10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Using Prosody to Spot Location Mentions

Gerardo Cervantes, Nigel G. Ward

University of Texas at El Paso

gcervantes8@miners.utep.edu, nigelward@acm.org

Abstract

Identifying location mentions in speech is important for many information retrieval and information extraction tasks; here we explore the use of prosody for location spotting. 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 has not been measured. Using the Switchboard corpus and LSTM modeling we obtain above baseline performance. Further, we identify specific prosodic features that tend to mark locations in American English.

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

1. Introduction

Words of different classes and with different functions may have different typical prosodic forms. Previous work has examined such tendencies, but generally regarding either specific words, or regarding broad syntactic or functional categories, such as content words, fillers, and backchannels. In this work we instead investigate the prosodic aspects of a semantic category: locations. Locations are convenient for a study of semantic-class prosody for three reasons. First, locations are a well-defined semantic category, and thus suitable for a big-data study. Second, location mentions can occur in any language, so it is a suitable topic for cross-language investigation. Third location spotting is of practical importance for many tasks, including information retrieval, information extraction, question answering, summarization and translation. For example, there is a practical motivation in finding mentions of disasters and the locations of those disasters in radio broadcasts [1].

For speech, location spotting by the usual methods, namely with speech recognition and a gazetteer (list of locations) is not always applicable and effective. First, for many low-resource languages there are no good recognizers, or no recognizers at all [2]. Second, even when good speech recognizers exist, many locations will be out-of-vocabulary, making the recognizer unable to find the location. Even without speech recognition, it can be useful 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 Texas in English and Tekisasu in Japanese — cross-language ASR using acoustic models trained on other languages [3], 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 it may be possible to identify such general patterns, and then leverage this information across languages

2. Related Work

There have been numerous computational studies of the prosodic properties of words and word classes. For example, Lai and others have shown the utility of prosody for spotting important words to include in summaries [4, 5]. Word-characteristic prosodic patterns and contextual prosodic tendencies have also been exploited in language models [6, 7, 8]. More specifically relevant to locations are studies of the value of prosodic information for named entity recognition. We briefly overview the three most relevant previous studies.

Hakkani-Tur and colleagues did the first study of this [9], 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 from prosody, and found that the benefit came largely from distinguishing content words from function words, rather than from distinguishing entity mentions from other content words.

Rangarajan and Narayanan [10] obtained good results using prosody for detecting person names, although their task was relatively easy because the inputs were read speech, 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.

Work by Katerenchuk and Rosenberg [11], on the Wall Street Journal corpus, also found 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.

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 doing more than just enabling 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

Our hypothesis is that prosody is informative for spotting location mentions. We formalize the task as one of identifying places in speech where locations are likely being said. Classical formulations of the task of named entity recognition assume that transcripts are available and exact word boundaries are given [12], which is not realistic in general. Instead, we formulate the task as one of identifying speech frames that have location mentions. Specifically, we aim to classify each 50 millisecond frame of audio as including part of a location mention (1) or not

 10.21437/SpeechProsody.2020-187

(0). In real-world 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 [13, 14].

Location mentions are however not labeled in the transcripts. To find locations from the transcripts, we used spaCy[15], a natural language processing library. SpaCy has multiple downloadable neural network models that identify named entity types. We applied spaCy to the transcripts and noted which words it classified as geopolitical entity (GPE) or location (LOC). These locations as output by the spaCy model were not exact. For example, the word *Dallas* in *the Dallas Cowboys* was tagged as a location mention, although this word here is not a location but part of the team name. (However, depending on the intended purpose [16], spotting the word *Dallas* in this context as a location could still be useful.)

To judge whether the spaCy-generated tags would be adequate to support our experiments, we did a small evaluation, in two parts. First, we hand-labeled the first 100 location mentions in 16 Switchboard conversations. Of these, 86 were tagged as locations by spaCy; thus the recall was 86%. Seconds, in a sample of 98 words tagged as locations by spaCy, we found 12 false positives, and thus the precision was 88%. Thus only slightly noisy, so we chose to use them uncorrected, both for purposes of training and 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 spaCy found 9673 location mentions.

5. Prosodic Features

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 the two models, described below, different featuresets were used.

For the linear regression model, we 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 [17]. Like other feature sets [18], this feature set has been shown in previous work to be informative regarding many semantic and pragmatic functions [17, 19, 20, 21]. 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. 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 will be available.

	Linear Regression	LSTM
Threshold	0.329	0.033
Precision	0.532	0.532
Recall	0.950	0.945
F1-measure	0.682	0.681

Table 1: *Model comparison on balanced datasets*

For the LSTM models 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. LSTMs have been shown to require only a few frame-level prosodic features to achieve good results [22]. For the LSTM, we accordingly used only 5 features per speaker, each computed frame-by-frame, namely absolute pitch, z-normalized pitch, voicing, energy, and cepstral flux (as an indicator for both speaking 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 [23].

6. Training and Testing

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 that gave 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 have locations in this data. To enable learning in linear regression, we accordingly downsampled 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. For evaluation, 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 sequences of prosodic features were fed to the model together with the label sequences, of 0 or 1 for every frame. The neural network was bidirectional, so the output could depend on both the left context (past), and right context (future) information. Based on informal experimentation on the

	Random Baseline	Speaking Baseline	Content-Word Baseline	LSTM Model	Single-Track LSTM Model
Precision	7.2%	10.9%	16.5%	18.9%	20.1%
Recall	43.2%	49.1%	49.9%	43.2%	38.6%
F1-measure	12.5%	17.8%	24.8%	26.3%	26.5%

Table 2: *LSTM models compared with baselines*

training and dev sets, we chose an network architecture with 4 hidden layers of 16, 8, 8, and 4 units respectively, each a bidirectional LSTM layer. After the LSTM layers, there was a simple dense feedfoward layer. The input layer was the prosodic features and the output was the location likelihood estimate. Cross-entropy was used as the loss function. L2 regularization of 0.0001 was used. The code to train and evaluate the model is available on GitHub [1].

7. Results

7.1. Comparison of Models

Table 1 compares the 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: the non-speech frames were different as they were randomly selected with a different random seed, and in different ways, as follows. For the linear regression model, we downsampled the negative frames, as described above. The LSTM model had to be tested on 10-second segments, for which it made a prediction for each frame, 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 linear regression model performs just slightly better than the LSTM model.

7.2. Comparison to Baselines

To understand the level of performance for the LSTM, Table 2 shows results for three baselines: a) Random baseline: we wanted see if the model was doing better than random. b) Speaking baseline: we were interested in finding if the model was doing better than a baseline that has perfect knowledge of whether there is speech or not. This baseline predicts randomly but only when there is speech, as given by the transcripts. c) Content-Word baseline: we were interested in a smarter baseline that has knowledge of whether there is speech, and also whether the word being said was a function word or a content word. Function words are used to express grammatical relationships and can not be locations mentions. We defined function words to be those on the NLTK stoplist. Thus this baseline only predicted randomly when there were content words, according to the transcript, and predicted false otherwise.

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. We expected better performance for the two-track model because it might enable the LSTM to learn to correct for the cross-track bleeding present in some conversations, and because the interlocutor's listening behavior and responses could be informative. However as seen in the rightmost column of Table 2, the performance of this single-track model was slightly higher, thus, contrary to expectation, considering interlocutor-track features gave no added benefit.

7.3. Locations and Other Entities

Previous work had not specifically shown the value of prosody for identifying location frames, rather than identifying frames with entities in general. We therefore decided to test the hypothesis that the prediction values for frames that were locations would tend to be higher than the prediction values for other named entities. We wrote a script to gather all capitalized words; these were in general names of people and organizations, and we used this set, uncorrected, as our list of entities. (This worked because 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 all other (non-location) entity frames. The means were 0.146 and 0.127, 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 we did not have timestamped transcripts for any of these, 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 excluded 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 dividing the number of locations found by 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 [21]. 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 two datasets were Spanish and Japanese Callhome telephone conversation corpora, approximately 10 and 49 hours respectively. As seen in Table 3, the model performed above baseline for Spanish, but below for Japanese. Though very small-scale, the results suggest that prosody of locations in English could have similarities with Spanish.

[1] https://github.com/gcervantes8/location-spotting-using-prosody

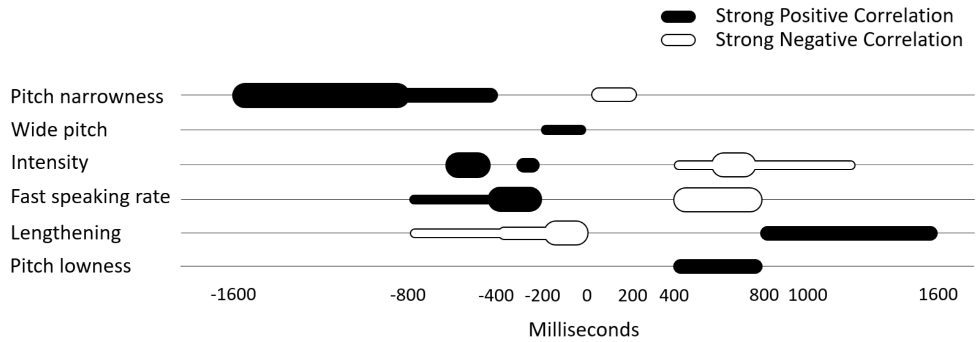

Figure 1: *Prosodic features best correlating with location mentions. Width indicates strength of correlation.*

	News	Conversation		
	English	English	Spanish	Japanese
Speaking Baseline	2.5%	0.8%	0.9%	2.0%
Content-Word B.	3.0%	–	1.0%	4.0%
Model	9.0%	3.0%	3.0%	0.0%

Table 3: *Precision of trained English conversation model evaluated over different datasets*

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. We took 20 timepoints in the data subset described in Section 7.2 from those to which our model ascribed the highest likelihoods of being locations, but which in fact were not. Of these 20, 7 were, although not precisely within a location mention, very close, for example, within the underlined words of: *in Texas*, and *Dallas uh*. 2 of them were mentions of sports teams, *Bears* and *Buccaneers*, which for Americans are often metonymic for cities and regions. 2 of them had a location mention but in the other track, with the name spoken by the interlocutor.

Second, we examined 20 misses (false negatives): timepoints where there was a location, but our model ascribed very low likelihood there. There was no evident pattern in these misses.

Third, we examined 20 of the strongest hits, timepoints to which our model ascribed very high likelihood of being a location, and which were in fact locations. 7 of these occurred in questions, and in five of these the location was the last word in the question, for example *live in Richardson?* and *in California?*. 3 of the 20 were found in answers to questions, for example *uh Clareen County* and *from Indiana*. 7 of the locations were found in truly grounded location mentions, where the speakers were stating or confirming where they were living, rather than, for example, discussing cities or states heard about in the news. Success across these various dialog acts suggests that the model had successfully learned the common properties of location prosody, regardless of 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 correlations 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. Figure 1 shows all features whose correlation's absolute value was greater than 0.02, ordered by time: the times are the window starts and ends relative to the frame being classified. All correlations shown were significant ($p < 10^{-12}$).

All features were of the speaker and not of the interlocutor, no interlocutor features had such high correlations. In location mentions there is usually a wide pitch at the location being said, so we were not surprised that there was a tendency to wider range of pitch at the frame being predicted. Before the frame being predicted we saw that there was narrow pitch around 400 millisecond before the predicted frame. We also saw there was a faster speaking rate before the frame being said. We found higher intensity correlation before the frame being predicted and intensity lowers after the frame.

10. Conclusions and Future Work

We have shown that prosodic information is useful for spotting location mentions, and that this ability is somewhat location-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.

The precision, while not high, is significantly better than baseline, and likely to be useful in larger workflows.

Based on a very small sample, the performance of an English-trained model appears respectable also for Spanish, and 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 recognition, for languages for which that technology is available.

11. Acknowledgements

We thank Olac Fuentes for discussion, and Isabel Ward and Aaron Alarcon 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

[1] N. G. Ward, J. A. Jodoin, A. Nath, and O. Fuentes, "Using prosody to find mentions of urgent problems in radio

broadcasts," in *Speech Prosody*, 2020.

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

[3] M. Wiesner, C. Liu, L. Ondel, C. Harman, V. Manohar, J. Trmal, Z. Huang, S. Khudanpur, and N. Dehak, "Automatic speech recognition and topic identification for almost-zero-resource languages," in *Interspeech*, 2018.

[4] C. Lai and S. Renals, "Incorporating lexical and prosodic information at different levels for meeting summarization," in *Fifteenth Interspeech*, 2014, pp. 1875–1879.

[5] S. Kafle, C. O. Alm, and M. Huenerfauth, "Fusion strategy for prosodic and lexical representations of word importance," in *Interspeech*, 2019, pp. 1313–1317.

[6] N. G. Ward, A. Vega, and T. Baumann, "Prosodic and temporal features for language modeling for dialog," *Speech Communication*, vol. 54, pp. 161–174, 2011.

[7] S. R. Gangireddy, S. Renals, Y. Nankaku, and A. Lee, "Prosodically-enhanced recurrent neural network language models," in *Interspeech*, 2015.

[8] S. Toyama, D. Saito, and N. Minematsu, "Use of global and acoustic features associated with contextual factors to adapt language models for spontaneous speech recognition." in *Interspeech*, 2017, pp. 543–547.

[9] D. Hakkani-Tur, G. Tur, A. Stolcke, and E. E. Shriberg, "Combining words and prosody for information extraction from speech," in *Proc. Eurospeech, vol. 5*, 1999, pp. 1991–1994.

[10] V. Rangarajan and S. Narayanan, "Detection of non-native named entities using prosodic features for improved speech recognition and translation," in *Multilingual Speech and Language Processing*, 2006.

[11] D. Katerenchuk and A. Rosenberg, "Improving named entity recognition with prosodic features," in *Interspeech*, 2014, pp. 293–297.

[12] J. D. Burger, D. Palmer, and L. Hirschman, "Named entity scoring for speech input," in *Proceedings of the 17th International Conference on Computational Linguistics – Volume 1*, 1998, pp. 201–205.

[13] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of ICASSP*, 1992, pp. 517–520.

[14] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone, "Resegmentation of Switchboard," in *ICSLP*, 1998, pp. 1543–1546.

[15] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, 2017.

[16] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, and J. M. Gómez-Berbís, "Named entity recognition: fallacies, challenges and opportunities," *Computer Standards & Interfaces*, vol. 35, no. 5, pp. 482–489, 2013.

[17] N. G. Ward, *Prosodic Pattterns in English Conversation*. Cambridge University Press, 2019.

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

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

[20] N. G. Ward and S. Abu, "Action-coordinating prosody," in *Speech Prosody*, 2016.

[21] N. G. Ward, J. C. Carlson, and O. Fuentes, "Inferring stance in news broadcasts from prosodic feature configurations," *Computer Speech and Language*, vol. 50, pp. 85–104, 2018.

[22] G. Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks," in *Sigdial*, 2017.

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