Cross-Language Prominence Detection

Andrew Rosenberg[1], Erica Cooper[2], Rivka Levitan[2], Julia Hirschberg[2]

[1]Department of Computer Science, Queens College / CUNY, USA
[2]Department of Computer Science, Columbia University, USA

Speech Prosody 2012

Background: Prominence Detection

Pitch accent detection is useful for many speech and language tasks:

- Part-of-speech tagging
- Syntactic disambiguation
- Text-to-speech synthesis
- Reducing language model perplexity for speech recognition
- Salience detection
- Distinguishing between given and new information
- Identifying turn-taking behavior and dialogue acts

Motivation and Experimental Goals

- Prosodically-labeled data is not available for most languages
- Can prominence detection models trained on labeled data from one language can be adapted successfully for other languages?
- How much data is needed for adaptation?
- Does language family predict cross-language prominence detection accuracy?

Languages and Corpora

- English: Boston Directions Corpus. Read (60min) and spontaneous (50min) speech from four non-professional speakers
- French: C-PROM Corpus. 70min of speech from 28 speakers in a mix of seven different tasks.
- German: DIRNDL Corpus. 2.5hrs of radio news from 3 professional speakers.
- Italian: 25min of read speech from one male professional speaker.

Automatic Prominence Detection with AuToBI

- AuToBI: an open-source toolkit for hypothesizing pitch accents and phrase boundaries
- L2-regularized logistic regression classifier for prominence detection
- Developed for Standard American English
- Features: Pitch, Intensity, Spectral Balance, and Pause/Duration

Experiments

- Cross-language prominence detection
- Language-independent prominence detection
- Analysis of language similarities and differences
  - Which acoustic features are most predictive of prominence in each language?
  - What are the distributions of acoustic features in each language?
- Adaptation with augmented data

Cross-Language Prominence Detection

Experiment: Train on one language, evaluate on another.

	Training Corpus – Full			
	BDC	C-PROM	DIRNDL	Italian
BDC	-	71.99(0.34)	76.28(0.44)	62.95(0.12)
C-PROM	80.35(0.28)	-	84.29(0.42)	79.28(0.24)
DIRNDL	76.97(0.53)	82.90(0.65)	-	82.08(0.64)
Italian	80.22(0.56)	77.20(0.49)	80.95(0.57)	-

Accuracy and (in parentheses) relative error reduction using models trained on full corpora in one language and testing on another.

	Training Corpus – 25 mins			
	BDC	C-PROM	DIRNDL	Italian
BDC	-	71.88(0.33)	78.07(0.48)	62.95(0.12)
C-PROM	79.44(0.24)	-	83.50(0.39)	79.28(0.24)
DIRNDL	78.43(0.55)	82.27(0.64)	-	82.08(0.64)
Italian	80.40(0.56)	78.40(0.52)	80.95(0.57)	-

Accuracy and (in parentheses) relative error reduction using models trained on 25 minutes of material from one language and testing on another.

Language-Independent Prominence Detection

Train on three languages, evaluate on the fourth.

Test Corpus	Accuracy	Majority Class Baseline
BDC	74.86%	57.86
C-PROM	81.81%	72.9
DIRNDL	84.24%	55.42
Italian	84.69%	50.85

Prominence detection accuracy training on three languages and testing on the test-split of the fourth.

- Performance on Italian and DIRNDL is improved
- BDC and C-PROM do worse than just training on DIRNDL
- *Model selection* approach would be appropriate

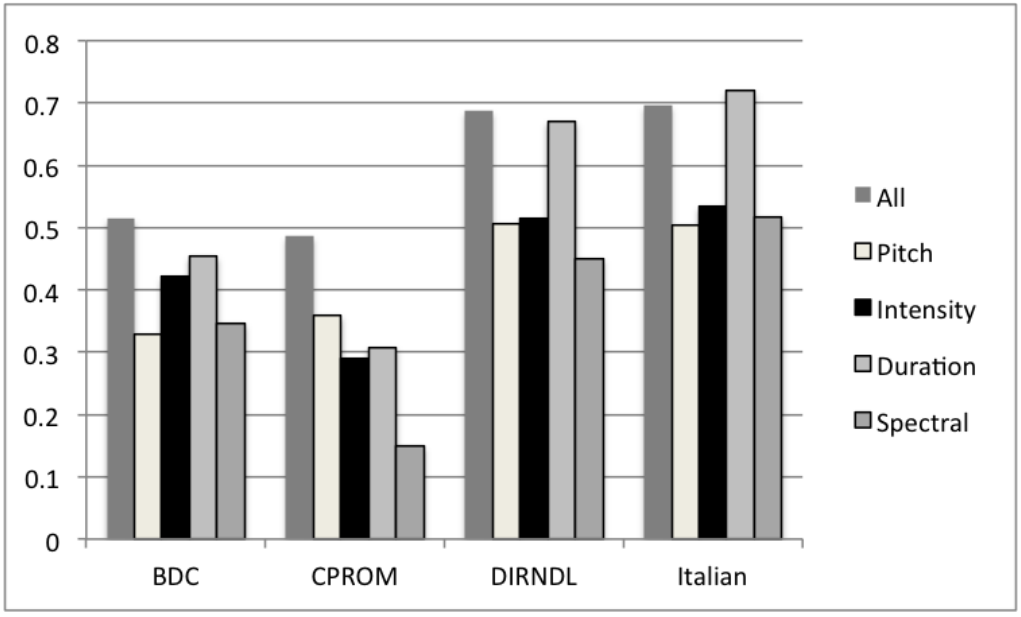
Feature Analysis

- Train prominence detection models with one feature set at a time: Pitch, Intensity, Duration, and Spectral
- Hypothesis: Two languages that demonstrate similar relative prominence prediction performance across the four feature sets may be more compatible for cross-language prediction

Corpus	All	Pitch	Intensity	Duration	Spectral
BDC	79.55	71.68	75.61	77.00	72.45
C-PROM	86.11	82.63	80.73	81.26	76.94
DIRNDL	84.65	75.72	76.13	83.84	73.02
Italian	86.51	77.85	79.22	87.51	78.49

Accuracy using feature subsets.

Feature Analysis



Relative reduction of error using feature subsets.

- German and Italian: Durational features are most predictive
- BDC: Intensity and pause/duration
- C-PROM: Pitch features
- These relationships are not predictive of the cross-language results.

Comparing Feature Distributions

- Mean and standard deviation of four representative features
- Measure KL-divergence between each pair of languages for each feature

Corpus	feature	prom.	non-prom.
BDC	pitch	0.166±0.85	-0.134±1.02
	int.	0.079±0.41	-0.155±0.63
	spec.	0.133±0.46	-0.211±0.41
	dur.	0.328±0.15	0.159±0.10
C-PROM	pitch	0.287±0.64	-0.204±0.80
	int.	0.075±0.41	-0.064±0.65
	spec.	0.098±0.43	-0.066±0.53
	dur.	0.425±0.19	0.186±0.14
DIRNDL	pitch	0.075±0.57	-0.216±0.79
	int.	0.053±0.32	-0.121±0.56
	spec.	0.027±0.31	-0.079±0.42
	dur.	0.529±0.22	0.230±0.13
Italian	pitch	0.032±0.71	-0.025±0.86
	int.	0.017±0.34	-0.038±0.66
	spec.	3.568±0.35	-0.032±0.49
	dur.	0.561±0.23	0.190±0.13

Comparing Feature Distributions

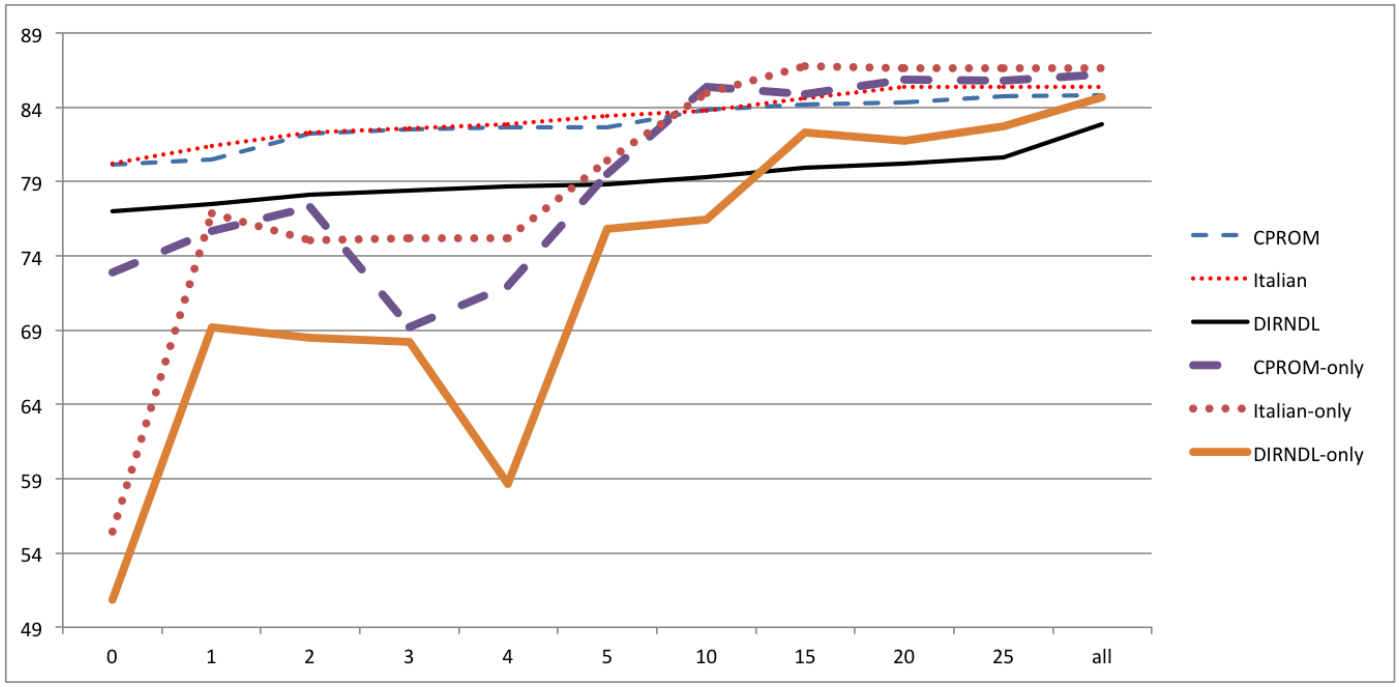
Corpus	BDC	C-PROM	DIRNDL	Italian
BDC	0	0.126	0.493	0.402
C-PROM	-	0	0.375	0.266
DIRNDL	-	-	0	0.056
Italian	-	-	-	0

Total KL divergence between each pair of languages based on supervised GMM based on four features.

- DIRNDL and Italian show similar distributions
- C-PROM and Italian show similarities
- Does not explain good performance of DIRNDL-trained models on BDC.

Adaptation with Augmented Data

- Semi-supervised domain adaptation to leverage large amount of English training data to improve models from other languages
- Base model trained on full BDC corpus
- Augment with increasing amounts of data from target language



Accuracy from models trained on BDC material augmented with variable amounts of target-language training data.

Conclusions

- Language family not predictive of cross-language prominence detection performance
- Nor is relative importance of features used in prominence prediction
- Augmented training data can successfully adapt American English models to other languages
- For some languages, using training data from multiple languages can improve performance over training on only a single language

Future Work

- Further exploration of domain adaptation
- Additional prosodic analysis tasks including prominence type classification, phrasing detection, and phrase-ending classification
- Model selection