Detecting Pitch Accent Using Pitch-corrected Energy-based Predictors

Andrew Rosenberg & Julia Hirschberg
{amaxwell, julia}@cs.columbia.edu

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Pitch Accent

- Pitch Accent is the way a word is made to "stand out" from its surrounding utterance.
- Accenting words is done for many reasons
  - Contrast, Focus, Salience, Information Status
  - Syntactic/Semantic Disambiguation
- Pitch (f0), Duration, and Energy are known correlates of Pitch Accent.
- Human detection agreement between 85-90% [Wightman&Ostendorf94], [Silverman, et al.92]

Previous Work

- Spectral Tilt correlates with pitch accent in Dutch and Swedish. [Sluijter & vanHeuven96,97], [Heldner, et al.99,01] [Fant, et al.00]
- We examined the discriminative strength of the energy components of 210 frequency bands by constructing pitch accent detectors using only energy information on read speech. [Rosenberg & Hirschberg06]
  - There is a relatively small overlap in correct predictions even among similar frequency bands.
  - Best band: 2-20bark (75.5% accuracy)
  - >99% of pitch accents correctly detected by at least one energy-based classifier.
  - These classifiers can be combined (voting) to predict pitch accent with high accuracy (81.8%)

Question

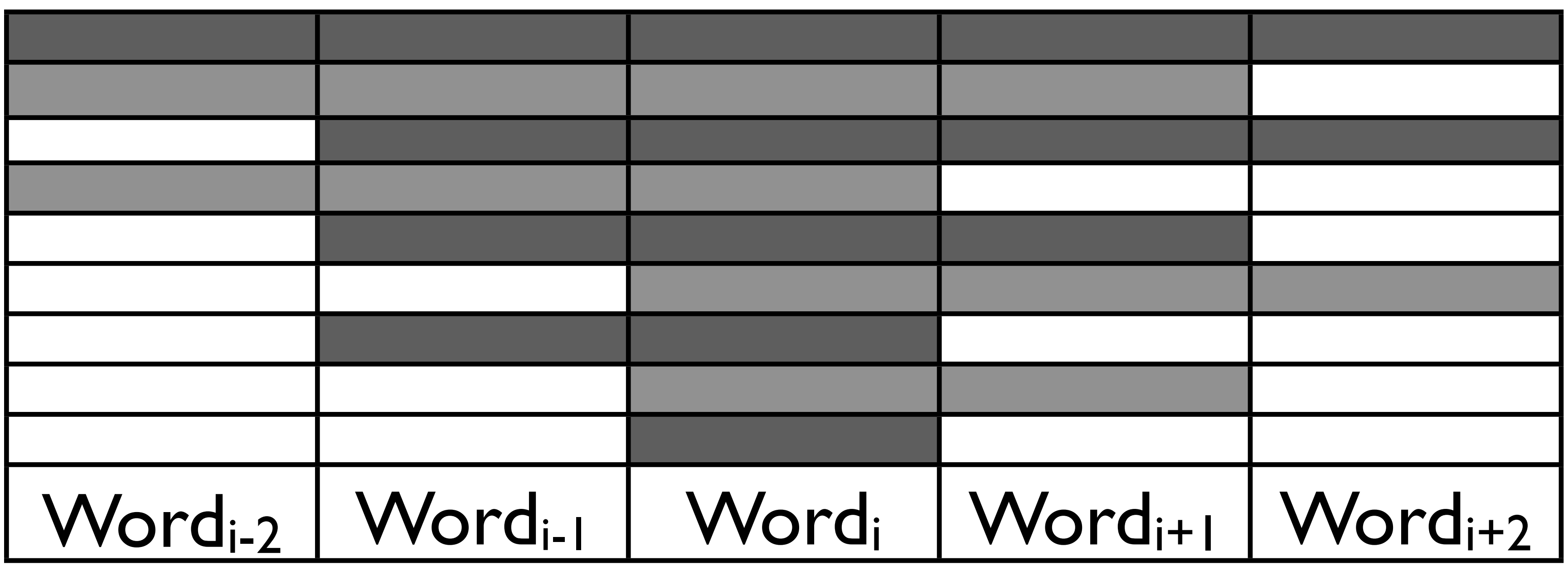
- Can we combine these energy-based classifiers with pitch and duration information to improve pitch accent detection further?

Extracted Features

- Pitch (f0): min, max, mean, stdev
  - Raw & speaker normalized
  - First order difference (Δf0)
- Energy: Min, max, mean, stdev
  - Extracted from 210 frequency regions from 0-20 Bark varying base frequency and bandwidth
    - Recall: Band between 2-20 Bark shows the best and most robust predictive power. [Rosenberg & Hirschberg '06]
- Duration
  - Length of the word and preceding and following pauses

Context Normalization

- Z-score normalized pitch and energy features based on acoustic information in surrounding words.
- 9 Context Windows



BDC Corpus

- Boston Directions Corpus (BDC) [Hirschberg & Nakatani ‘96]
  - Speech elicited from direction-giving tasks
  - Subjects delivered spontaneous-elicited monologues. 2 weeks later, subjects read transcribed versions of their monologues
    - 4 Speakers: 3 male, 1 female
    - 50 mins **Read Speech** (10818 wds)
      - 57% unaccented
    - 60 mins **Spontaneous Speech** (11627 wds)
      - 51% unaccented
  - Manually ToBI labeled including word boundaries

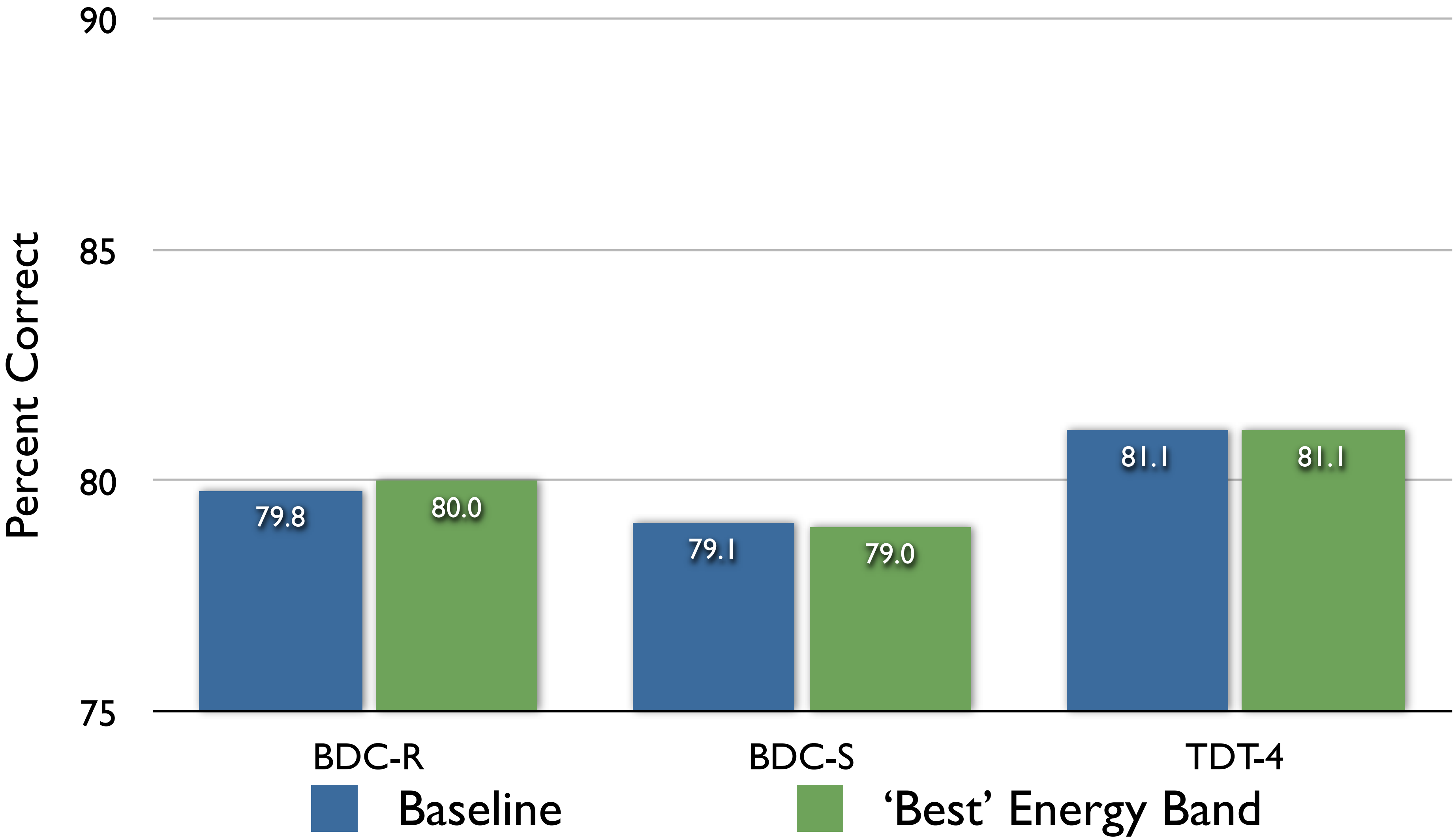
TDT-4 Corpus

- One 30-minute broadcast news (**BN**) show
- ASR word boundaries
- Automatic Speaker diarization
  - 25 speakers
- Manually labeled pitch accents and intonational phrase boundaries
- 20 mins of **speech** (3326 words)
  - 50.6% unaccented

Baseline Experiments

- Train a decision tree using all pitch and duration features and full-spectrum energy features.
- Instead of full-spectrum energy features, use only those from the "best" frequency region, 2-20bark.

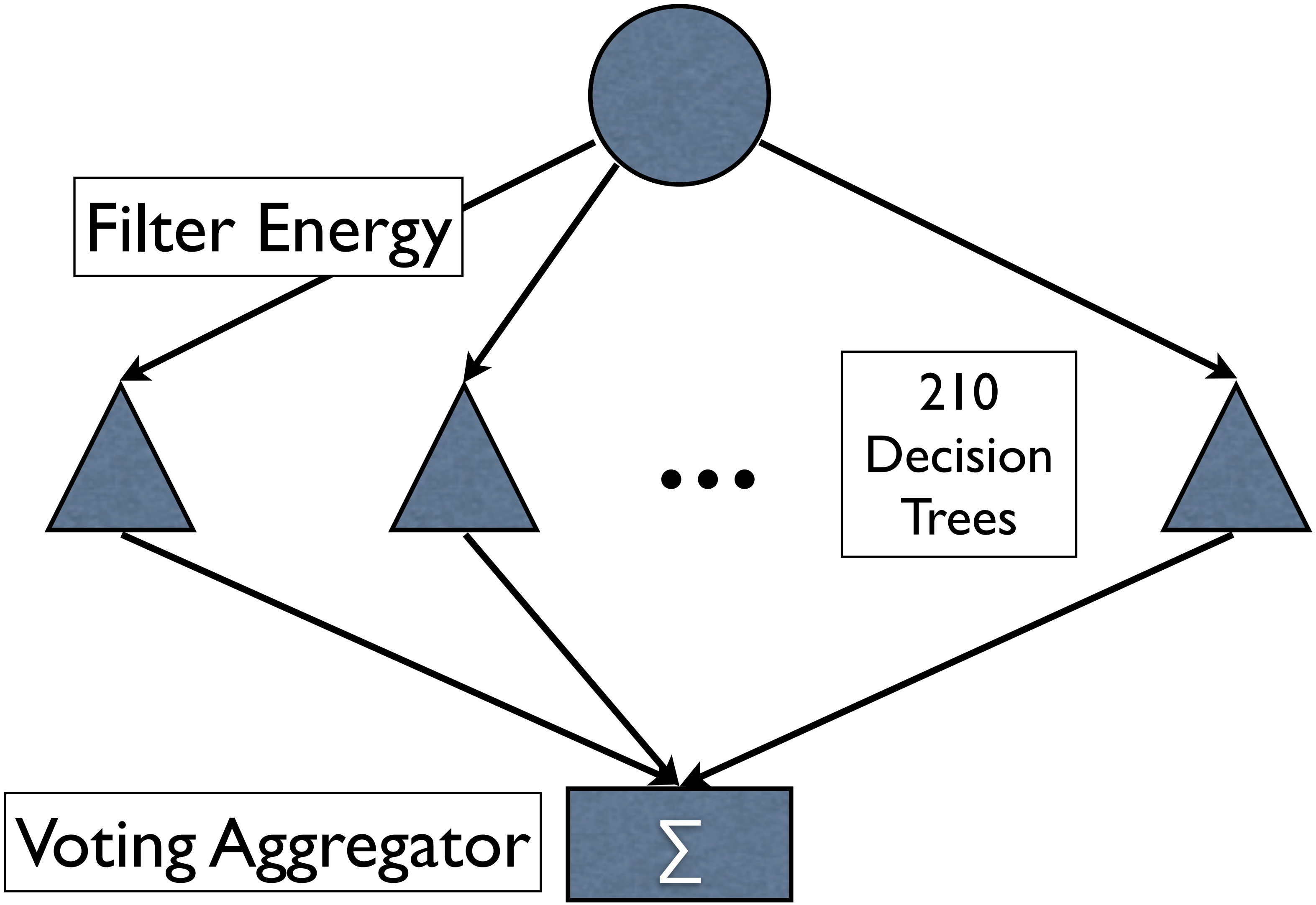
Results


90
85
80
75
Percent Correct
79.8
80.0
79.1
79.0
81.1
81.1
BDC-R
BDC-S
TDT-4
Baseline
'Best' Energy Band

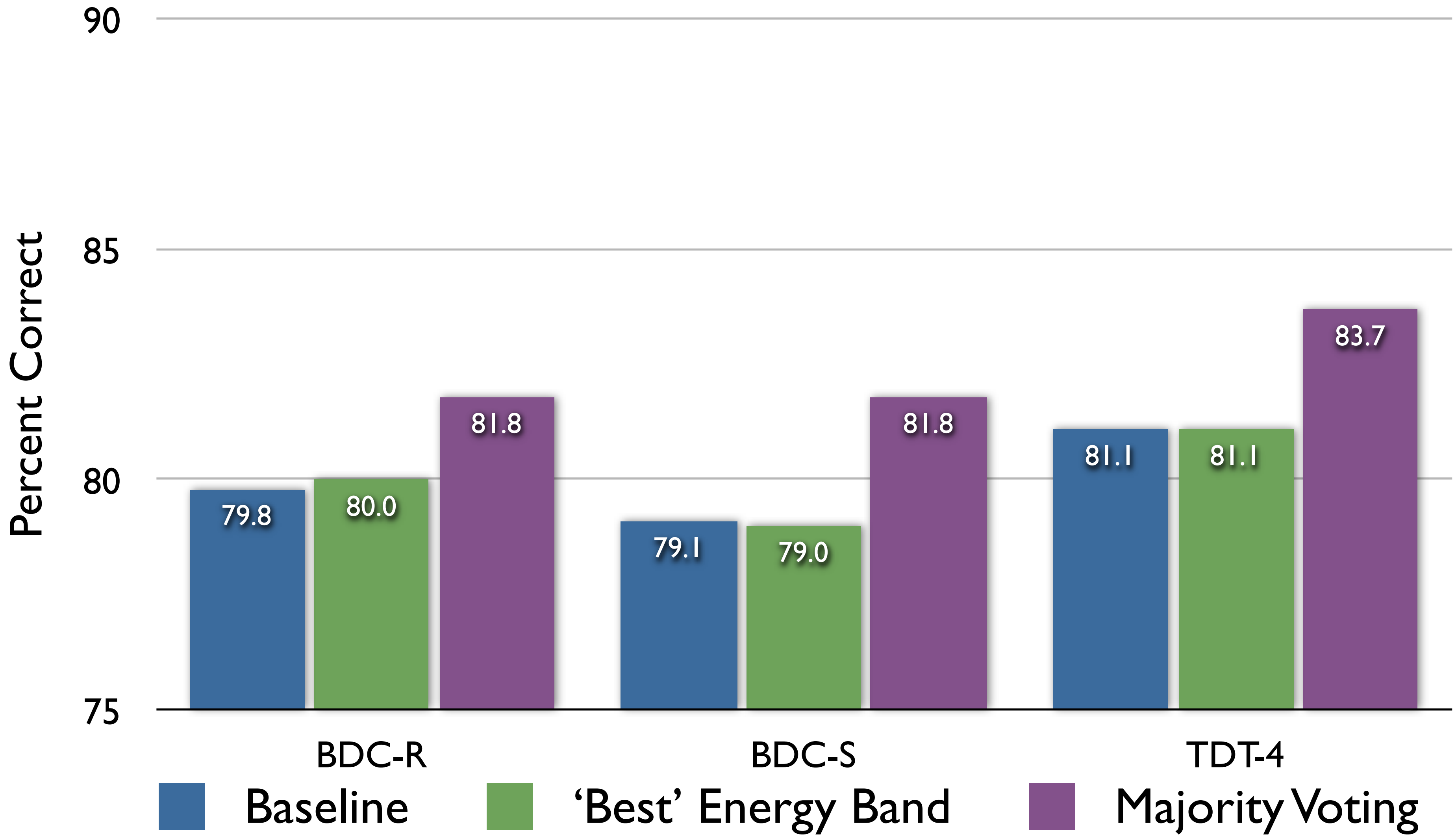
Energy-Based Predictors

- For each of 210 frequency regions, train a decision tree using **only energy features.**
  - 0-20bark, varying base frequency and bandwidth at 1 bark intervals.
- Combine these predictions using unweighted Majority Voting.

Majority Voting



Results

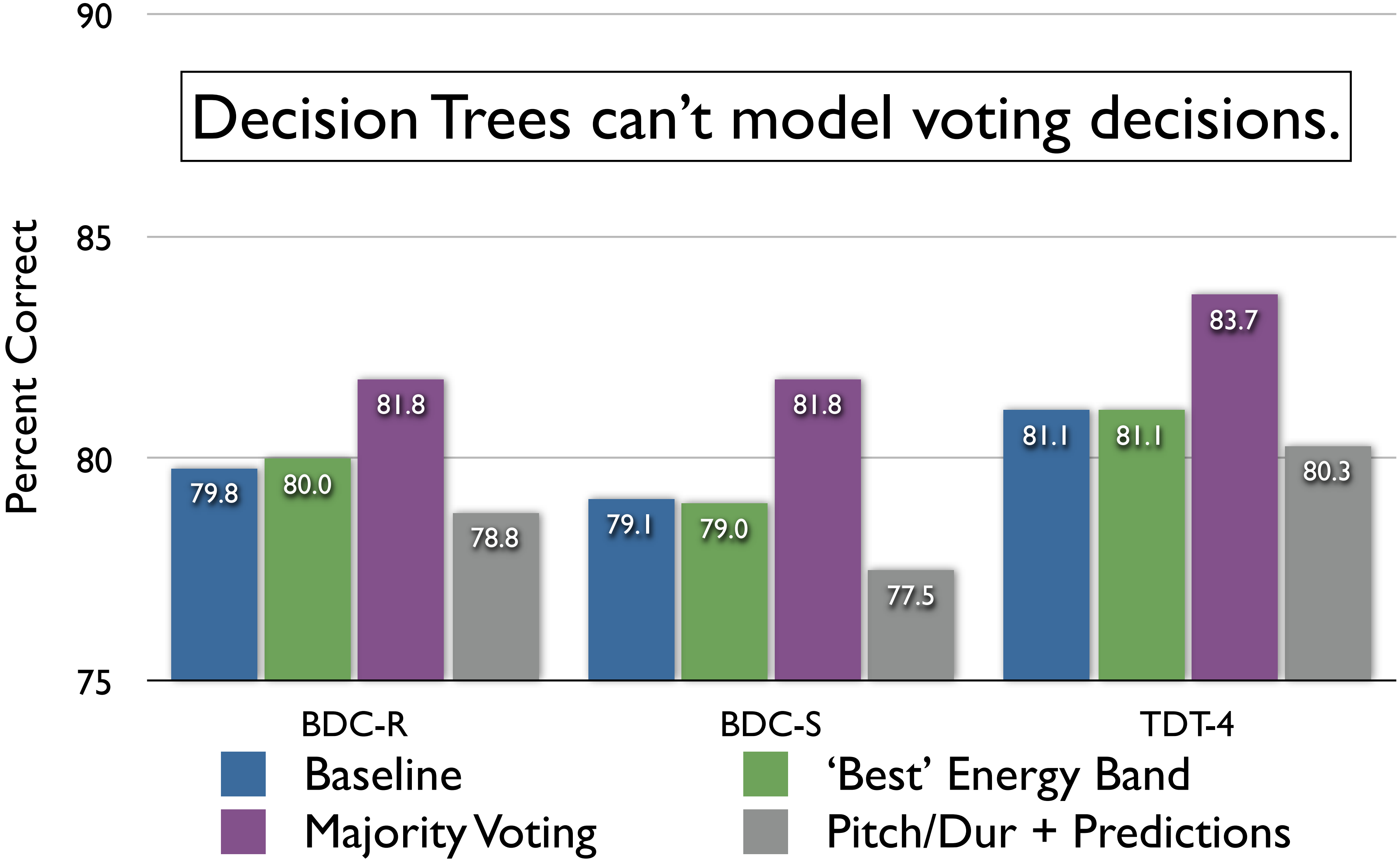


90
85
80
75
Percent Correct
79.8
80.0
81.8
79.1
79.0
81.8
81.1
81.1
83.7
BDC-R
BDC-S
TDT-4
Baseline
'Best' Energy Band
Majority Voting

Energy Predictions + Pitch and Duration Features

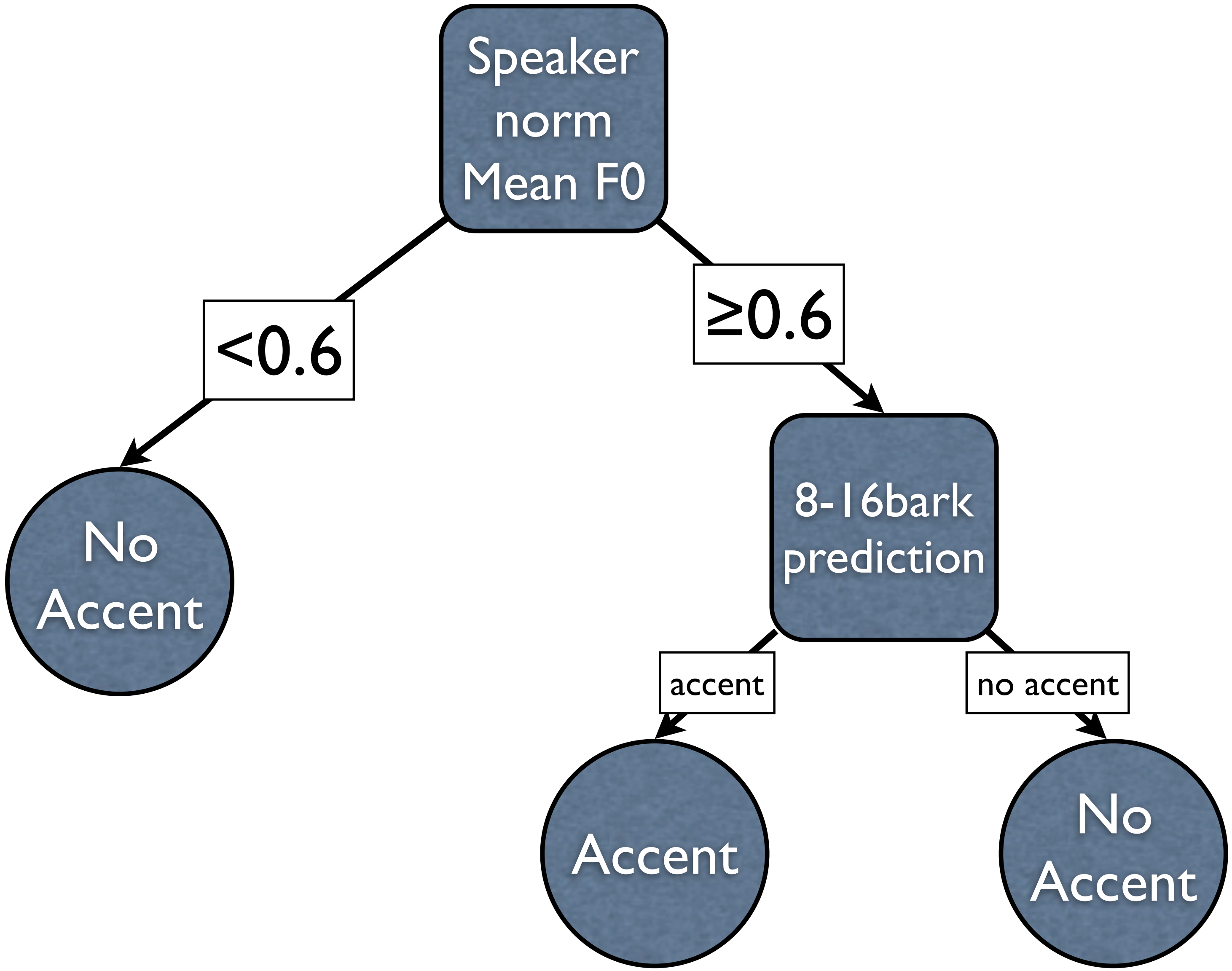
- Construct a feature vector using Pitch and Duration features as well as 210 Energy predictions.

Results



90
85
80
75
Percent Correct
Decision Trees can't model voting decisions.
79.8
80.0
81.8
78.8
79.1
79.0
81.8
77.5
81.1
81.1
83.7
80.3
BDC-R
BDC-S
TDT-4
Baseline
'Best' Energy Band
Majority Voting
Pitch/Dur + Predictions

Decision Tree Fragment

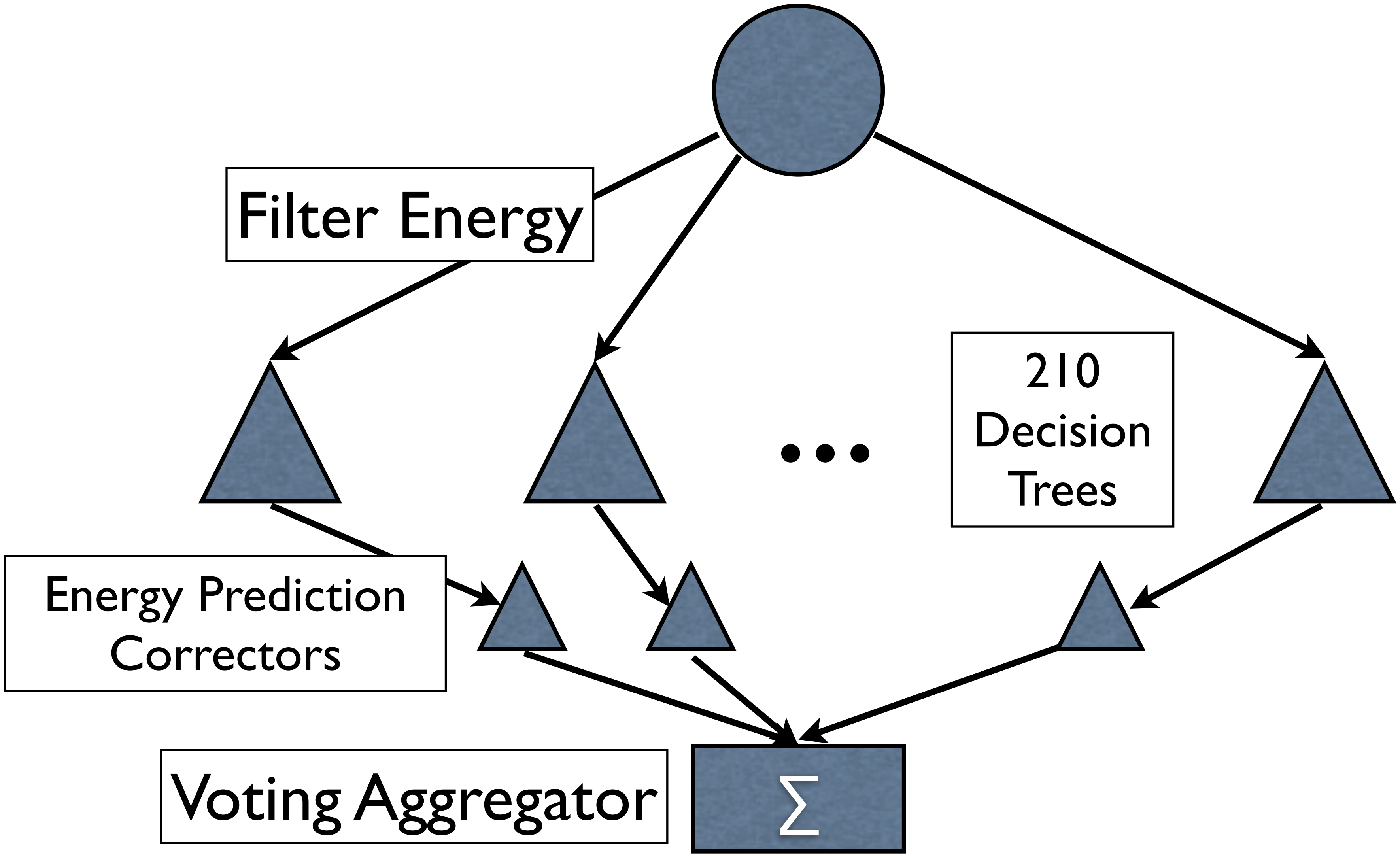


A. Rosenberg Interspeech '07

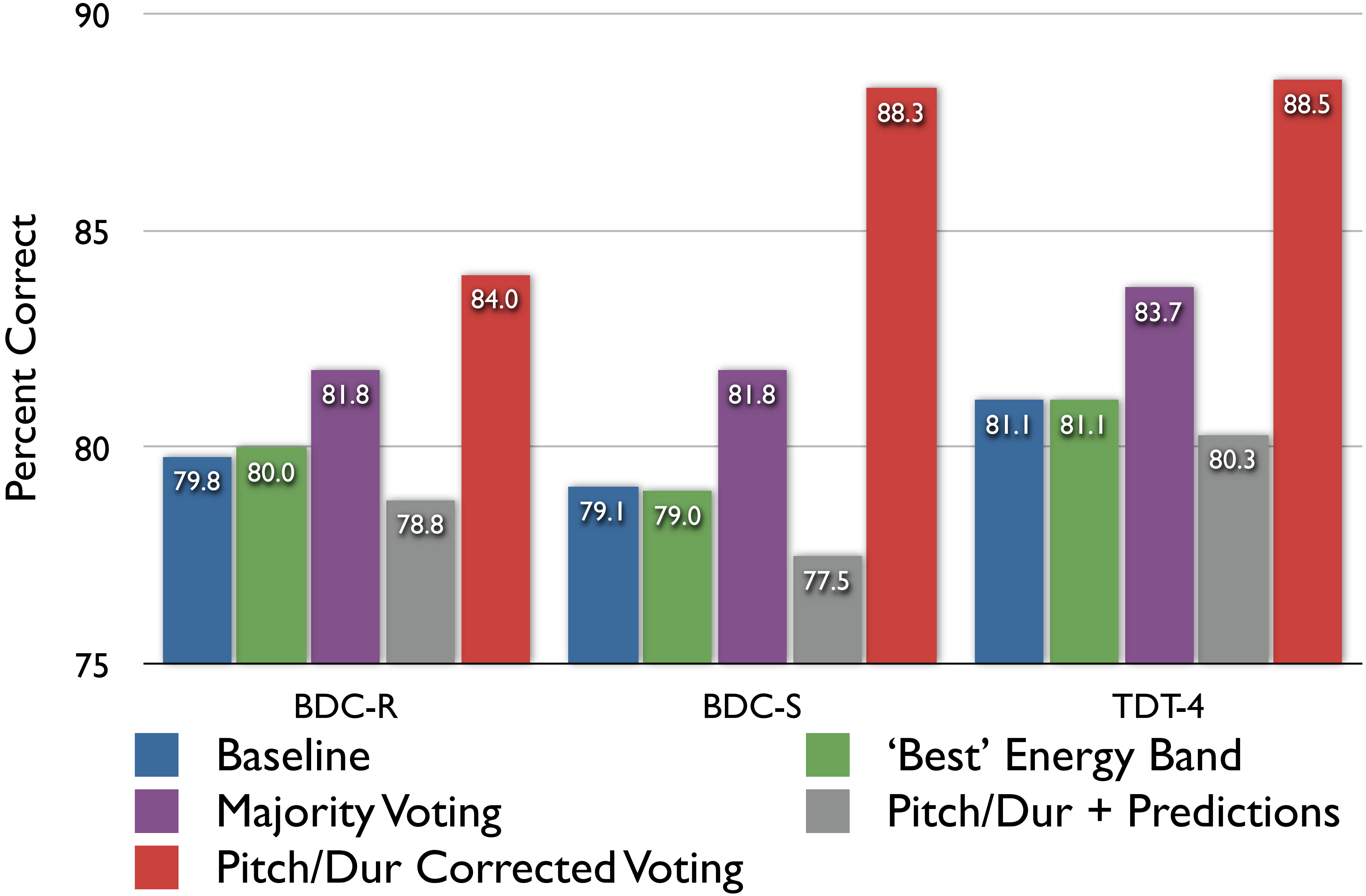
Pitch and Duration Correctors

- Can we predict, using pitch and duration information, whether an energy-based prediction will be correct or not?
- Train decision trees to predict “Correct” or “Incorrect”
  - Construct one corrector for each energy based predictor
  - Note: Corrector training data comes from cross-validation over **training data only**

Pitch and Duration Based Correctors



Results



90
85
80
75
Percent Correct
79.8
80.0
81.8
78.8
84.0
79.1
79.0
81.8
77.5
88.3
81.1
81.1
83.7
80.3
88.5
BDC-R
BDC-S
TDT-4
Baseline
'Best' Energy Band
Majority Voting
Pitch/Dur + Predictions
Pitch/Dur Corrected Voting

Conclusions & Future Work

- We presented a structured ensemble-based model that detects pitch accent with accuracy near human agreement
  - Speaker independent
  - Fairly robust to genre: Read, Spontaneous, BN
- Parallelizable, but computationally intensive
  - Identify redundant sets of frequency regions
- Include lexical and syntactic features
- Compare with other ensemble methods
- Evaluate on more corpora, particularly more BN
- Use hypothesized phrase boundaries to normalize acoustic features by phrase

Thank You.

Questions?

{amaxwell, julia}@cs.columbia.edu