MODELING VOCAL INTERACTION FOR TEXT-INDEPENDENT DETECTION OF INVOLVEMENT HOTSPOTS IN MULTI-PARTY MEETINGS

Kornel Laskowski
Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, USA
Cognitive Systems Lab, Universität Karlsruhe, Karlsruhe, Germany

Goal

Classify 60-second intervals of multi-party meetings as one of:

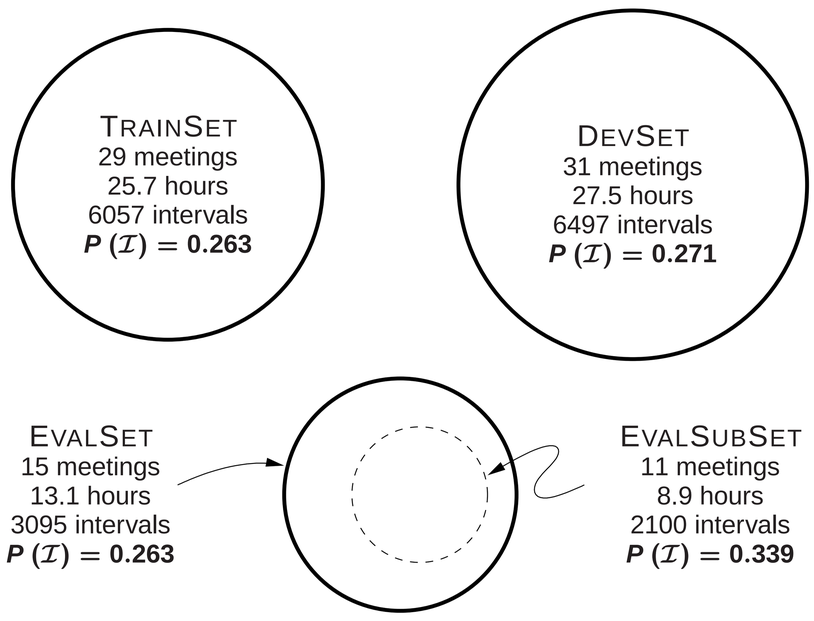
$$\begin{cases} \mathcal{I}, & \text{containing involved speech, or} \\ \neg\mathcal{I}, & \text{not containing involved speech.} \end{cases}$$

Definitions:

involvement	$\approx$	emotional activation	(Wrede & Shriberg, 2003)
hotspot	$\doteq$	region with involved speech	(multiple definitions exist)

Involvement hotspots are considered important for conversation understanding.

Data: ICSI Meeting Corpus (Janin et al, 2003)

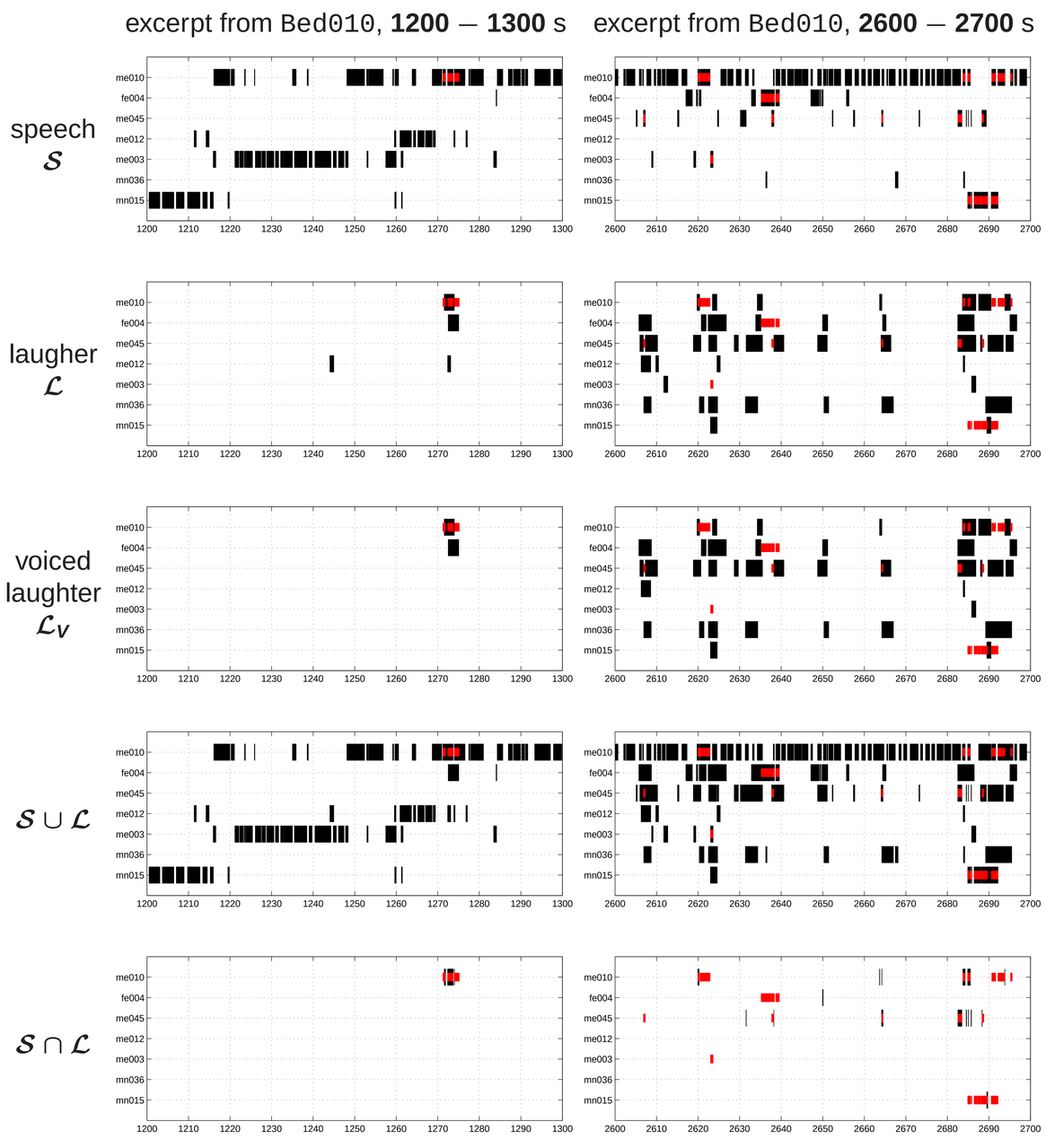


Corpus includes:

1. reference segmentation of speech (Shriberg et al, 2004)
2. reference segmentation of laughter (Laskowski & Burger, 2007)
3. reference segmentation of involved speech (Wrede et al, 2005)

Segmentation Types

Involved speech shown in red for comparison.



Features

Per 60-second interval:

$\{p_j^{\mathcal{V}}\}$	proportion of time each participant vocalizes, ranked in descending order of magnitude
$\{o_j^{\mathcal{V}}\}$	proportion of time spent in simultaneous vocalization (overlap), ranked in ascending order of degree of overlap
T_j^{PI}	single coefficient of interpolation between logistic vocal interaction model and randomness
$\{T_j^{PS}\}$	per-participant coefficients of interpolation between logistic vocal interaction model and randomness, ranked in descending order of magnitude

Classification Results on EVALSET

- Guessing randomly (TRAINSET prior): 61.23%
- Guessing majority class (TRAINSET prior): 73.67%
- Linear-kernel SVM classification
  - model training with TRAINSET
  - feature selection with DEVSET

Segm. Type	Feature Type				
	Static		Dynamic		all
	$\{p_j^{\mathcal{V}}\}$	$\{o_j^{\mathcal{V}}\}$	T_j^{PI}	$\{T_j^{PS}\}$	
$\mathcal{S}$	75.2	73.9	75.3	73.5	75.5
$\mathcal{S} \cup \mathcal{L}$	77.7	80.1	77.1	76.5	80.0
$\mathcal{L}$	80.6	81.2	80.8	75.5	80.0
$\mathcal{L}_V$	81.4	82.1	81.6	75.9	81.9
$\mathcal{L} \cap \mathcal{S}$	**83.0**	82.1	78.1	79.0	**84.2**
all	**83.4**	82.6	82.7	75.4	**84.0**

Observations:

1. All feature type and segmentation type combinations (except one) lead to accuracies which exceed majority class guessing.
2. Improvement in accuracy as increasingly smaller subsets of all laughter are used,
$$\mathcal{S} \subseteq \mathcal{S} \cup \mathcal{L} \supseteq \mathcal{L} \supseteq \mathcal{L}_V \supseteq \mathcal{S} \cap \mathcal{L} \ .$$
3. Static features frequently outperform dynamic features.
4. Feature combination results in improved performance.
5. **All feature and segmentation types: 84.0% $\equiv$ 39.2% relative reduction of error.**

Comparison with Human Performance on EVALSUBSET

Utterance-level agreement:

1. **pilot baseline study** (Wrede & Shriberg, 2003)
   - **two-way** distinction: INVOLVED versus NOT-INVOLVED
   - similar but smaller dataset than here
   - 5 **native** english-speaking raters: $\kappa = 0.63$
   - 3 **non-native** english-speaking raters: $\kappa = 0.52$
2. **baseline study** (Wrede & Shriberg, 2005):
   - **three-way** distinction: INVOLVED versus RAISED VOICE versus NEITHER
   - same data as used here (EVALSUBSET)
   - 2 raters: $\kappa = 0.58$

"60 second interval"-level agreement:

	A	*B*	ref	hyp
A		0.68	0.84	0.59
B	0.68		0.83	0.57
ref	0.84	0.83		0.54
hyp	0.59	0.57	0.54	

3. **current study**, using EVALSUBSET
   - **two-way** distinction: CONTAINING INVOLVED SPEECH or NOT CONTAINING INVOLVED SPEECH
   - human-human agreement: $\kappa = 0.68$
   - system-human agreement: $\kappa \in [0.57, 0.59]$
   - **difference $\Delta\kappa \approx 0.1$ similiar in magnitude to utterance-level native & non-native agreement**

Conclusions

1. Baseline (majority class guessing) accuracy: 73.67%
2. Automatic system accuracy: 84.0% (39.%rel reduction of error)
3. Of explored features, static laughter features far more relevant than any speech features.
4. Ceiling accuracy using vocal activity detection systems which do not distinguish between speech and laughter: 80.1% (24%rel reduction of error)

kornel@cs.cmu.edu