USING PROSODY TO SPOT LOCATION MENTIONS

GERARDO CERVANTES

Department of Computer Science

APPROVED:

____

Nigel Ward, Chair, Ph.D.

____

David Novick, Ph.D.

____

Olac Fuentes, Ph.D.

____

Stephen Crites, Ph.D.
Dean of the Graduate School

USING PROSODY TO SPOT LOCATION MENTIONS

by

GERARDO CERVANTES

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

May 2020

Acknowledgements

First, I want to thank my advisor Dr. Nigel Ward of the Computer Science Department at The University of Texas at El Paso, for all the guidance and support you have given me. I have learned so much by getting the chance to work with you. My analytical and writing skills have gotten so much better, and I always feel like there is much more for me to learn. Your input is always valuable, and your friendliness always makes it enjoyable to talk. Your support has motivated me so much and your encouragement helps drive me forward. It is difficult for me to say how much you have impacted me. Thank you.

I want to thank Dr. David Novick for all of his guidance and encouragement. I have learned so much about research from your input during the weekly research group meetings. You always ask very thought-inducing questions and always challenge everyone in the research group to do better. Your questions and comments have helped improve all work that I do and will do. I'm really happy you agreed to be a committee member.

I also want to thank the following individuals for their support:

Dr. Olac Fuentes

> I better understood my Computer Science fundamentals through your class. Discussions with you are always very informative, and your knowledge of machine learning has inspired me, and drives me to keep getting better at machine learning. I've grown a good passion for machine learning because of it, thank you.

Dr. Yoonsik Cheon

> I want to thank you for your enthusiasm and joyfulness, you made me much more interested in Computer Science. I learned so much as an undergraduate student from you. Your classes have really helped me improve my knowledge of Computer Science, and I can genuinely say that I write better software because of you.

I want to thank everyone at the research group (of Dr. David Novick and Dr. Nigel Ward), Interactive Systems Group, for being such good friends and great people to work with. We all support each other, and I'm very grateful to have such amazing friends.

This work was supported in part by DARPAs LORELEI program, but no official endorsement should be inferred.

Abstract

Identifying location mentions in speech is important for many information retrieval and information extraction tasks; here I 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 I obtain results indicating that prosody is useful in spotting location mentions. Further, I identify specific prosodic features that tend to mark locations in American English.

Table of Contents

Page

Chapter

Appendix

Chapter 1

Introduction

1.1 The Problem

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 occuring disasters and the location at which the disaster took place in radio broadcasts [22]. Location finding is important to machine translation. For example, in a rule-based approach to machine translation, if you spot a set of words to be a location, then rules specific to locations can be applied. For the purpose of summarization, spotting locations and adding them to the summary is important because locations tend to carry a lot of semantic information.

Besides being of practical importance, locations are convenient for a study of semantic-class prosody for two 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. Languages having shared prosodic patterns in the cross-language investigation I conduct would allow for location spotting in languages with similar prosodic patterns.

1.2 The State of the Art and Its Limitations

For speech, the usual method for location spotting is to use a speech recognizer and a gazetteer. A speech recognizer will turn speech to text, and the gazetteer (a list of locations) will tell you if the word is a location.

The usual method of using a speech recognizer and a gazetteer is not always applicable and effective. First, for many low-resource languages there are no good recognizers, or no recognizers at all [4]. Second, the gazetteer may be missing locations, such as lesser known locations like rivers or mountains not included in the gazetteer. Third, even when good speech recognizers exist, many locations will be out-of-vocabulary, making the speech 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 speech recognizers using acoustic models trained on other languages [25], and gazetteers in other languages may be effective.

1.3 Proposed Technique: Prosody

Prosody includes the intonation, rhythm, and stress of how something was said. I hypothesize that prosody will be a useful alternative to using a speech recognizer because words of different classes and with different functions may have different typical prosodic forms, which could apply to locations. Previous work has examined such tendencies, but generally regarding either specific words, or with respect to broad lexical categories such as content words, fillers, and backchannels. In this work, I instead investigate the prosodic aspects of a specific semantic category: locations.

With a prosody-based algorithm, you may have the advantage of being able to find locations in a way that does not require you to supply a list of locations. Using prosody to get locations will exploit you information on how something was said instead of what was said. Because both speech recognizer and prosody methods can fail at times, it may be important to have both methods to have an increased confidence.

For these reasons I am interested in ways to find locations without use of speech recognition. 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.

Chapter 2

Related Work

In this chapter, I survey related work on spotting named entities. I also define important terminology.

2.1 Semantic Classes

Words can be put into two semantic categories: being a content word or a function word. Content words carry semantic information with them; some examples of content words are nouns, verbs, adjectives, and most adverbs. Functional words have a functional role with little semantic information. They are very common in speech, and examples include articles, prepositions, pronouns, auxiliary verbs, and conjunctions. A named entity (NE) is a subset of content words, and is a real-world object that can be denoted with a proper name. Examples include person names, organizations, and locations. All named entities are content words, while a non-NE could be a content word or a function word. Spotting NEs is vital to speech understanding since they carry a lot of semantic information. There are three research papers most relevant to my work. Similarly, they investigate speech using prosody, but in contrast, they look into spotting NEs.

Distinguishing between content words and function words is an easier task than distinguishing between NE and non-NEs. This is because function words tend to have different prosody, for example, tending to be shorter in length. Since all NEs are content words, if you achieve a better than random chance at classifying content words from function words, then you will get some value for the task of classifying NEs from non-NEs. A more difficult task is to distinguish between NE and non-NE content words, as success in this task shows the ability to do more than just being able to tell content and function words apart.

2.2 Spotting Named Entities using Prosody

The prosodic properties of words and word classes have been studied in many ways. For example, Lai showed the utility of prosody for spotting important words to include in summaries [12]. Word-characteristic prosodic patterns and contextual prosodic tendencies have also been exploited in language models [23, 7, 16]. More specifically relevant to locations are studies of the value of prosodic information for named entity recognition.

Hakkani-Tur and colleagues did the first study of NE recognition using prosody in 1999 [9], motivated by the idea that name mentions would generally have “prominent” prosody. For broadcast news, they reported only a modest performance benefit when adding prosody to their entity tagger that used lexical information. In the study, they achieved a 69% accuracy in distinguishing between NE and non-NE on a balanced dataset using a hidden Markov model with only prosody. Removing function words from non-NE words showed that the high accuracy was likely achieved from being able to distinguish content words from function words. They noticed that the first mentions of named entities had more “prominent” prosody so they tried training a model with only first mentions, but they did not get better results from this model.

Rangarajan and Narayanan [14] obtained good results in 2006 by using prosody to detect non-native person names by using a support vector machine. They were able to distinguish non-native person names from content words that are non-NEs with an accuracy of 76% in a balanced dataset. This was the first study that showed it possible to differentiate content words from a NE type (non-native person names). Despite their good results, their task was made much easier 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] in 2014 showed that acoustic (prosodic) cues can help detect NEs when used in combination with a speech recognizer. They trained their speech recognizer on 78 hours of the English Wall Street Journal corpus. They mention that commercial systems are trained with orders of magnitude more speech data. Using five hours of CNN broadcast news as their test dataset, the speech recognizer got a word error rate of 49% on the dataset, and an F1 measure of 39% for recognizing NEs. After incorporating prosodic features, the F1 measure increased to 45%. They showed that in ASR systems that are deployed rapidly and/or with limited resources, prosody is able to help in detecting NEs.

Thus previous work has not shown whether prosody is useful for more than just enabling a general discrimination between content and function words nor whether prosody is useful for discriminating location mentions from NEs in general.

Chapter 3

Task

This chapter presents my formalization of the task, the corpus chosen for analysis, and the prosodic features selected.

3.1 Problem Description

My hypothesis is that prosody is informative for spotting location mentions. I 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 [1], which is not realistic in general. Instead, I formulate the task as one of identifying speech frames that have location mentions. Specifically, I aim to classify each fifty-millisecond frame of audio as including part of a location mention (one) or not (zero). In real-world applications, such labels would probably be smoothed or otherwise post-processed, however this task formulation is adequate for my aim here, namely, to evaluate the pure ability of prosody to discriminate location mentions from all other speech regions.

3.2 Data

3.2.1 Switchboard corpus

I used the Switchboard corpus of American English telephone conversations, as this is large (around two hundred hours) and fully transcribed with exact word boundaries [8, 5]. The corpus consists of two-sided telephone conversations. For each conversation the speakers are given a topic of discussion, and there were about seventy different topics given.

Models were trained with 1290 conversations, each about five to ten minutes long, in total about 124 hours of data, and tested with about 26 hours of data.

3.2.2 SpaCy tagging

Location mentions are however not labeled in the transcripts. To find locations from the transcripts, I used spaCy[10], a natural language processing library. SpaCy has multiple downloadable neural network models that identify named entity types from text. I applied spaCy to the transcripts and noted which words it classified as geopolitical entity (GPE) or location (LOC). Across all the data spaCy found 9673 location mentions. 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 the word Dallas by itself is a location, with the context of the surrounding words, it is not a location but part of the team name, which is a named entity that is an organization. However, depending on the intended purpose [13], 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 my experiments, I did a small evaluation, in two parts. First, I hand-labeled the first one hundred location mentions in sixteen Switchboard conversations. Of these, eighty-six were tagged as locations, thus the recall was 86%. Of the fourteen locations that were misclassified, spaCy classified five of them as person names and five as organizations. The other four misclassifications were because the locations were mispelled in the transcripts, there were mentions of Ruidoso being mispelled as “Riodosa”, the other mispelling was of LA (Los Angeles) being written as “L A”. Second, in a sample of ninety-eight words tagged as locations by spaCy, I found twelve false positives, so the precision was 88%. Of the twelve false positives, four were mentions of a cat breed, four were organizations, two were sport teams, one was a car name, and the other was a person name. The F1-measure was 87%, thus the labels were only slightly noisy, so I chose to use them uncorrected, both for purposes of training and evaluation.

3.3 Prosodic Features

I 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 [20]. For the two models, described below, different featuresets were used.

The code for computing the prosodic features for both models is available open-source in the Mid-Level Prosodic Feature Toolkit [17].

3.3.1 Linear regression features

For the linear regression model, I 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 [18]. Like other feature sets [6], this feature set has been shown in previous work to be informative regarding many semantic and pragmatic functions [18, 24, 19, 21]. Thinking that indications of location mention may be found not only on the word itself or its immediate neighbors, I 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, I used prosodic features for both speakers. I 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.

3.3.2 LSTM features

For the LSTM models I 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 [15]. For the LSTM, I accordingly used only five 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 ten (five + five) prosodic features.

Chapter 4

Training and Evaluation

In this chapter, I give details on the models I built, and describe the evaluation metrics. For evaluation, I compare the models, further analyze the LSTM model, run a t-test to find if the model is location specific, and I test across languages and genres.

4.1 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 when processing the test set for evaluation.

4.1.1 Linear regression model

Location mentions are not that common: only one in 256 frames have locations in this data. To enable learning in linear regression, I 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 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.

	Linear Regression	LSTM
Threshold	0.329	0.033
Precision	53.2%	53.2%
Recall	95.0%	94.5%
F1-measure	68.2%	68.1%

Table 4.1: Model comparison on balanced datasets

4.1.2 LSTM model

Because LSTM models require sequence data, I prepared the training data differently. Still wanting to reduce the preponderance of negative frames, I 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. I chose as a compromise a sequence length of ten seconds. These training sequences were selected to be non-overlapping. Sequences of ten seconds without any location mentions were excluded from training. This gave a positive:negative ratio of 1:14, which I felt was acceptable for training.

In training, the sequences of prosodic features were fed to the model together with the label sequences, of zero or one 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 training and dev sets, I chose a network architecture with four hidden layers of sixteen, eight, eight, and four units respectively, each a bidirectional LSTM layer. After the LSTM layers, there was a simple dense feedforward 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].

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

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

Table 4.2: LSTM models compared with baselines

4.2 Results

4.2.1 Comparison of models

Table 4.1 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, I downsampled the negative frames, as described above. The LSTM model had to be tested on ten-second segments, for which it made a prediction for each frame, but before computing precision and recall I downsampled the negative-class frames so that the data was balanced in this case also.

Both models have higher precision than baseline (0.50). The linear regression model performs just slightly better than the LSTM model.

4.2.2 Comparison to baselines

To better understand the level of performance for the LSTM, I made three baselines as shown in Table 4.2.

For the random baseline, the baseline predicted randomly at all time-points since I wanted to find out if the model was doing better than a naïve model.

In the speaking baseline, I wanted to find out how a baseline would perform if the model had knowledge of whether the user is speaking or not. This baseline predicts randomly only when the user is speaking, as determined from the transcripts. This baseline perfectly distinguishes speech timepoints from nonspeech timepoints since it uses the transcripts to know when speech is being said.

The content word baseline used transcripts to be able to identify whether a content word or a function word was being said. Similar to the speaking baseline, when they were not speaking, the model predicted there was no location. If the word was a function word, then the baseline predicted those times also as non-locations. A function word is used to express grammatical relationships and cannot be a location mention. I defined them to be the words on the NLTK stoplist. Thus this baseline only predicted randomly when there were content words, according to the transcript, and predicted false otherwise.

For all baselines, the random predictions were done using an actual random number generator predicting either zero or one, then the precision and recall were computed. (However a random number generator was not needed, because expected precision and recall can be computed without it. Since the baseline predicts randomly at all timepoints where there is a location, the recall must be 50%. The expected precision can be computed using the ratio of number of locations to number of timepoints it will predict randomly in. For example, since the data imbalance is one in fourteen, the random baseline will have an expected precision of 7.1%.)

The results shown in Table 4.2 show that the LSTM outperformed all three baselines. The model was able to spot location mentions better than the baseline model which can perfectly separate function word and content words. Thus, prosody is indeed useful in identifying locations in spontaneous speech.

Further, to evaluate whether the interlocutor-track features were informative, I built another LSTM model using only one track, excluding features computed from the audio track of the other speaker. I 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

	LSTM Model	Single-Track LSTM Model
Precision	18.9%	20.1%
Recall	43.2%	38.6%
F1-measure	26.3%	26.5%

Table 4.3: Single track LSTM model compared with a two track model

informative. However as seen in Table 4.3, the performance of this single-track model was slightly higher, this suggests that, contrary to expectation, considering interlocutor-track features has no benefit for performance.

4.2.3 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. I 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. I wrote a script to gather all capitalized words; these were in general names of people and organizations, and I used this set, uncorrected, as the list of entities. This worked because capitalization in the transcriptions was used only for proper names and titles, with sentence-initial words not capitalized. For the LSTM, I then compared the prediction values at the location frames to those at all the other (non-location) entity frames. The means were 0.146 and 0.127, respectively, which were significantly different by a t-test ($p < 0.0001$).

Thus the model outputs higher likelihoods of being a location at frames with a location than at frames with a named entity in general. This shows that locations are prosodically different from other named entity types, and the model is able to utilize some of these location-specific prosodic patterns.

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

Table 4.4: Precision of trained English conversation model evaluated over different datasets

4.2.4 Generality across languages and genres

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

Since I did not have timestamped transcripts for any of these, the evaluation was done in a post hoc fashion, based on examination of timepoints for which the model had high location estimates. I 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, I excluded frames within one second of those already examined. For each data set, I examined the top one hundred timepoints the LSTM model predicted in this way and computed the precision.

For comparison, I annotated randomly selected points in the audio until I found one hundred 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, I also examined one hundred predictions for the Switchboard corpus in the same way.

The first comparison dataset was an English news broadcast dataset: six hours of local news broadcasts data from different stations [21]. As these had only a single audio track, I used the single track model as seen in Table 4.3. As seen in columns 1 and 2 of Table

4.4, there appear to be many more locations in this data, and the model appears useful for identifying them.

The other two datasets were Spanish [3] and Japanese [2] Callhome telephone conversation corpora, approximately ten and fourty-nine hours respectively. As seen in Table 4.4, the model performed above baseline for Spanish, but below for Japanese. Though very small scale, this result suggests that prosody of locations in English could have similarities with those of Spanish.

Chapter 5

Further Analysis

In this chapter I present a further analysis of the results. I analyzed instances where the model performs the best and the worst. I also performed a feature analysis to discover which prosodic features are important in location spotting.

5.1 Failure Analysis

Seeking to learn more about how the model works and when it fails, I looked at its performance in specific cases. First, I examined false alarms. I took twenty timepoints in the data subset described in Section 4.2.2 from among those to which the model ascribed the highest likelihoods of being locations, but which in fact were not. Of these twenty, seven were very close, although not precisely within a location mention, for example, within the underlined words of: *in Texas*, and *Dallas uh*. Two of them were mentions of sports teams, *Bears* and *Buccaneers*, which for Americans are often metonymic for cities and regions. Two of them had a location mention but in the other track, with the name spoken by the interlocutor.

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

Third, I examined twenty of the strongest hits, timepoints to which the model ascribed very high likelihood of being a location, and which were in fact locations. Seven 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?*. Three of the twenty were found in answers

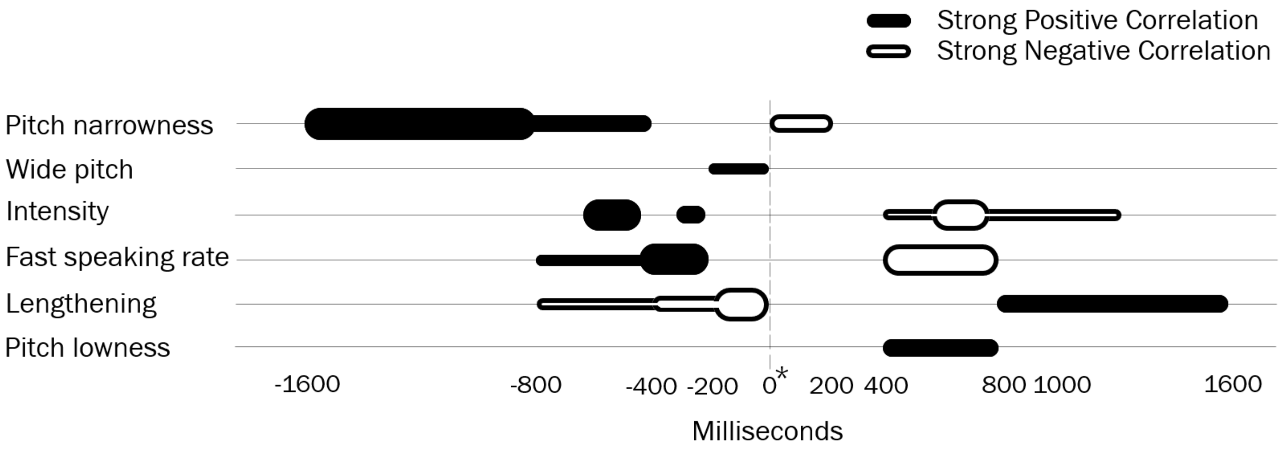

Figure 5.1: Prosodic features best correlating with location mentions. Width indicates strength of correlation. * is the frame being judged as a location or not

to questions, for example *uh* *Clareen* *County* and *from* *Indiana*. Seven 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.

5.2 Feature Analysis

To get a rough idea of how prosody was enabling detection of locations, I inspected the coefficients of correlation of the features with the presence or absence of a location frame (one/zero). 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 5.1 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}). Table A.1 in the appendix has fifteen correlations of the speaker with the smallest p-values, the full list is available on GitHub, the link can be found in section 4.1.2.

All these features were of the speaker and not of the interlocutor, as no interlocutor features had such high correlations. In location mentions there is usually a wide pitch at the location being said, so I was not surprised that there was a tendency to wider range of pitch at the frame being predicted. Before the frame being predicted I saw that there was narrow pitch around 1600 to 400 milliseconds before the location frames. I also saw there was a faster speaking rate before the frame being said. I found higher intensity correlation before the frame being predicted and a lower intensity after the frame.

Chapter 6

Concluding Remarks

6.1 Significance of the Results

I 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.

6.2 Future Work

Future work might explore the possible value of partly shared network training and the presence of possible universals. Future work might also run a larger scale cross-lanuage study with different language location spotting models to find which languages share location mention prosodic patterns. 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.

References

[1] John D. Burger, David Palmer, and Lynette Hirschman. Named entity scoring for speech input. In *Proceedings of the 17th International Conference on Computational Linguistics – Volume 1*, pages 201–205, 1998.

[2] Alexandra Canavan and George Zipperlen. *CALLHOME Japanese Speech*. Linguistic Data Consortium, 1996. LDC Catalog No. LDC96S37, ISBN: 1-58563-077-2.

[3] Alexandra Canavan and George Zipperlen. *CALLHOME Spanish Speech*. Linguistic Data Consortium, 1996. LDC Catalog No. LDC96S35, ISBN: 1-58563-083-7.

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

[5] Neeraj Deshmukh, Aravind Ganapathiraju, Andi Gleeson, Jonathan Hamaker, and Joseph Picone. Resegmentation of Switchboard. In *ICSLP*, pages 1543–1546, 1998.

[6] Florian Eyben, Martin Wöllmer, and Björn Schuller. OpenSmile: the Munich versatile and fast open-source audio feature extractor. In *Proceedings of the International Conference on Multimedia*, pages 1459–1462, 2010.

[7] Siva Reddy Gangireddy, Steve Renals, Yoshihiko Nankaku, and Akinobu Lee. Prosodically-enhanced recurrent neural network language models. In *Interspeech*, 2015.

[8] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520, 1992.

[9] Dilek Hakkani-Tur, Gokhan Tur, Andreas Stolcke, and Elizabeth E. Shriberg. Combining words and prosody for information extraction from speech. In *Proc. Eurospeech, vol. 5*, pages 1991–1994, 1999.

[10] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 2017.

[11] Denys Katerenchuk and Andrew Rosenberg. Improving named entity recognition with prosodic features. In *Interspeech*, pages 293–297, 2014.

[12] Catherine Lai and Steve Renals. Incorporating lexical and prosodic information at different levels for meeting summarization. In *Fifteenth Interspeech*, pages 1875–1879, 2014.

[13] Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489, 2013.

[14] Vivek Rangarajan and Shrikanth Narayanan. Detection of non-native named entities using prosodic features for improved speech recognition and translation. In *Multilingual Speech and Language Processing*, 2006.

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

[16] Shohei Toyama, Daisuke Saito, and Nobuaki Minematsu. Use of global and acoustic features associated with contextual factors to adapt language models for spontaneous speech recognition. In *Interspeech*, pages 543–547, 2017.

[17] Nigel G. Ward. Midlevel prosodic features toolkit. https://github.com/nigelgward/midlevel, 2017.

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

[19] Nigel G. Ward and Saiful Abu. Action-coordinating prosody. In *Speech Prosody*, 2016.

[20] Nigel G. Ward, Diego Aguirre, Gerardo Cervantes, and Olac Fuentes. Turn-taking predictions across languages and genres using an LSTM recurrent neural network. In *IEEE Spoken Language Technology Conference*, 2018.

[21] Nigel G. Ward, Jason C. Carlson, and Olac Fuentes. Inferring stance in news broadcasts from prosodic feature configurations. *Computer Speech and Language*, 50:85–104, 2018.

[22] Nigel G. Ward, James A. Jodoin, Anindita Nath, and Olac Fuentes. Using prosody to find mentions of urgent problems in radio broadcasts. In *Speech Prosody*, 2020.

[23] Nigel G. Ward, Alejandro Vega, and Timo Baumann. Prosodic and temporal features for language modeling for dialog. *Speech Communication*, 54:161–174, 2011.

[24] Nigel G. Ward, Steven D. Werner, Fernando Garcia, and Emilio Sanchis. A prosody-based vector-space model of dialog activity for information retrieval. *Speech Communication*, 68:86–96, 2015.

[25] Matthew Wiesner, Chunxi Liu, Lucas Ondel, Craig Harman, Vimal Manohar, Jan Trmal, Zhongqiang Huang, Sanjeev Khudanpur, and Najim Dehak. Automatic speech recognition and topic identification for almost-zero-resource languages. In *Interspeech*, 2018.

Appendix A

Additional Information

Feature	Correlation	p-value
Lengthening -200 to 0 ms	-0.0262	0
Narrow pitch -1600 to -800 ms	0.0267	1.55E-32
Intensity 600 to 800 ms	-0.0238	4.55E-31
Speaking rate -400 to -200 ms	0.0190	1.48E-29
Intensity -600 to -400 ms	0.0227	1.61E-29
Speaking rate 400 to 800 ms	-0.0184	2.77E-29
Narrow pitch -800 to -400 ms	0.0241	3.67E-29
Narrow pitch 0 to 200 ms	-0.0253	3.05E-28
Lengthening 800 to 1600 ms	0.0193	3.62E-28
Intensity -300 to -200 ms	0.0264	1.92E-27
Lengthening -400 to -200 ms	-0.0187	1.69E-26
Pitch lowness 400 to 800 ms	0.0191	4.85E-22
Intensity 800 to 1200 ms	-0.0189	4.59E-21
Lengthening -800 to -400 ms	-0.0160	1.90E-20
Intensity 400 to 600 ms	-0.0182	5.22E-17

Table A.1: Feature correlations of the speaker sorted by p-value (smallest 15)

Curriculum Vitae

Gerardo Cervantes was born on August 30, 1996. He graduated from Coronado High School, El Paso, Texas in 2014 and entered the Univeristy of Texas at El Paso the following fall. In the fall of 2016, he started working as an undergraduate assistant at UTEP where he worked on mobile applications for classroom environments. Following this, he started research as an undergraduate in natural language processing beginning in the summer of 2017. He graduated from UTEP with a Bachelors of Computer Science in the spring of 2018.

Planning to pursue a Masterś degree in Computer Science, he acquired a summer internship at the Army Research Lab in Maryland in 2018, where he implemented a Neural Machine Translation module for mobile applications. In the fall of 2018, he entered Graduate School at the University of Texas at El Paso and continued his research as a graduate student. In the summer of 2019, he interned at Texas Instruments working as a data scientist on two projects focusing on natural language processing and time series forecasting.

Permanent address: 345 Shadow Mountain Dr. Apt. 502
El Paso, Texas 79912