Detection of Non-Native Named Entities Using Prosodic Features for Improved Speech Recognition and Translation

Vivek Rangarajan, Shrikanth Narayanan

Speech Analysis and Interpretation Lab
Department of Electrical Engineering
University of Southern California Viterbi School of Engineering
http://sail.usc.edu
vrangara,shri@sipi.usc.edu

Abstract

In this work, we describe the use of acoustic-prosodic features to detect and localize non-native named entities spoken by a native speaker in the target language (English) for the purpose of improved speech recognition and translation. The exaggerated variation in accent and duration introduced by the speaker for non-native names is exploited in the detection process through the use of prosodic features like f0 excursions, durational variations and pause information. First, we validate the use of prosodic features in classifying non-native named entities (person names in Chinese, Japanese, Russian, Spanish, Italian, Persian, Indian) in the first mention spoken by native English speakers. We set up the problem as a binary classification task between the non-native named entities and other content words spoken by the speakers in the native language. Results based on a Support Vector Machine (SVM) classifier indicate a 80% classification accuracy for such events. Second, we use the prosody-based SVM classifier to detect and localize named entities at the output of an Automatic Speech Recognizer (ASR).

1. Introduction

The extraction of important entities in speech thus far has been addressed from an information extraction perspective to bridge the gap between automatic speech recognition and speech understanding [1, 2, 3]. The problem can be decomposed into detection, localization and extraction of the entity from speech. The detection and localization of such events in speech has applications besides information extraction. For instance, even rough knowledge of salient information regions in a speech stream opens up possibilities for incorporating alternate decoding and knowledge integration strategies to the speech recognition problem.

Named entities (NEs) are a key part of any language and typically include person names, locations, organization names, monetary amounts, dates and times. They carry salient information and are desired to be recognized with high accuracy in speech streams. The localization of these entities is also beneficial in speech-to-speech translation where the NE can be preserved in translation. Speech summarization [4] is another task where the extraction of NE is vital to the overall performance.

Named entity extraction from speech began as an evaluation metric complementary to WER in typical automatic speech recognizers (ASR) with the NE recognition performance found to degrade linearly with WER [5]. The problem was seen as an information extraction from text task within the natural language community. Hence, previous work on NE extraction relies mostly on lexical information [1, 3, 6]. These systems were grammar-based and relied on attaching names to vocabulary items like punctuation, capitalization and numeric characters. They also required large lexicons to associate words with names. However, the output of a speech recognizer typically lacks these typographic cues.

On the other hand, the speech signal carries rich suprasegmental information, that is beyond words, in the form of energy, intonation and duration, i.e., acoustic-prosodic features. Prosody is used by humans to disambiguate similar words and emphasize the importance of words or phrases. Hence, acoustic correlates of prosody are likely to aid as a cue in several speech related tasks. Prosodic features have been found to be relevant in tasks such as topic segmentation [7], discourse structure and disfluency detection in spontaneous speech [8], voicemail summarization [9] and emotion recognition [10]. The discrimination capability of suprasegmental cues in named-entity recognition from speech can be considered to be supplementary to the information derivable from the linguistic structure.

Given, a vocabulary **V**, the words in it can be divided into function words **F** and content words **C** [11]. Function words include pronouns, articles, prepositions, conjunctions and auxiliary verbs. Linguistically, they are a closed class of words that have a functional role. Content words include nouns, verbs, adjectives, and adverbs. They are an open class of words and convey semantic information. The NE-instances are a subset of the content words. The relation is depicted in Figure 1. It is debatable if all named entities are content words, but in our experiments, we are interested only in person names and from part-of-speech categorization, they are deemed content words.

From a linguistic perspective, it can be expected that stressed syllables in prominent words, and thus also the vowels, are louder, longer and show more pitch variation than non-prominent words [12]. Prosodic features such as f0, intensity and duration have been shown to have an influence on word prominence. Studies have also proven that brand-new entities and new inferred entities in discourse bear phonological prominence [13].

One of the first efforts on NE extraction based on both word content and prosodic features was presented in [2]. In a binary classification task (NE versus non-NE) using prosody alone they found the accuracy was 69%. However, this gain disappeared when the function words were removed from their classifier, suggesting that the gain came from classifying func-

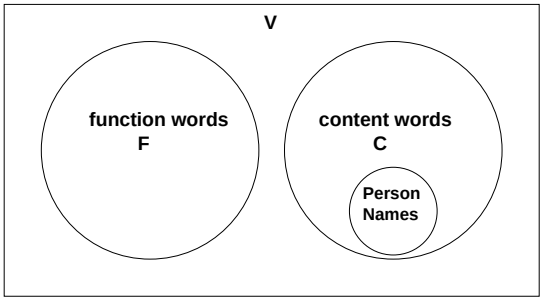

Figure 1: Relation between person names and function, content words based on part-of-speech information

tion and content words, rather than NE-content words and non-NE content words. They used a HMM for NE tagging by incorporating the likelihoods from the prosodic model in the HMM as additional state likelihoods.

In this paper, we are interested in addressing the problem of detecting and localizing non-native named entities. ASR acoustic models are typically trained for a particular demographic or set of speakers and they do not necessarily perform well for new accents and speaker variations. The ability to detect and localize non-native NE in the speech stream using prosodic information alone offers the flexibility to perform additional processing to improve recognition or preserve the location for speech translation. Further, such a procedure does not rely on the hypothesized word sequence of the ASR which may be erroneous due to recognition errors.

We first report results for the NE (person names) classification task based on prosodic and lexical features. The classification task is performed as binary classification between non-native NE and other content words spoken by the speaker using a SVM classifier [14]. This would enable us to understand if non-native named entities are prosodically any different from other content words. Further, we apply the classifier to the output of ASR for the recognition of non-native NE and the word boundary associated with it based on the posterior probabilities of the recognized words and the prosodic classifier.

The paper is organized as follows. In Section 2, we describe the speech corpus used in this work and the prosodic and lexical features chosen for the NE classification task followed by a description of the classifier and its accuracy in Section 3. Section 4 demonstrates the application of the prosody based classifier to the output of an ASR for NE recognition. Finally we provide a summary and directions for future work in Section 5.

2. Experimental Setup

2.1. Data

To test the relevance of prosodic information in NE detection, we used a speech corpus consisting of non-native person names spoken by native English speakers in natural utterances. The database, collected at USC, consists of 70 speakers, with 20 utterances for each speaker. The speakers were prompted with sentences randomly chosen from 100 templates with a variety of syntactic constructs, and populated with names picked from a database of person names in Chinese, Japanese, Russian, Spanish, Italian, Persian and Indian. The corpus was carefully designed to provide good distribution of name positions within the sentence. The utterances consist of only first mentions of the names. The speakers are all native English speakers and hence, the names spoken from other languages can be expected to exhibit pitch accents and durational variations. Figure 2 shows an example utterance.

Where was Kensaku on the night of the fourteenth?

c f NE−c f f c f f c

Figure 2: An example utterance illustrated with function and content word tags

2.2. Annotation of content and function words

The sentences in the corpus were part-of-speech (POS) tagged using a log-linear POS tagger as described in [15]. The POS tags were then used to classify the words as function and content words [11]. A wide range of prosodic and lexical features were extracted for the content words.

2.3. Prosodic and lexical features

2.3.1. Prosodic features

NE words usually carry salient information within a sentence and speakers tend to emphasize them in the first mention. At the word level, prominence is characterized by prosodic features like f0 excursions, increased syllable durations and intensity. This is especially more apparent for native speakers speaking non-native names [16]. Using this as a motivation and also based on descriptive literature [2, 7, 9] we used the following prosodic features in our classifier:

- f0 onset: first non-zero pitch value in the segment
- f0 offset: last non-zero pitch value in the segment
- f0 range: pitch range within the segment
- f0 slope: slope of f0 regression line over segment normalized by f0 slope of sentence
- Energy: mean rms energy of segment normalized by message
- Pause: preceding and succeeding pause information
- Duration: duration of final rhyme in the word (normalized by overall phone duration)

The f0 and energy features were calculated on a segment that included a window before and after the boundary of the content word and the raw values were normalized by speaker specific f0 mean. Other features like logarithm of the ratio of f0 onset and f0 offset, f0 maximum, f0 minimum were also included.

2.3.2. Lexical features

In addition to the prosodic features, we included context information, since content words are usually preceded by function words. The following lexical features were used in the classifier:

- Context: type of preceding and succeeding word (function/content)
- POS : part of speech tag of preceding and succeeding word
- Position: position of word in sentence

3. Classification Task

Let the prosodic features extracted from the ith content word be $\mathbf{f}_i^p$ and the lexical features extracted for the same be $\mathbf{f}_i^l$. The classification problem is a binary one, and involves selecting the class S_i (NAME versus NOT-NAME) for each content word based on the feature set. We trained a SVM classifier as well as a simple decision tree based on the C4.5 algorithm in [17] to predict the class for each of the content words. The reason for selecting a decision tree classifier in addition to a SVM classifier was the easy interpretation of results and the support for missing attributes. Since the algorithm is susceptible to locally optimal convergence, we used a feature selection algorithm [18] to search for an optimal subset of features that are described in the previous section.

The corpus was divided into a training and test set. The training set consisted of about 1000 utterances and the features were extracted for the content words in the training set. We generated forced alignments for the sentences using human transcriptions. The prosodic features were derived from the resulting phone-level alignments and speech signal. The human transcriptions were used only to ensure accuracy in the extraction of prosodic features as ASR systems produce inaccurate time marks due to erroneous recognition. The SVM classifier was trained on 1500 samples and tested with 500 samples (both the training and test set had equal priors).

For the binary classification task of NE versus non-NE, the precision and recall are 76.7% and 86.2% respectively for the SVM classifier. The overall accuracy for the test samples is 80% (significantly higher than chance performance). This suggests that non-native NEs are prosodically different from other content words, at least in the first mention.

		Hypothesis	
		NOT-NAME	NAME
Reference	NOT-NAME	73.8	26.2
	NAME	13.8	86.2

Table 1: Confusion matrix for test data using SVM classifier with prosody only (results in %)

Table 1 shows the confusion matrix for the classification task. In Table 2 we show the prosodic feature usage as the percentage of decisions that have queried the feature. The classification accuracy of the decision tree classifier is 76%. Even though the accuracy is less than that for the SVM classifier, it offers easier interpretation of feature usage. The feature that gets queried the most is f0 range, followed by f0 offset, f0 onset, pause information, energy within the word and rhyme duration (duration of the final vowel in the word or final vowel followed by consonant). The f0 features are indicative of the pitch excursions within words, pause information characterizes a speaker's attention to saying a prominent word, and the energy slope captures the emphasis on the particular word.

The classification was also performed using the lexical features and combined prosodic-lexical features for the same training and test set. Table 3 shows the precision, recall and the overall accuracy of prosodic, lexical and combined features on the test set. The results show that the combined model performs just as well as the lexical model with a marginal improvement.

The high accuracy using the lexical features is due to the limited syntactic variability of the corpus. In general the output of spontaneous speech ASR tends to be noisy (grammatically inaccurate) and relying on just lexical features for recognizing these entities is difficult. The NEs may also be out-of-vocabulary (OOV) words that cause recognition errors. However, since the prosodic features are independent of the hypothesized word sequence, the classifier can be used to detect and localize these entities in speech. In the next section we employ the SVM classifier on ASR output and evaluate the NE detection performance.

Prosodic feature	Percentage of queries
f0 range	25.30
f0 offset	24.52
f0 onset	20.04
pause information	16.79
energy slope	9.81
rhyme duration	3.12
log ratio of f0 onset and f0 offset	0.41

Table 2: Prosodic feature usage in terms of percentage queries (decision tree)

Model	Precision (%)	Recall (%)	Accuracy (%)
Prosody only	76.7	86.2	80.0
Lexical only	87.3	88.0	87.6
Combined	88.6	87.2	88.0

Table 3: Performance of models (SVM classifier)

4. Named Entity Detection from Speech with ASR

The problem of Named Entity recognition in text can be formulated as tagging a sequence of words $W = \{w_1, \cdots, w_k\}$ with the NE tags $NE = \{ne_1, ne_2, \cdots, ne_k\}$ such that $P(W, NE)$ is maximized.

$$NE^* = \arg\max_{NE} P(NE/W) \quad (1)$$

$$= \arg\max_{\{ne_1, ne_2, \cdots, ne_k\}} P(W/NE).P(NE) \quad (2)$$

By using a bigram NE language model and a context dependent channel model (making some conditional independence assumptions), we can decompose the above equation.

$$NE^* \approx \arg\max_{\{ne_1, ne_2, \cdots, ne_k\}} \prod_{i=1}^{k} P(w_i/ne_{i-1}, ne_i, w_{i-1}). P(ne_i/ne_{i-1}) \quad (3)$$

The probabilities are learned from annotated data by using appropriate back-off mechanism. The most probable sequence of named entities is identified by tracing the Viterbi path across the tag-word trellis.

However, the output of the recognizer is the hypothesized word sequence $W^{'} = \{w_1^{'}, \cdots, w_m^{'}\}$ which may have insertion, deletion or substitution errors. The recognizer [19] also outputs a word graph posterior probability $p_{w_i^{'}}$ for each word $w_i^{'}$. Our approach to NE recognition is to select candidate segments s and then apply the prosodic classifier described in section 3 to classify them as NAME or NOT-NAME. Firstly, we tag each hypothesized word $w_i^{'}$ with a tag $t_{w_i^{'}}$ where $t_{w_i^{'}} \in \{\mathbf{f,c}\}$. The tagging is done in a context independent fashion as the hypothesized word sequence maybe grammatically inaccurate. The potential segments are defined as

$$s = \begin{cases} w_i^{'} & \text{if } t_{w_i^{'}} = \mathbf{c}\,; p_{w_i^{'}} < thr; \\ w_i^{'} \cdots w_{i+k}^{'} & \text{if } w_i^{'} ... w_{i+k}^{'} = \{\mathbf{f...f}\}; \\ & p_{w_i^{'}} ... p_{w_{i+k}^{'}} < thr; k \geq 1; \end{cases} \quad (4)$$

For each of the selected segments s, prosodic features are computed and classification is done.

To evaluate the classifier on the output of a speech recognizer, we designed an ASR for the task. The training data from the speech corpus was used to interpolate the language model with one built from the CMU lexicon. We used acoustic models trained on the Wall Street Journal (WSJ) adapted to the training data using maximum likelihood linear regression (MLLR). The test data comprised of 500 utterances not included in the training set and the WER on the test set is 20.1%. The test set consists of 9.07% OOV words and the WER is primarily due to the OOV person names.

The hypothesized word sequence was tagged with NE tags based on just setting the posterior probability threshold thr and tagging all content words and two or more consecutive function words less than the threshold, to be NAME. The resulting segments were then classified using the prosodic classifier. Finally, the reference and hypothesized NE tagged word sequences were aligned[1] and the tags were compared. The method was chosen simply to illustrate the discrimination capability of the prosodic classifier, as it yields a high percentage of false positives which are eventually rejected by the classifier. Applying the prosodic classifier to every segment (thr 1.0) also yields a high percentage of false positives. The proposed scheme of selecting potential segments and applying the prosodic classifier reduces the false positives though localization performance is also slightly affected. The results are summarized in Table 4.

We also evaluated our NE recognition results by using the NIST toolkit for NE-scoring [20]. Table 5 shows the recognition accuracy in terms of *content, extent and type* on applying the prosodic classifier to the selected segments at thr 0.8. *Content* evaluates the performance of classifier on correctly recognized words, *extent* compares the alignment of the reference and hypothesized words and *type* checks for correctness of NE type. The *extent tolerance*, defined as the degree to which the first and/or last word of the hypothesis need not align exactly with the corresponding word of the reference was set to 1.

Model	Accuracy (%)			
	Content	Extent	Type	F-measure
Prosody only	65	76	88	77.81

Table 5: Named Entity tagging performance on ASR output using NIST NE-scoring

It is important to note that ASR WER improvement is not our focus here. We are interested in detecting named entities in speech despite the WER. The *extent* measure which characterizes the alignment of the reference and hypothesized NE tags is more informative. With prosody alone we can localize the NE events in the test set with 76% accuracy at tolerance of 1. At an *extent tolerance* of 2, we found that the *extent* accuracy rose to 81%. In the problem we address, one is more concerned about the approximate boundary of the NE, so the tolerance can be set based on the sentence length. This is a considerable gain as the accuracy of the speech recognizer on NE (person names in our case) as characterized by *content* is not very high. Though we did not try a word-based model (we refer to language model based NE tagging), we believe that prosodic features either in isolation or in conjunction with word-based models could prove beneficial in NE recognition, especially in speech recognizers with high WER. It could serve as a starting point for further processing on the detected regions to improve the performance of the recognizer. The localization of these events is also encouraging from a speech translation perspective, since the named-entity information can be preserved during the translation.

[1] SCLITE (http://www.nist.gov/speech/tools/) alignment tool

5. Conclusion and Future Work

The detection and localization of named entities is important for developing new strategies for decoding speech. While grammar-based approaches have been tried in the past, in this paper, we investigated the problem of NE recognition, particularly for non-native person names, using prosodic features. Experimental results show that for a binary classification task (NE versus non-NE), the accuracy of the prosody classifier is 80%. The classification was done for content words only and the results are indicative of the overlap between named entity content words and words that speakers perceive as prosodically prominent. This is encouraging too, as most speech recognition and translation applications are used in such a context.

NE recognition results on a speech recognizer output show that we can detect the approximate boundary of non-native names using prosody alone with an accuracy of 76% and 81% for *extent tolerance* of 1 and 2, respectively. Our proposed method of selecting segments and applying the prosodic classifier reduces the false positives considerably. Thus, a prosody based NE recognition scheme could serve as a useful tool for analysis of the speech signal. The results also suggest that a first-pass speech analysis can be relevant to building better speech recognizers and aid in speech translation.

Since the prosodic profile of the signal remains the same irrespective of the hypothesized word sequence from the ASR, we believe that such a scheme can aid in OOV detection. Additional information derived from the prosodic classifier can also be incorporated in confidence scoring measures for the ASR output.

The results presented in this paper are preliminary. The experiments were conducted on a limited domain only for non-native person names. However, as a first approximation for the detection of named entities in speech through prosodic cues, the results are encouraging. We need to investigate our method on a larger corpus and also on spontaneous speech. Finally, a unified approach to prosody based NE detection and speech recognition based on the information from the proposed analysis, is a future direction of our work.

6. References

[1] D. Bikel, R. Schwartz, and R. Weischedel, "An algorithm that learns what's in a name," in *Machine Learning*, 1999.

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

[3] Y. Gotoh, S. Renals, and G. Williams, "Named entity tagged language models," in *Proc IEEE ICASSP*, (Phoenix, AZ), pp. 513–516, 1999.

Model	Exact Localization	False Positives	Localization ± 1 word	False Positives
thr 0.7	73.2	41.3	79.7	29.3
thr 0.8	79.5	42.7	84.0	26.4
prosodic classifier + thr 0.7	64.5	23.2	72.1	13.6
prosodic classifier + thr 0.8	72.0	26.8	79.0	12.3
prosodic classifier + thr 1.0	78.3	65.8	86.1	48.7

Table 4: Named Entity tagging performance on ASR output using SCLITE (results in %)

[4] C. Hori and S. Furui, "A new approach to automatic speech summarization," *IEEE Transactions on Multimedia*, vol. 5, no. 3, pp. 368–378, 2003.

[5] F. Kubala, R. Schwartz, R. Stone, and R. Weischedel, "Named entity extraction from speech," in *Proceedings of DARPA Broadcast News Workshop*, (Landsdowne, VA), 1998.

[6] D. E. Appelt and D. Martin, "Named entity extraction from speech: Approach and results using the textpro system," in *Proceedings Of The DARPA Broadcast News Workshop*, pp. 51–54, 1999.

[7] J. Hirschberg and C. Nakatani, "Acoustic indicators of topic segmentation," in *Proc. Inter. Conf. on Spoken Language Proc.*, pp. 976–979, 1998.

[8] E. E. Shriberg, R. A. Bates, and A. Stolcke, "A prosody-only decision-tree model for disfluency detection," in *Proc. Eurospeech 97*, (Rhodes, Greece), 1997.

[9] K. Koumpis and S. Renals, "The role of prosody in a voicemail summarization system," in *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, (Red Bank, NJ, USA), 2001.

[10] C. M. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 293–303, March 2005.

[11] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik, *A comprehensive grammar of the English language*. Longman, 1985.

[12] B. M. Streefkerk, L. C. W. Pols, and L. F. M. ten Bosch, "Up to what level can acoustical and textual features predict prominence," in *Proc. Eurospeech*, pp. 811–814, 2001.

[13] G. Brown, "Prosodic structure and the given/new distinction," in *Prosody: Models and Measurements* (A. Cutler and R. Ladd, eds.), pp. 67–77, Springer-Verlag, 1983.

[14] V. Vapnik, *The Nature of Statistical Learning Theory*, ch. 5. Springer-Verlag, 1995.

[15] K. Toutanova and C. D. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, (Hong Kong).

[16] S. Fitt, "The pronunciation of unfamiliar native and non-native town names," in *Proc. Eurospeech*, (Madrid, Spain), 1995.

[17] R. Quinlan, *Programs for Machine Learning*. Morgan Kaufmann, 1993.

[18] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," in *Speech Communication*, no. 32 in Special Issue on Accessing Information in Spoken Audio, pp. 127–154, 2000.

[19] B. Pellom, "Sonic: The university of colorado continuous speech recognizer," tech. rep., University of Colorado, Boulder, Colorado, 2001.

[20] J. Burger, P. D., and H. L., "Named entity scoring for speech input," in *COLING-98*, (Montreal), 1998.