Carnegie Mellon

FINDING EMOTIONALLY INVOLVED SPEECH USING IMPLICITLY PROXIMITY-ANNOTATED LAUGHTER

Kornel Laskowski
Language Technologies Institute
Carnegie Mellon University
Pittsburgh PA, USA

Goal & Approach

Find meeting intervals containing **emotionally involved speech**. (ignore who is involved)

HOW?

Use laughter as a cue.

PROBLEMS TO OVERCOME?

- the most relevant laughter is that which is **produced simultaneously with the *laugher's* speech**:
  - acoustically and pragmatically difficult to model
  - rare ($<$ **0.4**% of all vocalization by time)

Analysis of Hotspot Detection, ICSI Meeting Corpus

Does a 60-second interval of meeting time contain involved speech?

1. Given a vocalization segmentation, compute the interval proportion that each participant vocalizes.
2. Train an SVM using TRAINSET; feature select using DEVSET.

Feature Set	Accuracy, %		F-Score, %	
	DEVSET	EVALSET	DEVSET	EVALSET
guess, priors	*60.9*	*61.2*	—	—
guess, major.	*72.9*	*73.7*	—	—
$\mathbf{f}(\mathcal{S})$	*74.4*	*75.3*	34.4	28.0
$\mathbf{f}(\mathcal{L})$	*80.4*	*80.8*	64.6	64.8
$\mathbf{f}(\mathcal{L} \cap \mathcal{S})$	*81.2*	*83.3*	68.9	70.6
$\mathbf{f}(\mathcal{L} \cap \neg\mathcal{S})$	*80.4*	*80.8*	64.6	64.8
$\mathbf{f}(\mathcal{L}_V)$	*81.5*	*81.6*	65.1	64.3
$\mathbf{f}(\mathcal{L}_V \cap \mathcal{S})$	*82.9*	*85.6*	69.5	67.1
$\mathbf{f}(\mathcal{L}_V \cap \neg\mathcal{S})$	*81.5*	*81.4*	65.0	67.1
$\mathbf{f}(\mathcal{L}_U)$	*76.4*	*77.4*	56.3	55.1
$\mathbf{f}(\mathcal{L}_U \cap \mathcal{S})$	73.7	72.6	27.6	21.9
$\mathbf{f}(\mathcal{L}_U \cap \neg\mathcal{S})$	76.4	77.4	56.3	55.1

Questions

1. Can conversational hotspots be detected with similar accuracy even if "speech-laughs", the most relevant but also the likely most difficult-to-detect type of laughter, are **ignored**? How?
2. Is there a characteristic of laughter which is discriminative of laughter's **relevance** to this problem?

(Hypothesis: proximity to **laugher's** speech.)

Proximity Annotation via $\mathcal{S}$-Cosegmentation Masking Operators

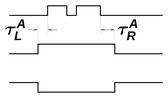
voiced laughter, $\mathcal{L}_V$
speech, $\mathcal{S}$

a snippet of vocal activity

$\mathcal{S}$

$\Upsilon_1 = \sigma\left(\tau_L^A, \mathcal{S}, \tau_R^A\right)$

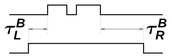
$\Upsilon_2 = \neg\sigma\left(\tau_L^A, \mathcal{S}, \tau_R^A\right)$



STEP 1. instants at least (τ_L^A, τ_R^A)-seconds away from speech

$\mathcal{S}$

$\Upsilon_3 = \sigma\left(\tau_L^B, \mathcal{S}, \tau_R^B\right)$

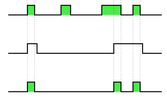


STEP 2. instants at most (τ_L^B, τ_R^B)-seconds away from speech

$\mathcal{L}_V$

$\Upsilon_4 = \Upsilon_2 \cap \Upsilon_3$

$\Upsilon_5 = \Upsilon_4 \cap \mathcal{L}_V$



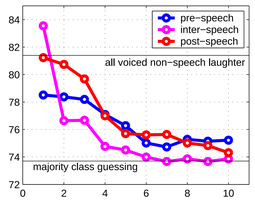
STEP 3. laughter in the intersection

Findings

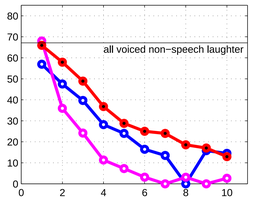
1. Annotating voiced laughs with **proximity to the laugher's speech** improves performance, and approaches or exceeds performance using speech-laughs.
2. Laughter which is distant from the laughter's speech appears deleterious to the task of detecting involved speech, **regardless of that laughter's distance to the speech of others.**

Experiment I: 1-second slices, τ-seconds away

EVALSET accuracy, vs τ

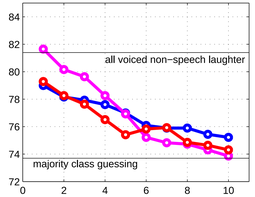


EVALSET F-score, vs τ

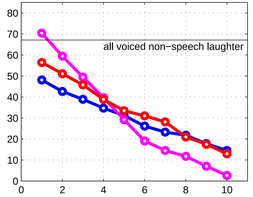


Experiment III: $(10 - \tau)$-second slices, τ-seconds away

EVALSET accuracy, vs τ



EVALSET F-score, vs τ



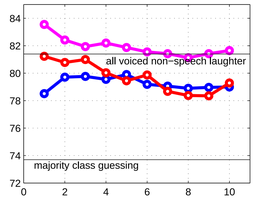
Conclusions & Impact

I. Downstream **conversational speech understanding systems** can benefit from the exclusion of laughter which is distant from the laugher's speech, even long and loud intervals of group laughter (unless the most recent or most imminent speaker is also laughing).

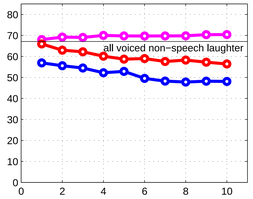
II. **Laughter detection** should focus on laughter which is proximate to laugher's speech, and, to be useful to the upstream nalysis of emotionally involved speech, can safely ignore all other laughter.

Experiment II: τ-second slices, 0-seconds away

EVALSET accuracy, vs τ



EVALSET F-score, vs τ



Experiment IV: all 1-second slices in $\tau \in [1, 10]$

Use all 1-second pre-/inter-/post-$\mathcal{S}$ slices $\in$ **[1, 10]** (270 features)

Feature Set	Accuracy, %		F-Score, %	
	DEVSET	EVALSET	DEVSET	EVALSET
$\mathbf{f}(\mathcal{L} \cap \mathcal{S})$	81.2	83.3	68.9	70.6
$\mathbf{f}(\mathcal{L}_V \cap \mathcal{S})$	82.9	85.6	69.5	67.1
$\mathbf{f}(\mathcal{L}_V \cap \neg\mathcal{S})$	**81.5**	**81.4**	**65.0**	**67.1**
$\mathbf{F}(\mathcal{L}_V \cap \neg\mathcal{S})$	**83.2**	**84.4**	**71.2**	**70.3**

IEEE International Conference on Acoustics, Speech, and Signal Processing, Dallas TX, USA

kornel@cs.cmu.edu