COMPARING THE CONTRIBUTIONS OF CONTEXT AND PROSODY IN TEXT-INDEPENDENT DIALOG ACT RECOGNITION

Kornel Laskowski[1] and Elizabeth Shriberg[2,3]

[1] Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, USA
[2] SRI International, Menlo Park CA, USA
[3] International Computer Science Institute, Berkeley CA, 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?

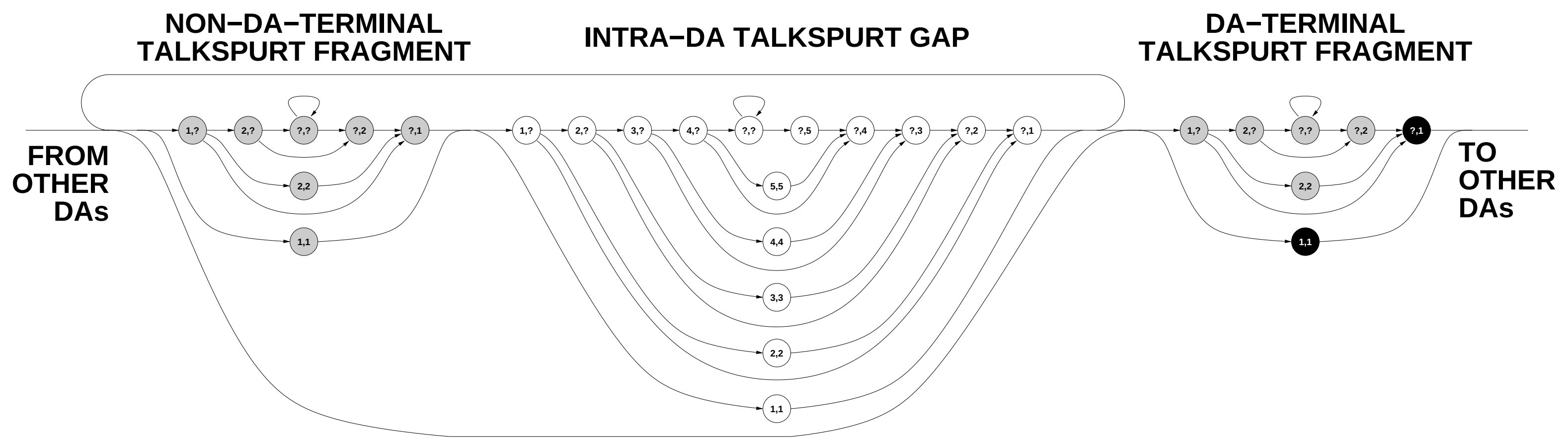
- anchor feature computation to **unrecognized** speech
- construct an acoustic ASR-like decoder, whose states are
  - **not** phonemic sub-segment units
  - but **prosodic sub-phrase** units
  - vocabulary consists not of words, but of **dialog acts**

Questions

0. *(implicit) Is DA recognition at all* ***possible****, using only features describing loudness, intonation, voice quality, speaking rate, intra-talkspurt location, and inter-participant timing?*
1. How much does context versus prosody contribute to text-independent DA recognition?
2. To what extent are context and prosody features complementary and does this depend on DA type?
3. How do these text-independent systems compare to a system that uses the words?

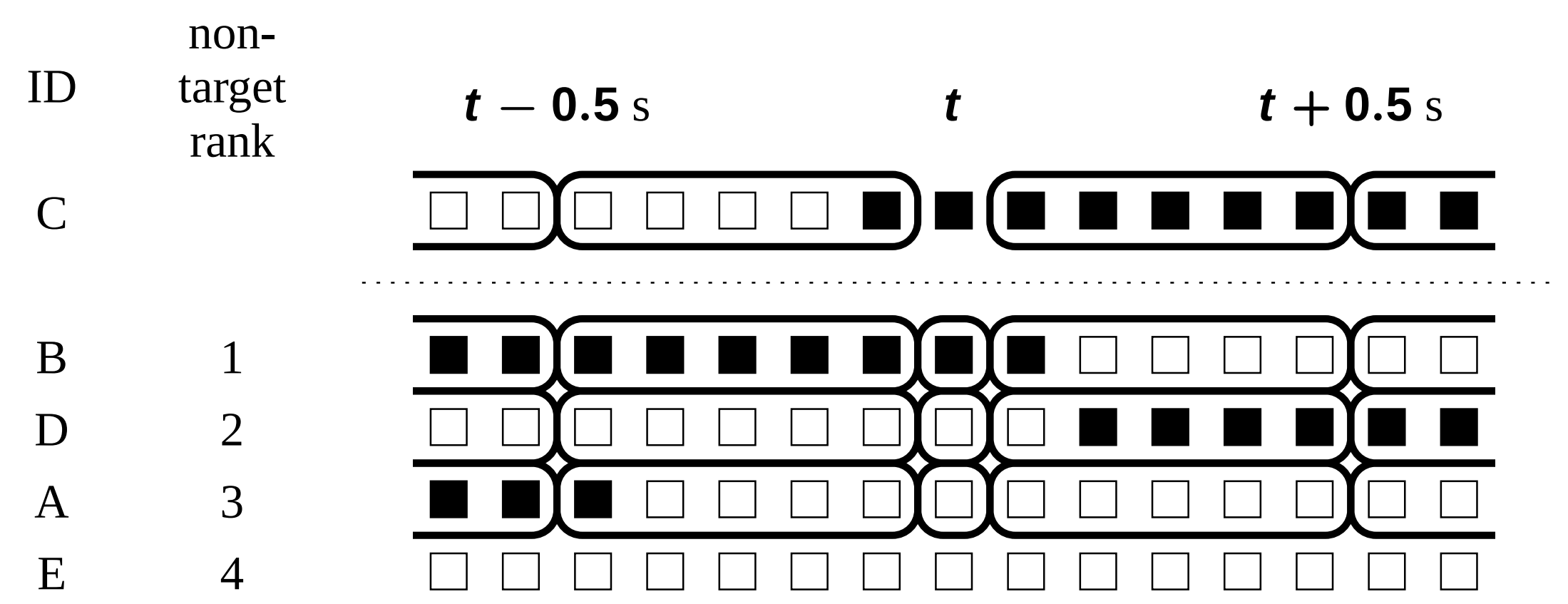
HMM Topology

- any DA subtopology can transition to any DA subtopology
- DA subtopologies connected via inter-DA gaps
- total number of states: 1220
- transitions probabilities trained from forced-alignment
- speech state emissions modeled with GMMs


NON–DA–TERMINAL TALKSPURT FRAGMENT
INTRA–DA TALKSPURT GAP
DA–TERMINAL TALKSPURT FRAGMENT
FROM OTHER DAs
TO OTHER DAs

Speech/Non-speech Context Features

- 10-second S/NS posterior context, in 0.5-second tiles
- target speaker and 3 locally most talkative interlocutors



Instantaneous Prosody Features

- frame-level features for target speaker only
- features computed only for speech states
- 12 features:
  - energy
  - delta-energy
  - normalized autocorrelation maximum
  - Mel-filterbank magnitude cosine difference
  - Mel-filterbank log-magnitude cosine difference
  - 7 FFV intonation filterbank features

Findings

1. **Speech/non-speech context and instantaneous prosody achieve comparable performance.**
   ⟶ mean ***F***-scores using prosody are ≈2.5% higher.
2. **The two feature streams are (∼ perfectly) complementary.**
   ⟶ mean ***F***-scores (excl. effect of topology) are additive.
3. **Combined performance approaches lexical system performance in several cases.**
   ⟶ DA types: questions, 78%rel; backchannels, 87%rel.
   ⟶ DA boundary types: compl, 92%rel; interr, 131%rel.

Conclusions & Impact

I. Automatic DA recognition **is possible** in privacy-sensitive settings; the presented techniques achieve surprisingly good results **without words or word boundary information**.

II. **Conversational prosody can be modeled** directly, using standard acoustic modeling techniques and HMM decoding, **independently of automatic speech recognition**.

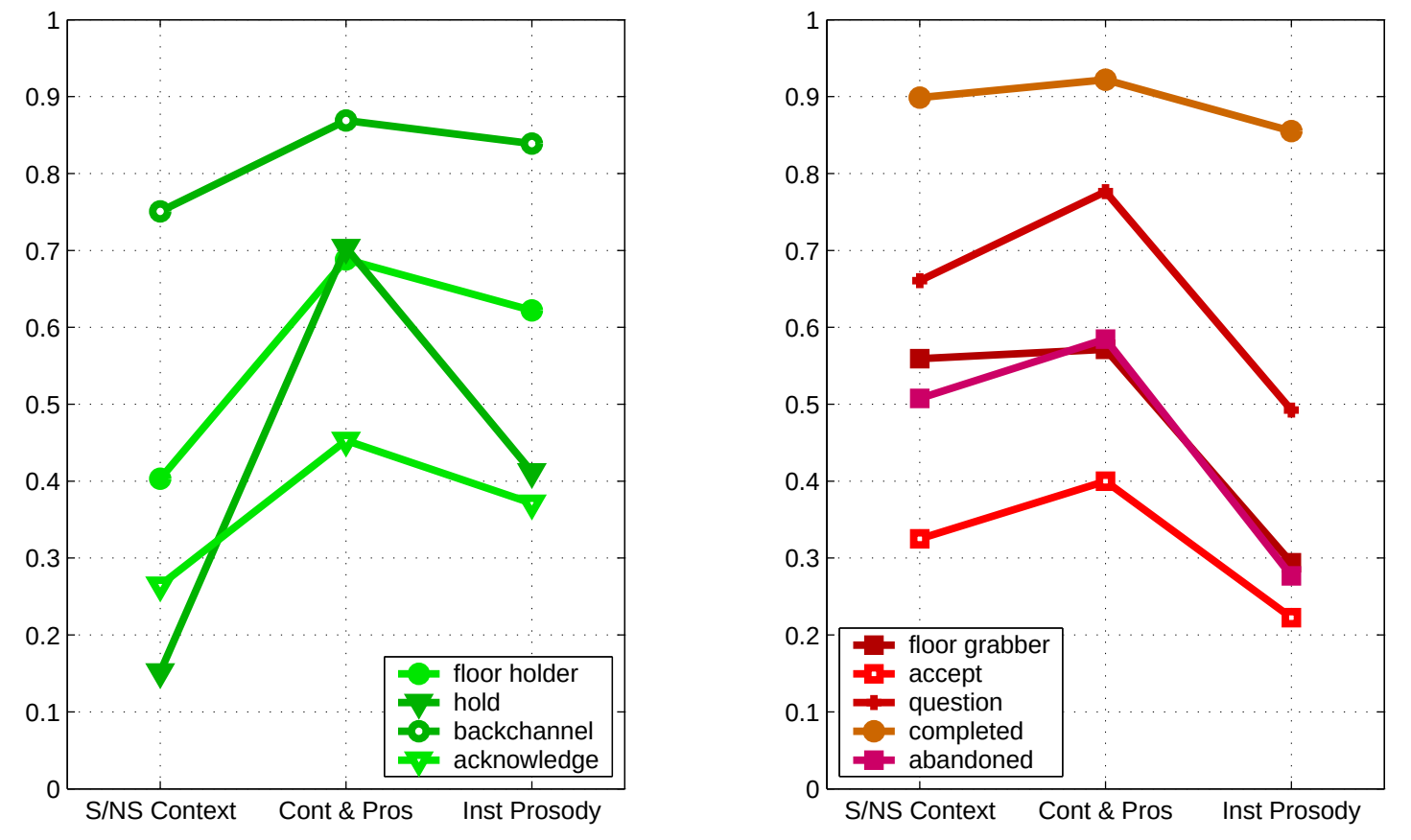
III. **Instantaneous prosody and speech/non-speech context** provide important and complementary DA-discriminative information, whose joint utility **approaches that of lexical information**.

Experiments on ICSI Meeting Corpus, *F*-scores on EvalSet (11 meetings)

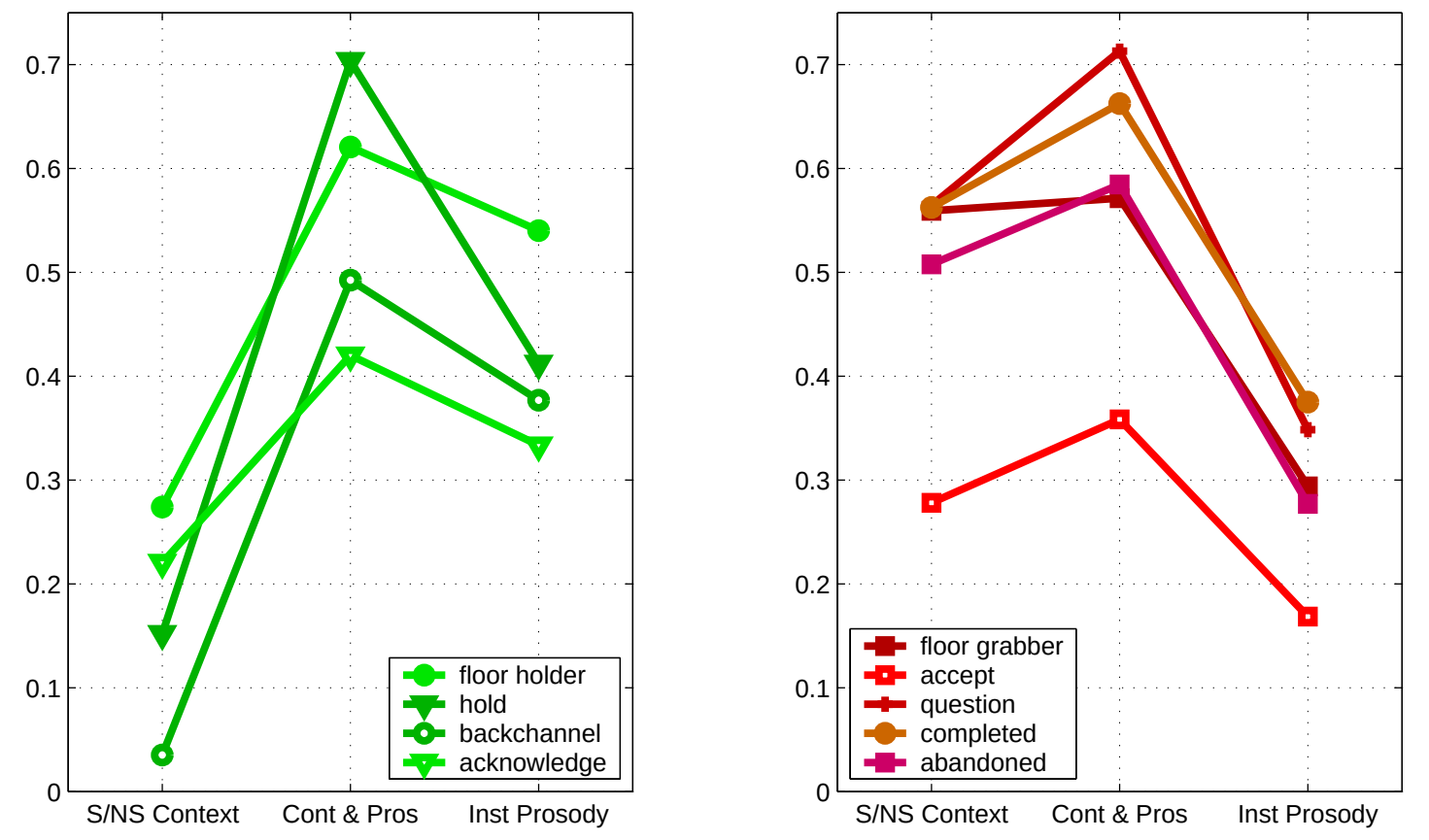
		Topo	Context		Prosody		Cont & Pros		Lex 2-grams	
			g-Opt	c-Opt	g-Opt	c-Opt	g-Opt	c-Opt	g-Opt	c-Opt
DA Types										
mean	prior	21.8	29.3	31.1	31.5	33.7	38.4	39.8	53.0	54.5
floor holder	2.7%	11.3	24.0	25.6	37.7	**39.5**	43.5	**43.7**	62.3	63.5
hold	0.3%	†0.0	8.5	†6.3	25.0	**17.1**	31.8	‡**29.2**	33.9	‡41.5
floor grabber	0.6%	0.0	12.5	†**13.7**	7.2	7.2	11.6	†**14.0**	24.5	24.5
backchannel	2.8%	†57.1	54.7	†57.8	48.0	**64.6**	64.5	**66.9**	77.0	77.0
acknowledge	1.5%	3.2	15.7	14.9	19.0	**20.9**	24.2	**25.6**	56.3	56.3
accept	1.1%	2.6	12.3	†**13.0**	9.5	8.9	14.0	†**16.0**	38.1	40.0
statement	84.5%	†91.4	82.3	†91.3	85.8	‡**91.8**	87.3	‡**91.8**	91.9	93.3
question	6.6%	8.8	23.9	**26.3**	19.6	19.6	30.4	**30.9**	39.8	39.8
DA Termination Types										
completed		53.1	58.3	**62.1**	59.1	59.1	63.4	**63.7**	68.0	69.1
interrupted		0.0	22.6	**22.6**	10.5	11.8	26.0	**28.7**	21.9	21.9
abandoned		0.0	6.6	† **6.6**	2.4	3.6	5.4	† **7.6**	11.4	13.0

(“g-Opt” systems = optimized for mean ***F***-score; “c-Opt” systems = optimized for specific conditions)

text-independent c-Opt performance, normalized by “Lex 2-gram”



text-independent c-Opt performance, normalized by “Lex 2-gram” and excluding effect of “Topo”



kornel@cs.cmu.edu, ees@speech.sri.com