

LING 575

Recognition and Understanding of Prosody

John T. McCranie
jtm37@uw.edu

28/Apr/2016

Prosody. Prosody? Prosody!

- non-lexical tone, intonation, rhythm, stress patterns
- suprasegmental
- pitch, duration, energy
- English only?

Prosody modeling for automatic speech recognition and understanding

- prosody is **exclusive** to spoken language
- additional info to text
- partially redundant source for error correction
- might be useful to improve systems: sentence segmentation, disfluency detection, topic segmentation, dialog acts, word recognition

- classification problem: $P(S|W, F)$
- raw features: F0, segment durations, energy
- derived features: F0 baseline, pitch range
- decision trees

- disfluency detection
- topic segmentation
- turn-taking in meetings

- sentence segmentation: in some cases the prosodic model alone performed better than the LM alone, pause duration
- dialog act labeling: disambiguation backchannel ("right") / agreement ("Right!")
- word recognition in conversation: improvements for task-oriented dialogs, but not large-vocabulary usage

Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog

- same prosody model as previous paper, added emotion labels
- language models are poor predictors of frustration
- highly accurate word recognition is not required for emotion recognition
- raised voice is a predictor for emotion
- hyperarticulation is not a predictor for emotion
- non-native speakers more forgiving of system failures

- hyperarticulation is often a signal of user correction of a system error
- eliminate redundant lexical material: "Did you say you wanted to go to Baltimore?" vs. "Baltimore?" with proper intonation
- information status, theme, topic / comment, focus:
John only introduced **Mary** to Sue.
John only introduced Mary to **Sue**.

Pragmatic Functions of Prosodic Features in Non-Lexical Utterances

- non-lexical items (uh-huh, um, hmmm) convey much by prosody
- back-channels , fillers, disfluency markers
- syllabification: uh (filler, disfluency) vs. uh-huh (backchannel)
- some correlation to usage in Japanese

Turn-taking and Backchannels

By Jeff Heath

Turn-taking



Turn-taking



Turn-taking



Turn-taking



Using silence

Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody. (Ferrer et al., 2002)

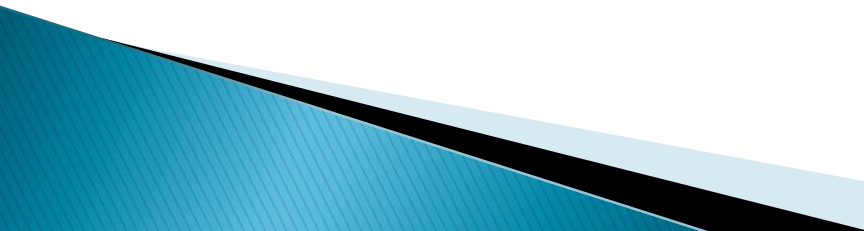
Proceedings of the ICSLP
pp. 2061–2064

Turn-taking

Some signals and rules for taking speaking turns in conversations (Duncan, 1972)

Journal of Personality and Social Psychology
1972, Vol. 23, No. 2, 283–292

Turn-yielding signals

- ▶ Rising or falling pitch
 - ▶ A “drawl” of final or stressed syllable
 - ▶ Termination of any hand gesture
 - ▶ Short phrase that doesn’t add information
 - ▶ Drop in pitch or loudness in paralanguage
 - ▶ Completion of subject–predicate clause
- 

Attempt-suppressing signal

- ▶ Hands engaged in gesticulating

Backchannel signals

Given by auditor (listener):

- ▶ Saying “mm–hmm”, “yeah” or “OK”
- ▶ Nodding head

Turn-taking

Turn-taking cues in task-oriented dialogue
(Gravano and Hirschberg, 2011)

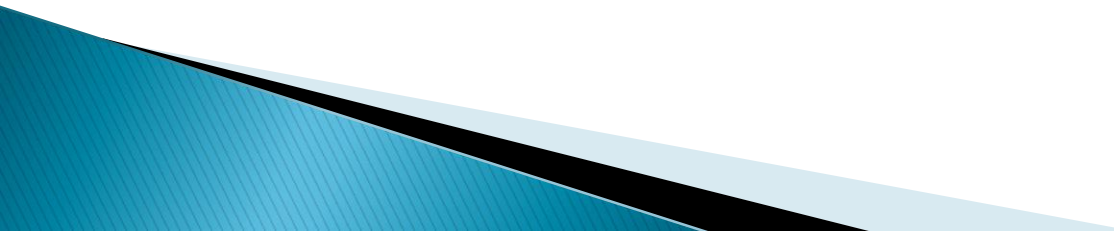
Computer Speech and Language
2011, Vol. 25, No. 3, 601–634

Backchannel

Backchannel–inviting cues in task–oriented
dialogue
(Gravano and Hirschberg, 2009)

Interspeech 2009
pp. 1019–1022

Backchannel–inviting cues

- ▶ Final rising intonation
 - ▶ A higher intensity level
 - ▶ A higher pitch level
 - ▶ A phrase ending in a noun preceded by a determiner, an adjective or a noun
 - ▶ Lower noise–to–harmonics ratio (NHR)
 - ▶ Longer phrase duration
- 

Turn-taking cues

Turn-taking cues in a human tutoring corpus
(Friedberg, 2011)

Proceedings of the ACL 2011 Student Session

pp. 94–98



Turn-taking cues


- ▶ Duration: YIELD's are shorter
 - ▶ Pitch: YIELD's are higher
 - ▶ RMS: energy of YIELD's are lower
- 

Turn-taking cues

Turn-taking cues in task-oriented dialogue
(Gravano and Hirschberg, 2011)

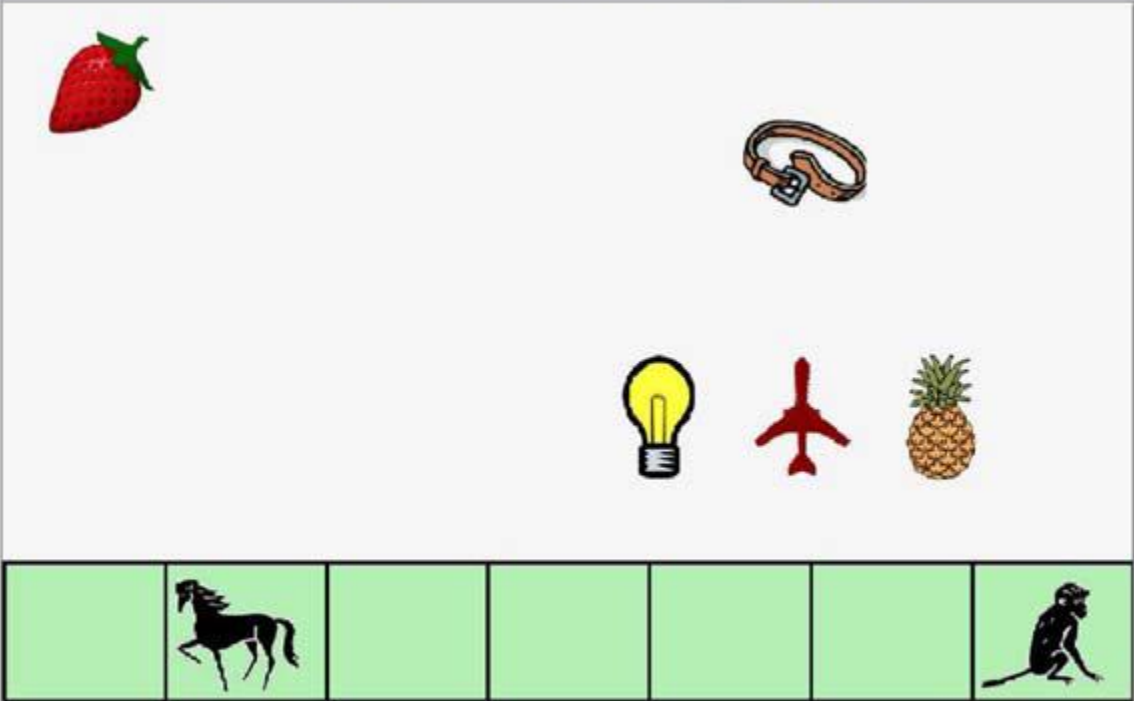
Computer Speech and Language
2011, Vol. 25, No. 3, 601–634

Challenges for IVR systems

- Q1. The system wants to keep the floor: how should it formulate its output to avoid an interruption from the user?
 - Q2. The system wants to keep the floor but to ensure that the user is paying attention: how should it produce output encouraging the user to utter a backchannel?
 - Q3. The system is ready to yield the floor: how should it convey this to the user?
 - Q4. The user is speaking but pauses: how can the system decide whether the user is giving up the turn?
 - Q5. The user is speaking: how does the system decide whether and when to produce a backchannel as positive feedback to the user?
- 

Columbia Games Corpus

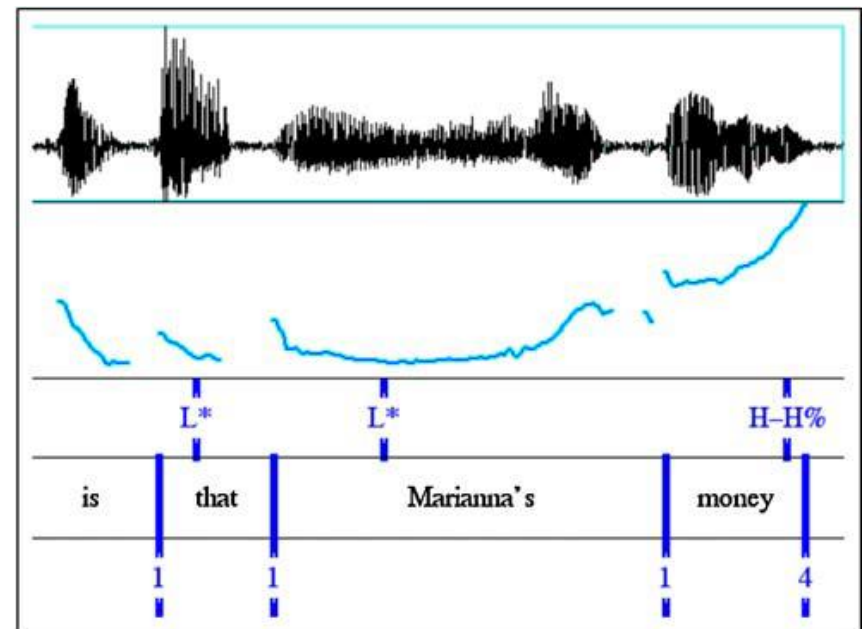
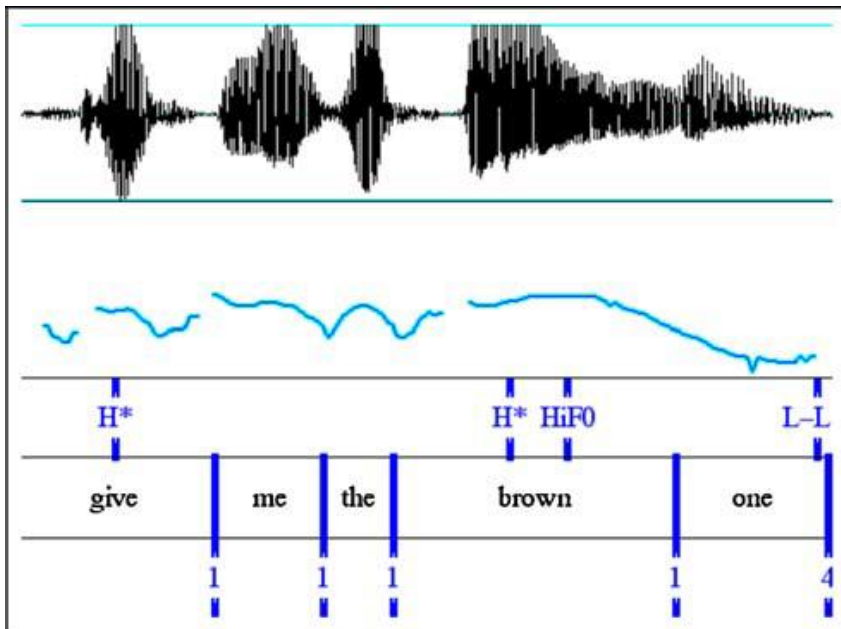
Describe the location of the blinking image.



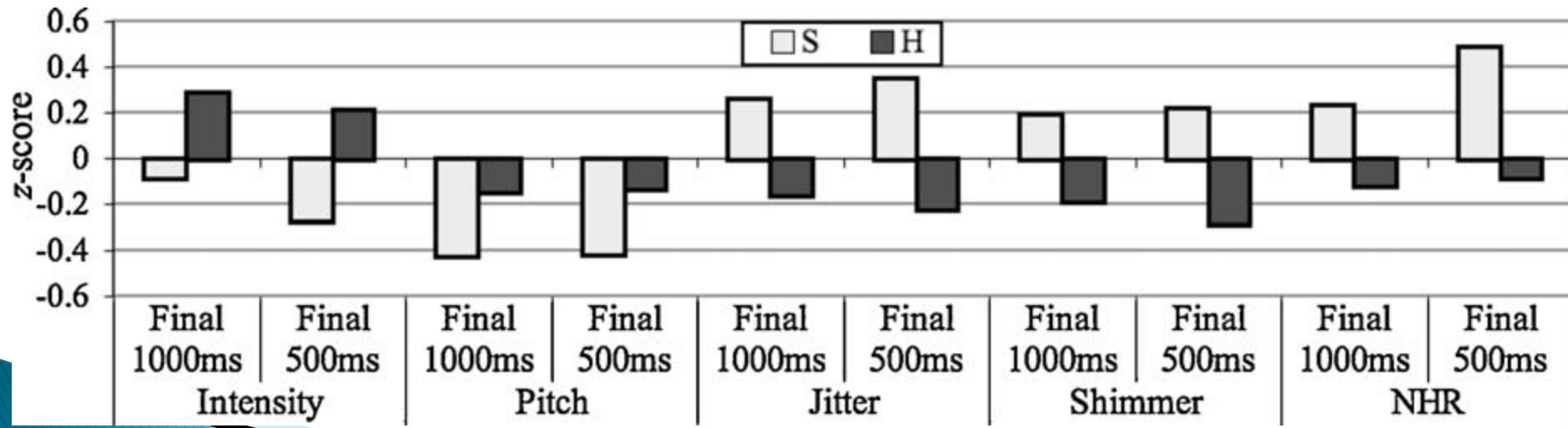
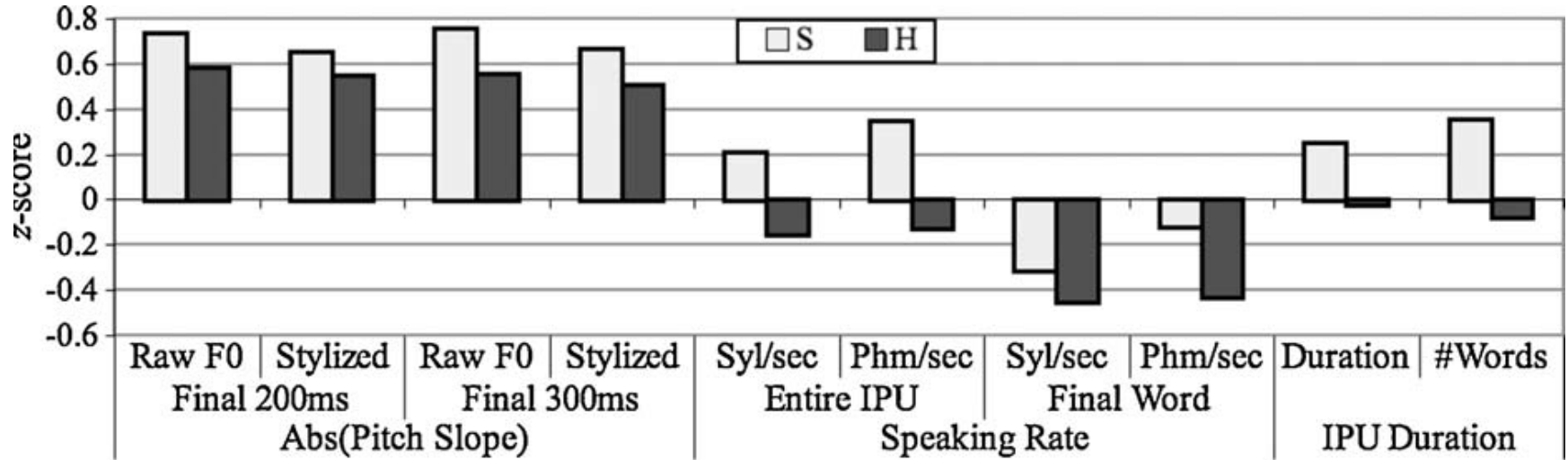
YOUR SCORE: 000
HIGH SCORE: 1322

DONE CONTINUE

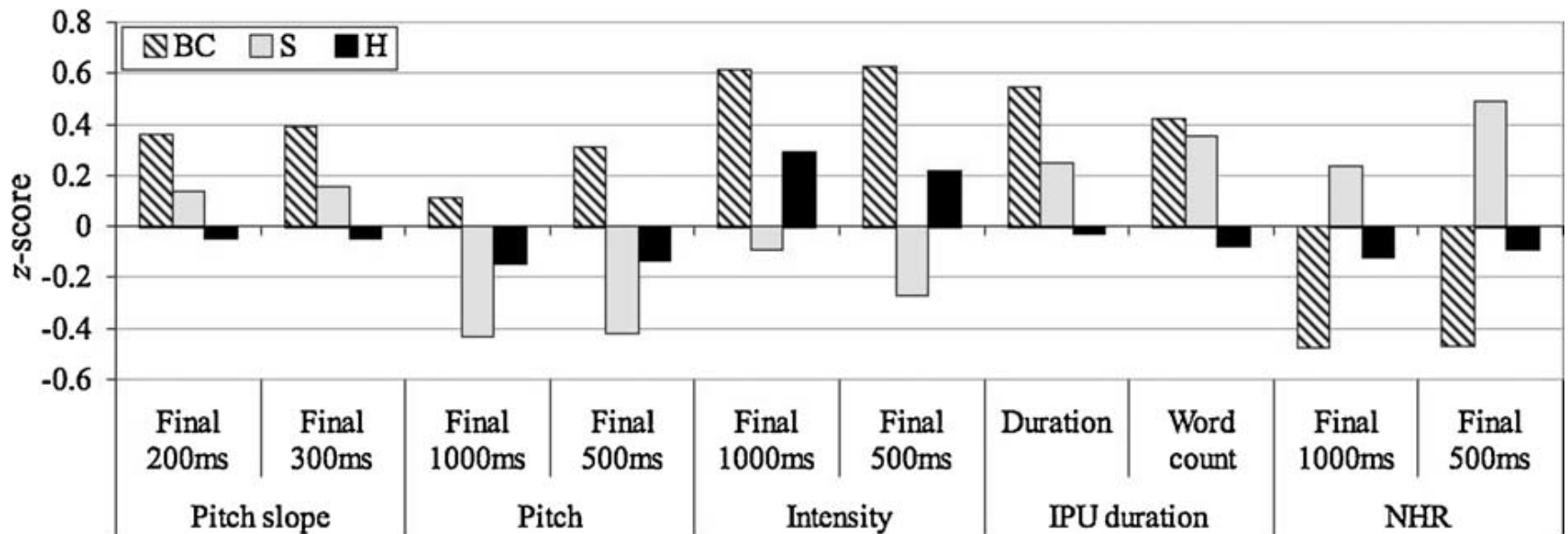
Annotation, feature extraction

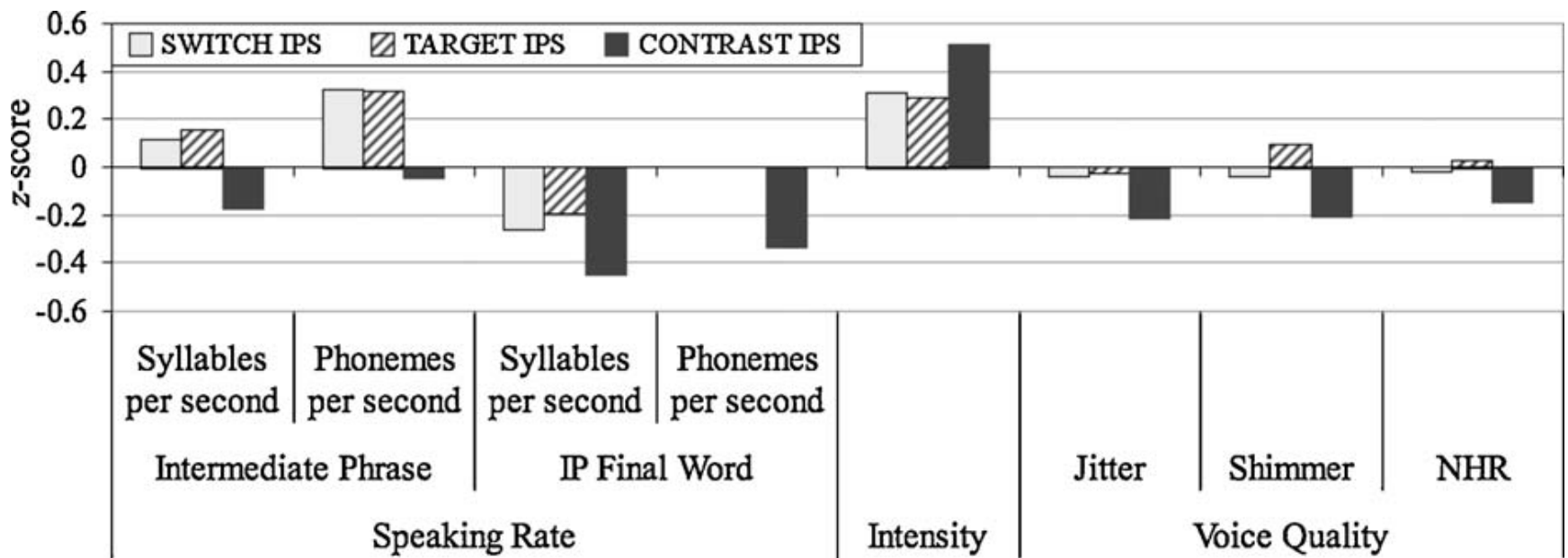


Turn-yielding cues



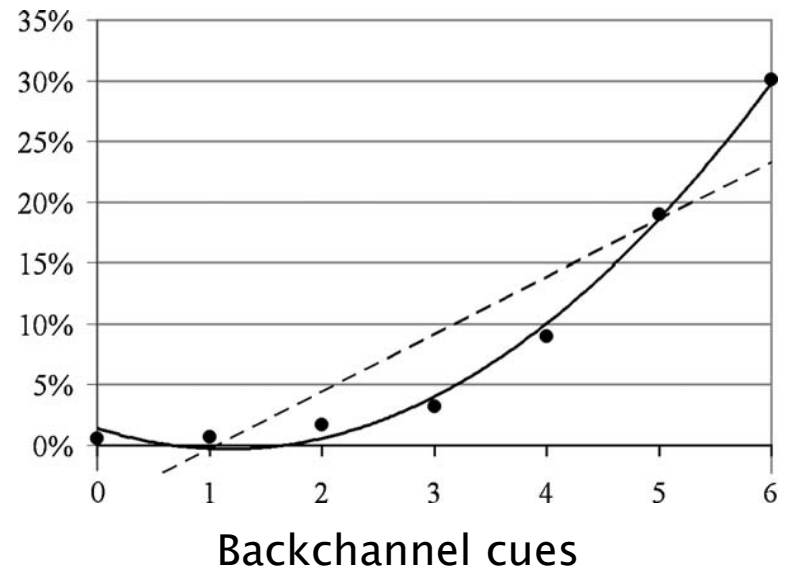
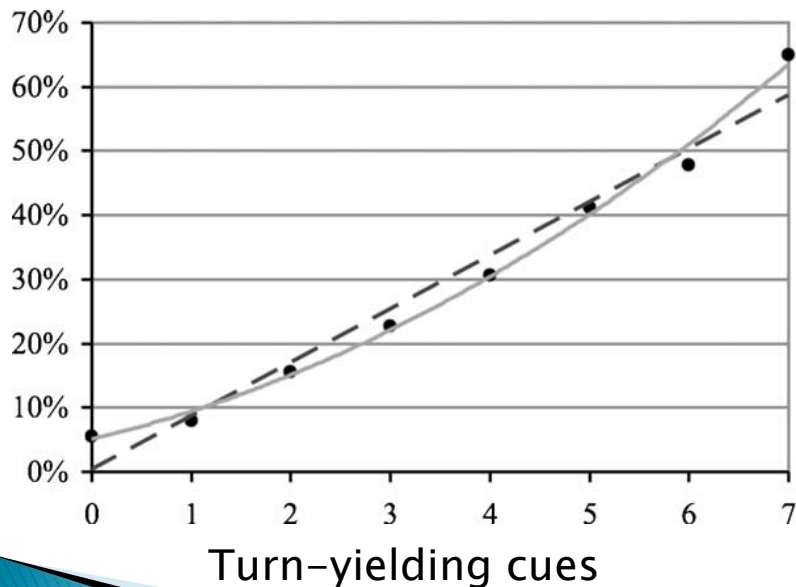
Backchannel-inviting cues



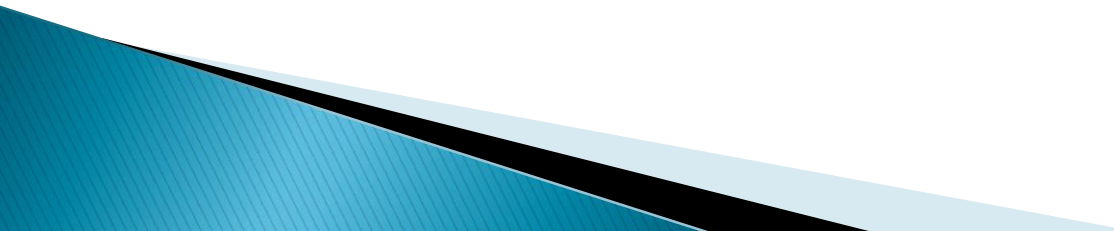


Room for improvement

- ▶ Diversify corpus
- ▶ Less disparaging comments about previous work done
- ▶ Less curve-fitting



Why does it matter?

- ▶ Effective communication – avoid awkwardness and frustration from unnatural pauses
 - ▶ Efficient communication – time is money!
- 

Questions

- ▶ Are computers fast enough to make these calculations in real time, e.g. POS tagging?
- ▶ Could a dialogue system tune to an individual speaker's turn-yielding behavior over time? Do people do this?
- ▶ Does speech synthesis have enough control to produce turn-yielding cues?
- ▶ It seems like semantic cues could be more important than prosodic cues for turn-taking. How long might it be before semantic processing improves to the necessary level?

Multi-party dialog

April 2016
Joanna Church



Roles of agents

- Participant roles: speaker, hearer, social
- Conversational Roles:
 - Active participant
 - Overhearer
 - Uninvolved
- Speaker Identification
 - voice, style, self-identification
 - Microphone array
 - Lips, gestures
-

Addressee recognition

- Volume level
- Router/network
- Direct indication
- Context
- Previous addressee
- Gaze, body orientation
- Attention getting

Interaction

- Turn management (when)
 - More agents competing
 - More actions: assign next turn, request turn
- Channel management (where)
 - Multiple main channels: one per topic/
conversation/set of participants
 - Simultaneous uninterrupted communication

- Thread management (what)
 - Multiple participants allows multiple topics
 - Multiple conversations (might depend on each other)
- Initiative management
 - less symmetric, not equal initiative
 - Leaders develop
 - Cross-initiative



Grounding and Obligation

- Any addressee grounds = optimistic
- Every addressee grounds = unrealistic
- Transfer of obligation

1. **If** utterance specifies addressee (e.g., a vocative or utterance of just a name when not expecting a short answer or clarification of type person)
then Addressee = specified addressee
2. **else if** speaker of current utterance is the same as the speaker of the immediately previous utterance
then Addressee = previous addressee
3. **else if** previous speaker is different from current speaker
then Addressee = previous speaker
4. **else if** unique other conversational participant
then Addressee = participant
5. **else** Addressee unknown

Fig. 1. MRE Agent Speech Addressee Identification Algorithm

Incremental Processing of Dialogue

Eslam Elsayy

Motivation

- Human spoken dialogue is highly interactive
 - Fluent turn-taking with little or no delays
 - Interruptions
 - Different overlapping behaviours
 - backchannels
- Most spoken dialogue systems wait until the user stops speaking before trying to understand and react to what the user is saying.
 - Adequate for system-initiative systems
 - Unnatural and inefficient for mixed initiative dialogue systems
 - like: multiparty negotiation training systems

Solution: Incremental Processing of Dialogue

Goal:

- Make the system able to prepare its action before utterance is complete

Requirements:

- Incremental interpretation of partial utterances
- The ability to predict the final meaning of the utterance

Questions:

- Many utterance partials ? which one to use ?
- How can the system decide that it reached maximum understanding of an on-ongoing utterance ?

Towards Natural Language Understanding of Partial Speech Recognition Results in Dialogue Systems

Kenji Sagae, Gwen Christian, David DeVault, and David
Traum. (2009)
Proceedings of HLT-NAACL

Overview

Contribution:

- They showed that using partial ASR results, relatively high accuracy can be achieved in understanding the meaning of an utterance before it's complete

Domain:

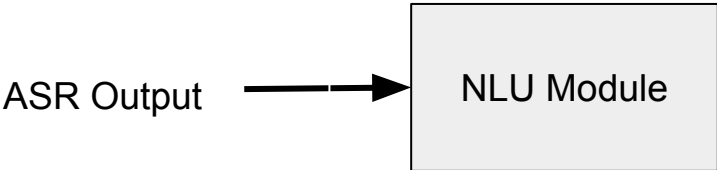
- Negotiation scenario

Dataset:

- Utterances collected from people playing the role of captain



NLU Module

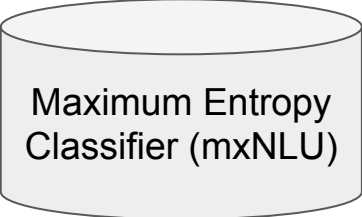


```
[ mood : declarative
  [ type : event
    agent : captain - kirk
    event : deliver
  ]
  sem : [ theme : power - generator
        modal : [ possibility : can ]
        speech - act : [ type : offer ] ] ]
```

AVM utterance representation (Semantic frames)

Features

- Bag of words
- Bigrams
- Pairs of every two words
- Number of words



Training:

- 4500 utterances
- 10 % testing set
- 136 distinct frames
- 10 % deployment set

Evaluation

The goal is to examine two aspects of the NLU:

- **Correctness:** how similar the NLU output with partial utterances is to the gold-standard manual annotation for the entire utterance
- **Stability:** how similar the NLU output with partial partial utterances is to what the NLU result would have been for the entire utterance.

Evaluation Experiment:

- Run audio of all utterances, recording partials of varying lengths for each utterance
- Use partial utterances to train separate models, such that each model is trained with partials of specific length
- Use these models to analyze partial utterances from test set, using F-score as the evaluation metric

Results

- NLU model trained on partial utterances is better than NLU model trained on complete utterances
- Allowing the system to start processing user input when four or five-word partial ASR results are available provides interesting opportunities.

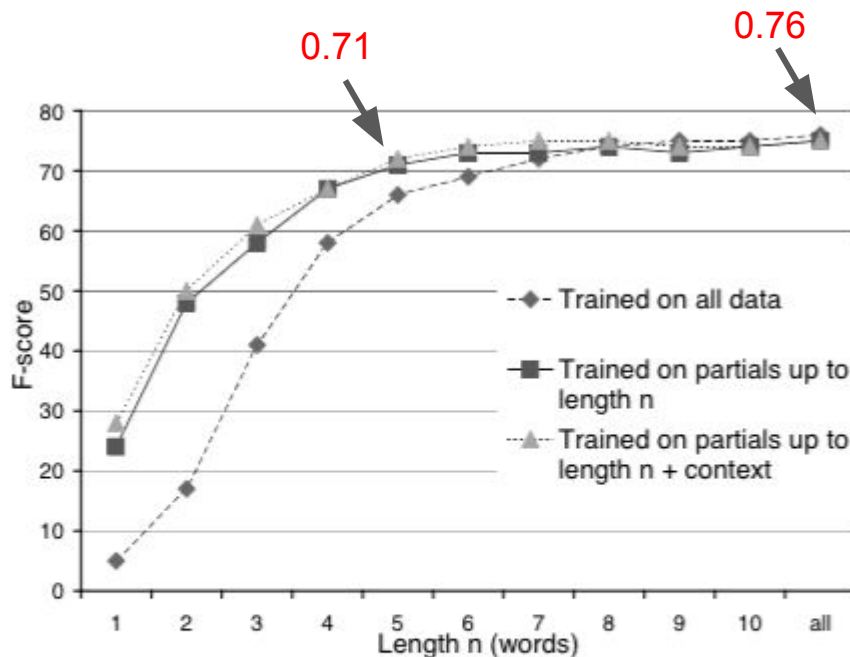


Figure 2: Correctness for three NLU models on partial ASR results up to n words.

Automatic Assessment of Partial Results

Goal:

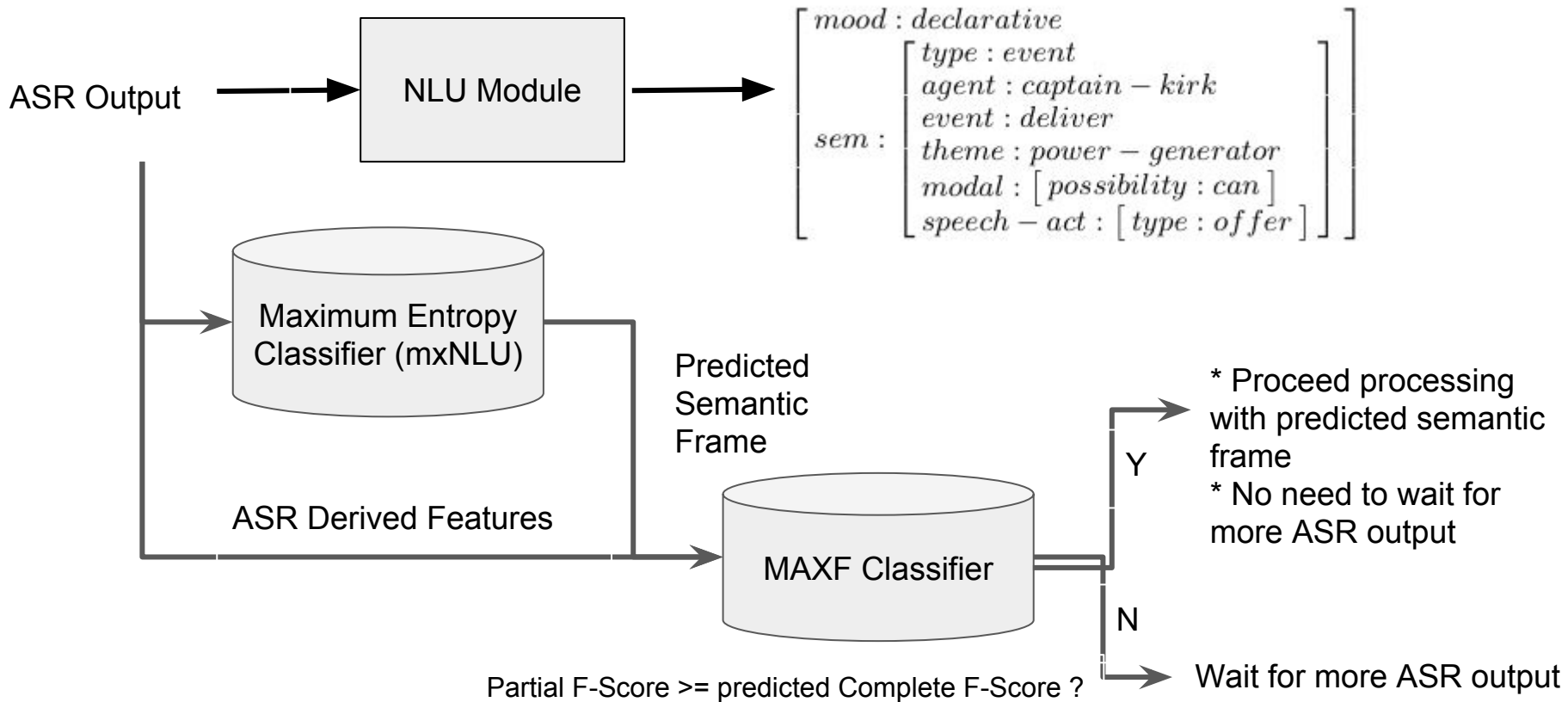
- Give the system the ability to assess whether or not it has already understood the utterance “well enough”, based on the partial ASR results that are currently available

Approach:

- Use a second classifier, MAXF classifier, which uses various features of the ASR result and the current mxNLU output to estimate whether the NLU F-score for the current partial ASR result is at least as high as the mxNLU F-score would be if the agent were to wait for the entire utterance.

MAXF Classifier

AVM utterance representation (Semantic frames)



MAXF Classifier

Features:

- K = the number of partial results that have been received from the ASR
- N = the length (word count) of the current partial ASR result
- The entropy in probability distribution assigned to alternative output frames
- $P(\max)$ = probability of most likely output frame
- NLU = most probable output frame

Target Label: MAXF (GOLD)

Boolean: F score of partial result \geq predicted F-score of final utterance

Training Goal:

Train the MAXF classifier, to predict the value of MAXF (GOLD) as a function of the input features.

Training Procedure:

Decision tree using Weka J48 algorithm, 10-fold cross validation, high precision and low recall

Evaluation Results

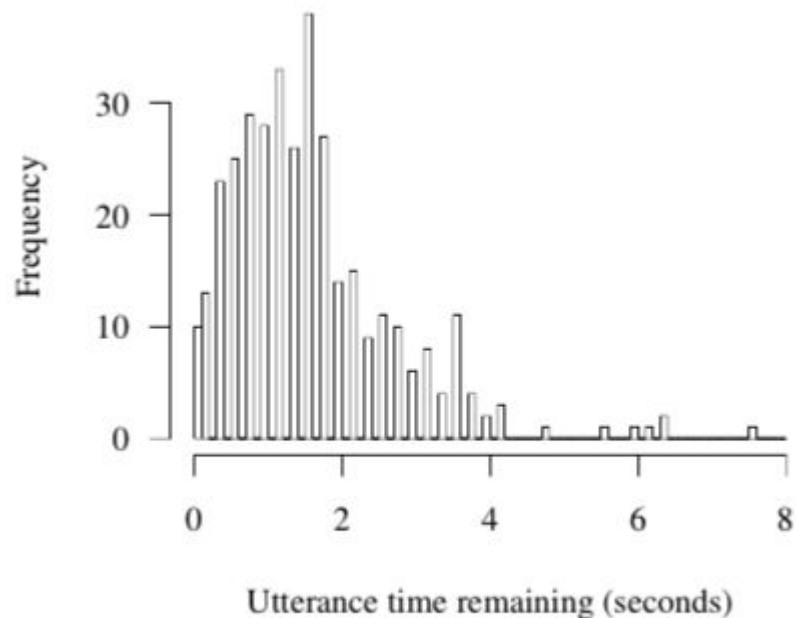


Figure 6: Distribution of $T(K_{\text{MAXF}})$.

$\Delta F(K_{\text{MAXF}})$ range	Percent of utterances
-1	10.67%
$(-1, 0)$	17.13%
0	62.35%
$(0, 1)$	7.30%
1	2.52%
$\text{mean}(\Delta F(K_{\text{MAXF}}))$	-0.1484
$\text{median}(\Delta F(K_{\text{MAXF}}))$	0.0000

Figure 7: The distribution in $\Delta F(K_{\text{MAXF}})$, the “loss” associated with interpreting partial K_{MAXF} rather than K_{final} .

Stability and Accuracy in Incremental Speech Recognition

Ethan Selfridge, Iker Arizmendi, Peter Heeman, and Jason Williams
Proceedings of the 12th Annual SigDial Meeting on Discourse and Dialogue,
Portland, Oregon.

Overview

Contribution

- Shows a method that increases the stability and accuracy of ISR output, without adding delay
- Presents a method for predicting the stability and accuracy of ISR result

Approach: decoding process

Three types of partial results are defined:

- Basic – most likely path through partially decoded Viterbi lattice.
- Terminal – most likely path ends at a terminal node.
- Immortal – all paths come together at a single node

Frequency, Stability and Accuracy Results

Immortal > Terminal > Basic

Immortal < Terminal < Basic

Table 2: Average Number of Partials per utterance

ISR	Group	RLM1	RLM2	SLM
Basic	All	12.0	9.9	11.6
	MW	14.6	12.3	29.7
Terminal	All	5.4	3.3	6.2
	MW	6.4	4.1	8.8
Immortal	All	0.22	0.32	0.55
	MW	0.42	0.67	0.63

Table 3: Stability and Accuracy Percentages

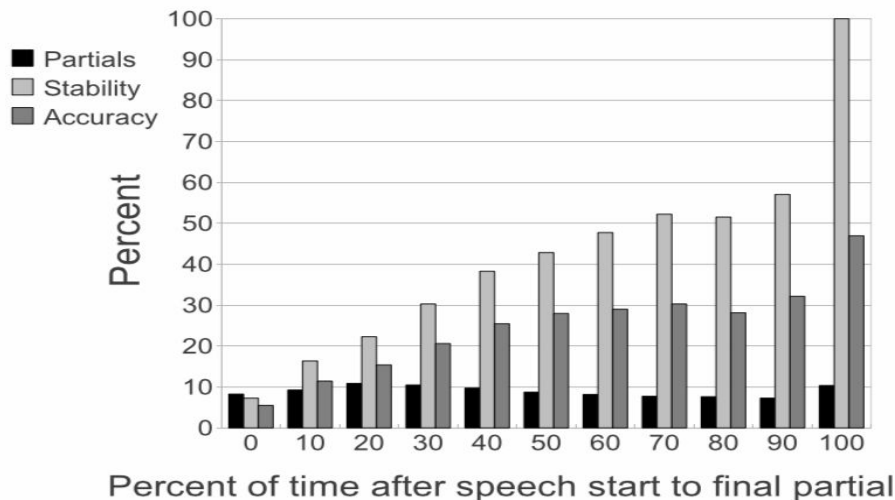
ISR	Group	RLM1	RLM2	SLM
Stability				
Basic	All	10 %	11 %	7 %
	MW	14 %	15 %	9 %
Terminal	All	23 %	31 %	37 %
	MW	20 %	28 %	36 %
Accuracy				
Basic	All	9 %	1 %	5 %
	MW	11 %	13 %	6 %
Terminal	All	13 %	21 %	24 %
	MW	12 %	17 %	21 %
Immortal	All	91 %	93 %	55 %
	MW	90 %	90 %	56 %

Hybrid Approach: LAISR

Lattice-Aware Incremental Speech Recognition

- Recognizes both Terminal and Immortal results; checks for Immortal result first, then backs off to Terminal result.
- Produces a steady stream of partials with better stability and accuracy.

Partials per Utterance			
	RLM1	RLM2	SLM
All	5.6	3.5	6.7
MW	6.7	4.5	9.6
Stability Percentage			
All	24 %	33 %	40 %
MW	24 %	35 %	41 %
Accuracy Percentage			
All	15 %	23 %	26 %
MW	16 %	22 %	24 %



Takeaways

- Incremental processing of dialog is essential to replicate many of the human dialog behaviours
- Incremental processing needs prediction of accuracy and stability of partials while the utterance is still progressing
- Prediction gives the system the ability to assess the strategic points of time where it can proceed using the partials

Discussion Point:

Should the system generate overlapping behaviour or interrupt at every opportunity ?

Thanks!

References

- Kenji Sagae, Gwen Christian, David DeVault, and David Traum. (2009). Towards Natural Language Understanding of Partial Speech Recognition Results in Dialogue Systems. In Proceedings of HLT-NAACL.
- David DeVault, Kenji Sagae, and David Traum. (2009).
- Can I finish? Learning when to respond to incremental interpretation results in interactive dialogue. In The 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009), London, UK.
- Ethan Selfridge, Iker Arizmendi, Peter Heeman, and Jason Williams. (2011). Stability and Accuracy in Incremental Speech Recognition. In Proceedings of the 12th Annual SigDial Meeting on Discourse and Dialogue, Portland, Oregon.