this language and this data set, we did some small-scale experimentation with other data sets.

Since for none of these did we have timestamped transcripts, our evaluation was done in a post hoc fashion, based on examination of timepoints for which our model had high location estimates. We started from the highest likelihood frame and worked down the list. However, as high-estimate frames tended to be clustered in time, to get a more diverse sampling we skipped over frames within one second of those already examined. For each language, we examined the top 100 timepoints the LSTM model predicted in this way and computed the precision. For comparison, we annotated randomly selected points in the audio until we found 100 random timepoints that had non-function words. For the speech-only baseline, laughter, music and silence timepoints were excluded. In each case, the precision was computed by taking the number of locations found over the number of timepoints examined. To enable comparison, we also examined 100 predictions for Switchboard in the same way.

The first comparison dataset was an English news broadcast dataset: 6 hours of local news broadcasts data from different stations [17]. As these had only a single audio track, we used the single track model as seen in Table 2. As seen in columns 1 and 2 of Table 3, there appear to be many more locations in this data, and the model appears useful for identifying them.

The other datasets were the Japanese and Spanish Callhome telephone conversation corpora, approximately 49 and 10 hours of data, respectively. As seen in Table 3, there is no evidence that the model is providing any value. This suggests that the prosody of locations is different across languages.

### 8. Failure Analysis

Seeking to learn more about how our model works and when it fails, we looked at its performance in specific cases.

First we examined false alarms, initially for just 20 timepoints in the data subset described in Section 6.2 to which our model ascribed some of the highest likelihoods of being locations, selecting among those as described above. Of the 20 timepoints, 6 of these were not technically within a location mention, but were very close, for example, within the underlined words in uh Richardson and Dallas is. 2 were location names that our script had failed to tag as such, namely Eastern Europe and VPI, a college campus mentioned as a geographical reference point. 4 were within mentions of people, curiously all intimately related to the city under discussion, for example, the mayor of Houston and Michael Jordan for Chicago. 3 were within location mentions that did not include specific mention names, as in from up North, in [Colorado] . . . that’s a great state and in what type of area is that. In summary, most of these false alarms are associated with locations in some way and thus, depending on the application, could still be of practical value. A further examination of high-estimate false alarms across all the test data suggests that indistinguishable prosody is also common with three dialog acts: grounding, agreeing, and comparing.

Second, we examined 20 misses (false negatives): timepoints where our script found a location, but to which the model

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Random</th>
<th>Speaking</th>
<th>Content-Word</th>
<th>LSTM</th>
<th>1-Track LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>6.7%</td>
<td>10.3%</td>
<td>15.4%</td>
<td>18.3%</td>
<td>17.3%</td>
</tr>
<tr>
<td>Recall</td>
<td>49.9%</td>
<td>49.3%</td>
<td>49.8%</td>
<td>40.7%</td>
<td>39.5%</td>
</tr>
<tr>
<td>F1-measure</td>
<td>11.9%</td>
<td>17.1%</td>
<td>23.5%</td>
<td>25.3%</td>
<td>24.3%</td>
</tr>
</tbody>
</table>

Table 2: LSTM models compared with various baselines
Ascribed very low likelihood of being a location. These include 1 person name that was the same as a city name, 3 uses of a city name in metonymy for a sports team, for example, a mention of Philadelphia going into the playoffs, and 3 uses of a state name not as a location but as a geo-political entity, as in Massachusetts reinstated the death penalty. Depending on the application, not tagging these as locations could be the proper choice.

Thus the quantitative performance values reported above (Tables 1 and 2) likely underestimate the true utility of prosody for location spotting.

Third, we examined 20 of the strongest hits: timepoints to which our model ascribed very high likelihood of being a location, which were in fact locations. These were predominantly truly grounded location mentions, where the speakers were stating or confirming where they had lived or were currently living, rather than, for example, discussing of cities or states in the news. Thus it seems that our model had learned a preference for certain kinds of location mentions; again, depending on the application, this could be a good thing. We also noted that the model successfully identified locations across diverse speech acts, including statements (e.g. I went to UT at Austin), confirmatory echoes (e.g. oh, Plane and questions (e.g. how long have you been in Dallas?). This suggests that the model was able to learn the commonalities of location prosody, regardless of other superimposed prosodic patterns conveying other pragmatic functions.

9. Feature Analysis

To get a rough idea of how prosody was enabling detection of locations, we inspected the coefficients of correlation with the presence or absence of a location frame (1/0). The first finding was that the correlations were time-dependent. For example, intensity correlated positively with upcoming frames being locations, but negatively with recent past frames being locations. Table 4 shows all features whose correlation’s absolute value was greater than 0.02, ordered by time: the times are the window (ms) correlation

<table>
<thead>
<tr>
<th>feature</th>
<th>window (ms)</th>
<th>correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>speaker pitch narrowness</td>
<td>–800 –400</td>
<td>+0.026</td>
</tr>
<tr>
<td>speaker intensity</td>
<td>–600 –400</td>
<td>+0.022</td>
</tr>
<tr>
<td>speaker speaking rate</td>
<td>–400 –200</td>
<td>+0.027</td>
</tr>
<tr>
<td>speaker lengthening</td>
<td>–400 –200</td>
<td>–0.022</td>
</tr>
<tr>
<td>speaker lengthening</td>
<td>–200 –000</td>
<td>–0.25</td>
</tr>
<tr>
<td>speaker pitch narrowness</td>
<td>0 +200</td>
<td>–0.21</td>
</tr>
<tr>
<td>speaker intensity</td>
<td>+600 +800</td>
<td>–0.25</td>
</tr>
<tr>
<td>interlocutor pitch wideness</td>
<td>+800 +1600</td>
<td>+0.20</td>
</tr>
</tbody>
</table>

Table 4: Prosodic features best correlating with location mentions.

Specific, beyond any generic benefit of being able to distinguish content words from function words, and even beyond any generic ability to spot entity mentions. We also present an LSTM model capable of performing better than a linear regression model.

Based on very small samples, the performance of an English-trained model appears low on other languages, but within English appears to generalize to the news genre.

Future work might explore the possible value of partly shared network training and the presence of possible universals. Future work should also quantify the extent to which the information provided by prosody is a useful (non-redundant) complement to that provided by speech recognizers, for languages for which those are available.

11. Acknowledgements

We thank Olac Fuentes for discussion and Isabel Ward for helping with the post hoc evaluation. This work was supported in part by DARPA’s LORELEI program, but no official endorsement should be inferred.

12. References


