A FRAME-SYNCHRONOUS PROSODIC DECODER FOR TEXT-INDEPENDENT DIALOG ACT RECOGNITION

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

Goal & Approach

Text-independent dialog act (DA) segmentation *and* classification in **privacy-sensitive** settings.

cannot compute ASR features
⟶ no **words** or **word boundaries**

HOW?

- anchor feature computation to **unrecognized** speech
- construct an acoustic ASR-like decoder, whose states are
  - **not** phonemic sub-segments
  - but **prosodic sub-phrases**

Questions

1. Can the current 32 ms instantaneous prosodic feature vector and 8 ms frame step be extended to the decoder frame step of 100 ms, without negative impact on DA recognition performance?
2. Does feature-space combination with temporally adjacent features from the target participant improve DA recognition?
3. Does feature-space combination with temporally adjacent features from non-target participants improve DA recognition?

Findings

1. **A larger frame size (256 ms) improves** DA classification only:
   ⟶ mean DA ***F***: +**2.5**%abs
   ⟶ boundary ***F***: −**0.2**%abs
2. **Target-speaker prosodic context (1 s) improves** DA recognition:
   ⟶ mean DA ***F***: +**1.7**%abs
   ⟶ boundary ***F***: +**4.0**%abs
3. **Non-target-speaker prosodic context (1 s) improves** DA segmentation only:
   ⟶ mean DA ***F***: −**1.5**%abs
   ⟶ boundary ***F***: +**1.8**%abs

Conclusions & Impact

I. Significant improvements in text-independent HMM-based DA recognition can be achieved with **longer audio frames** (256 ms vs 32 ms) and **feature stacking** (1 s).
II. Improvements observed despite the resulting **much smaller amounts of training material**.
III. Non-target-speaker prosody improves DA segmentation.

FUTURE WORK: *Does non-target-speaker prosody improve over only non-target-speaker speech activity?*

Techniques

Have:

- hidden Markov model decoder
  - 100 ms frame step, 100 ms frame size
  - specialized topology (split-and-merge talkspurts)
  - 8 DA types
  - 3 DA boundary types
- instantaneous, frame-level prosodic features
  - 8 ms frame step, 32 ms frame size
  - loudness (2): log-energy, delta-log-energy
  - speaking rate (2): cosine-Mel-energy, cosine-log-Mel-energy
  - voice quality (1): max-norm-autocorrelation
  - intonation (7): fundamental frequency variation (FFV) coefficients

Want **to model prosodic context**:

- decoder frame step = feature computation frame step
- consider adjacent target-speaker prosody
- consider adjacent non-target-speaker prosody

Experiments on ICSI Meeting Corpus, EVALSET (11 meetings)

	Baseline (BL) 32 ms / 8 ms (12.5 frames)		Experiment 1 256 ms / 100 ms (1 frame)		Experiment 2 256 ms / 100 ms (11 frames)			Experiment 3 256 ms / 100 ms (w/ non-target)	
	g-Opt	c-Opt	g-Opt	c-Opt	g-Opt	c-Opt	%rel, BL	g-Opt	c-Opt
DA Types									
mean ***F***	31.5	33.7	35.5	36.2	**36.6**	**37.9**	+12.5—	35.8	36.4
F, floor holder	37.7	39.5	45.2	45.2	**45.8**	**48.2**	**+22.0** **	45.5	47.8
F, hold	25.0	17.1	**25.3**	20.6	21.6	**21.6**	**+26.3** *	16.5	12.0
F, floor grabber	7.2	7.2	8.2	8.2	**10.1**	**10.1**	**+40.3** *	6.9	7.3
F, backchannel	48.0	64.6	57.5	64.4	59.9	64.2	**−0.6**	**63.5**	**64.9**
F, acknowledgment	19.0	20.9	25.2	25.2	25.3	25.3	**+21.1** **	**25.8**	**27.1**
F, accept	9.5	8.9	17.5	17.5	**21.9**	**22.5**	**+152.8** **	20.9	22.1
F, statement	85.8	91.8	**88.8**	91.9	88.5	**92.0**	**+0.2** **	88.6	91.9
F, question	**19.6**	**19.6**	16.6	16.6	19.4	19.4	**−1.0**	18.5	18.5
classification error	25.9	16.6	**21.0**	15.9	21.7	**15.8**	**−4.8**—	21.4	16.1
DA Termination Types									
F, completed	59.1	59.1	59.9	59.9	**62.7**	**63.8**	**+8.0** **	61.0	63.3
F, interrupted	10.5	11.8	6.7	9.6	14.6	14.6	**+23.7**	**28.4**	**28.4**
F, abandoned	2.4	3.6	2.4	4.3	**6.1**	**7.0**	**+94.4** **	4.6	5.4
any type, ***F***	62.6	62.6	62.4	62.4	66.4	66.4	**+6.1**—	**66.9**	**68.2**
NIST error	66.5	63.0	66.5	63.0	66.1	58.5	**−7.0**—	**65.9**	**56.0**

kornel@cs.cmu.edu