You Talking to Me?

Speech-based and multimodal approaches for human versus computer addressee detection

Andreas Stolcke
Speech & Dialog Research Group
Microsoft AI & Research
Mountain View, CA

With apologies to many …

- "`You Talking to Me?' Exploring Voice in Self-Service User Interfaces", Johnson & Coventry, *Intl. J. Human-Computer Interaction*, 13(2), 2001.
- "Are you talking to me? Dialogue systems supporting mixed teams of humans and robots", Dowding et al., AAAI Fall Symposium, 2006
- "You Talking to Me? A Corpus and Algorithm for Conversation Disentanglement", Eisner & Charniak, ACL 2008
- …

Outline

- Motivation & Task
- Prior work
- Speech-based AD
  - Lexical modeling
  - Stylistic/prosodic modeling
  - On-line/incremental processing
  - Portability across corpora
  - NN-based lexical modeling
- Multi-modal AD
- Conclusions & Outlook

Motivation

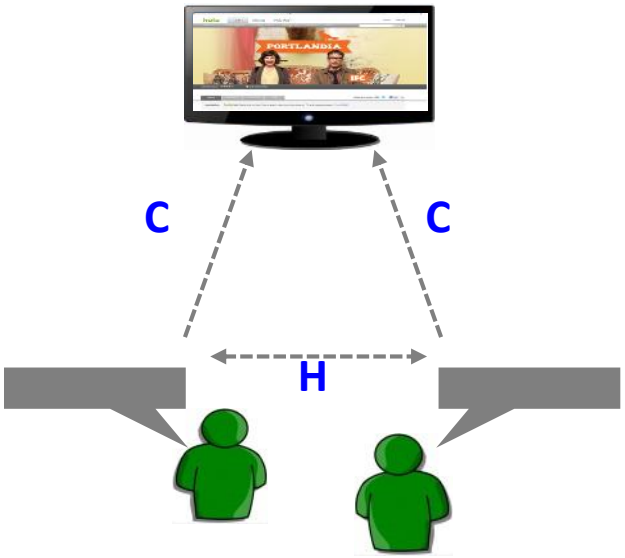
Without
- "Alexa"
- "OK Google"
- "Hey Cortana"

Motivation

- Ubiquity and naturalness of human-computer dialog systems creates new problems
  - Systems can interact with multiple users, or in environments where other speakers are present
  - Input is free-form and conversational, similar to how we talk to fellow humans
- New problem: *How to tell what speech is meant for the system?*
  - Interpreting human-directed speech is really bad
  - Other features like gaze are not always available and may not help when users look at the screen while conversing

Addressee Detection Task

- Test unit = audio segment, always human-produced
  - A.k.a. “clip” “utterance”
- Binary task: Classify each speech segment as H or C
  - C = addressed to computer
  - H = addressed to another human



Prior Work

- Most AD work focused on H-H conversation and meetings, and multimodality (AMI project, op den Akker et al. 2009)
- Prior H-C work
  - Usually single H
  - Utterances to the system are not open-ended; are often commands
- Little work on H-H-C in situated dialog (except: Bohus & Horvitz, SIGDIAL'11)
- **Common approach: utterance rejection based on poor recognition (Dowding et al., AAAI'00) or failure to interpret (Paek et al., ICSLP'00)**
- Prosody features
  - Utterance-level pitch/energy statistics (e.g., Reich et al., Interspeech '11)
  - Not "online": features normalized by overall statistics for session, speaker
- Some work on gaze (Katzenmaier et al., ICMI'04; Disney Labs)
- Some work on perception: e.g., humans have trouble using only lexical features (Lunsford & Oviatt)

Data: Conversational Browser

(Heck et al., Multimodal Conversational Search and Browse, IEEE SLAM Workshop, 2013)

CB multi-human corpus:

- 6.3 hours, 38 sessions, 36 unique speakers
- Hand-transcribed, annotated for addressee & commands
- 22 shortest sessions used for testing (maximize number of speakers); empty and unintelligible utterances removed
- Full sessions used as units in testing, jack-knifing

CB Corpus Statistics

- From **CB** prototype; **H = human-directed, C = computer-directed**

Statistic	Training Set	Test Set
Utterances	2577	2889
Recognized words	7026	7874
H utterances	40.8%	31.0%
C utterances, noncommand	31.9%	32.8%
C utterances, command	24.5%	32.0%
Mixed H/C *(grouped with C)*	3.7%	4.2%

Utterance Class	Example
H	Want to watch a movie?
C-noncommand	Show movies that are under two hours.
C-command	Scroll down. Go back.
Mixed H/C	Did you see this already … Stop listening!

Lexical Modeling

- *What* people say correlates with addressee
  - Commands, domain-specific words, etc.
- Features = word N-grams
  - Capture vocabulary, grammar differences
  - Unigrams, bigrams, trigrams (with <s>, </s>)
  - Errorful, based on ASR (WER = 19%, SER = 28%)
  - Control experiment: human (reference) transcripts

Lexical Likelihood Ratio Scoring

- Train class-specific (C vs. H) language models
  - Maximum entropy trigram LMs
- Detection scores: length-normalized log likelihood ratios

$$S = \frac{1}{L} \log \frac{P(W \mid M_C)}{P(W \mid M_H)}$$

- W are recognized words
- P is LM joint utterance probability, L is number of words

Stylistic-Prosodic Modeling

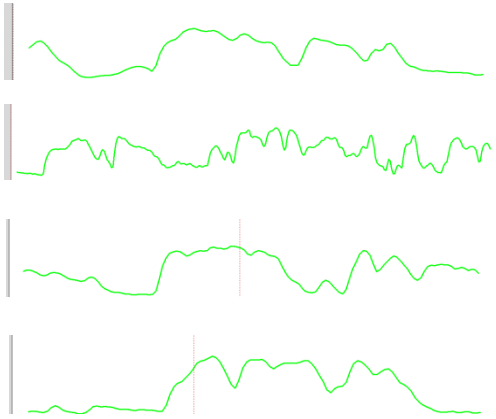
(Shriberg et al., 2012-2013)

- *How* people speak also correlates with addressee
  - Rhythmicity (temporal properties)
  - Vocal effort: raised voice (spectral properties)
  - H-C speech is more rhythmic & louder
- Desiderata for prosodic features:
  - Word independent: no ASR needed.
  - Context-independent: use only current segment. No system state, no information from other segments, no session-level normalization.
  - Speaker-independent: no speaker normalization or modeling
  - Enable early (e.g., client-based) processing
- Pitch-based features are problematic
  - Signal quality makes it hard to track F0 robustly
  - Pitch models usually require speaker-level normalization

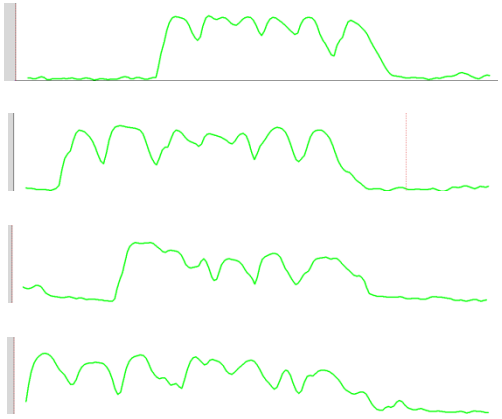
Style: Rhythmicity

Intensity Contours (Praat)
Contours are mean-normalized within utterance

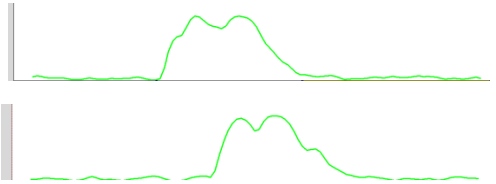
Human



Computer, noncommand



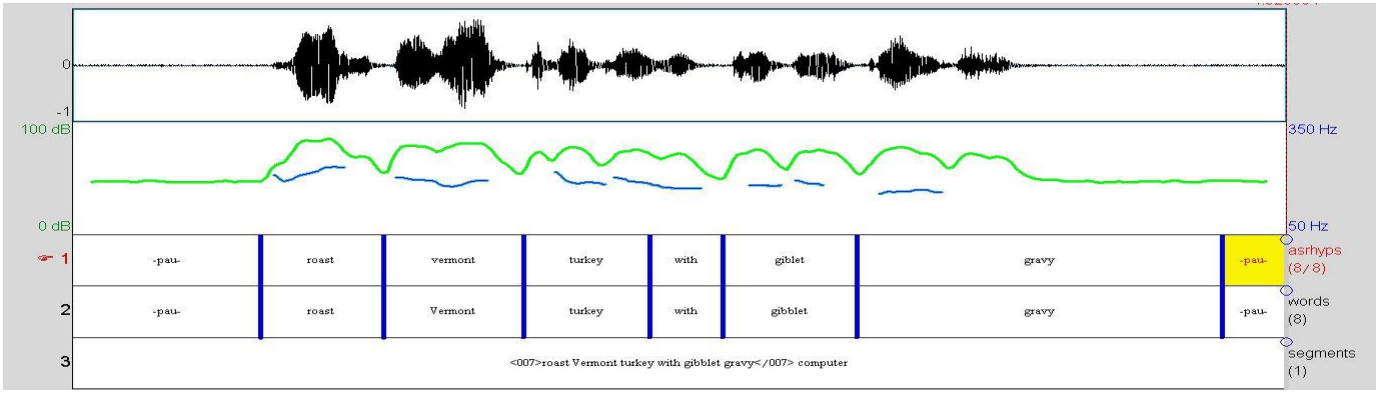
Computer, command

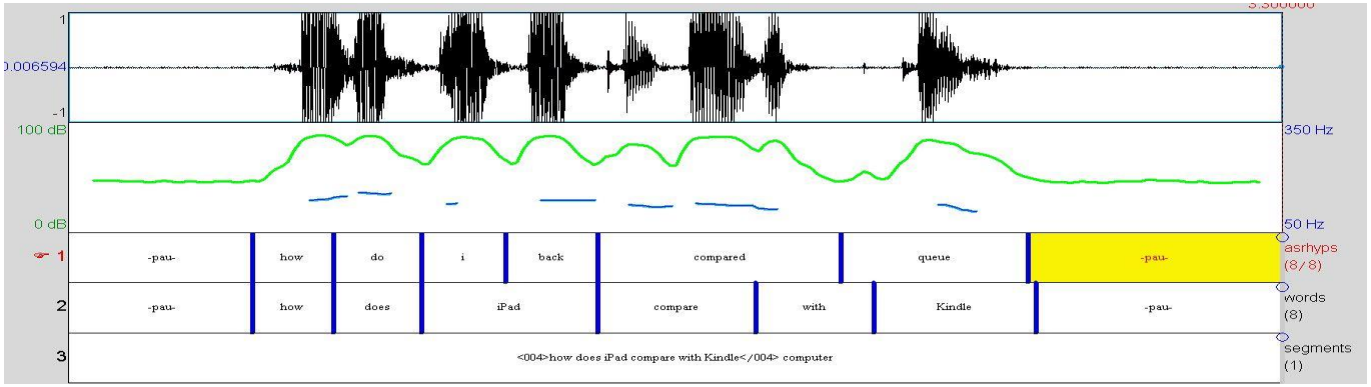


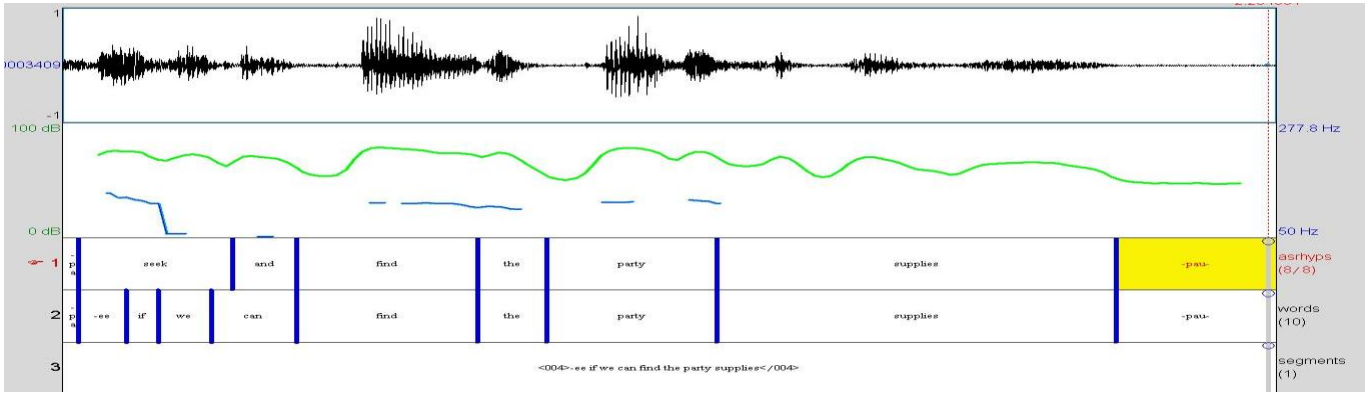
Computer-addressed is overall more "rhythmic"

- More regularly spaced peaks
- Rounder peaks (slower rate)
- Higher and more regular peak line and valley line
  → valleys not far from peaks

Sample Utterances







Modeling Rhythmicity

- Energy contour GMM
  - **Captures shape of energy contour** using a signal processing approach
  - Slide **fixed-length window** over segment and compute **DCT in temporal domain**
  - Use c0 and c1 output from standard MFCC analysis
  - 200 ms window, 50% shift, utterance-level mean-subtraction
  - First 5 DCT coefficients for c0 contour, 2 coefficients for c1 contour
  - Model 7-dim feature vector by GMMs (20 gaussians, full covariances)
  - Best using whole utterance, including nonspeech regions
- Energy contour boosting model
  - Extract **utterance-level statistics** of same 7 DCT features
  - Mean/min/max/variance
  - Feed into boosting classifier (ICSIboost)

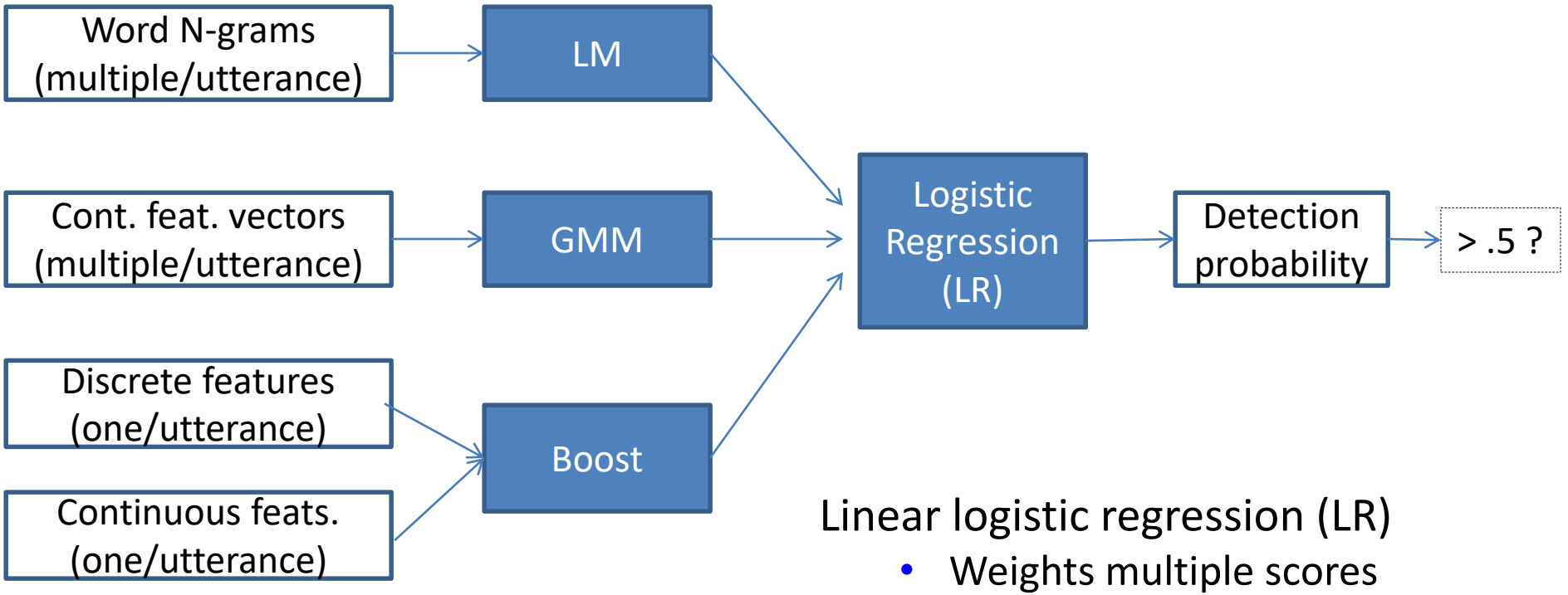
Prosodic Likelihood Ratio Scoring

- Detection scores for GMM-based systems are length-normalized log likelihood ratios

$$S = \frac{1}{L} \log \frac{P(X \mid M_C)}{P(X \mid M_H)}$$

- X are features for all frames in utterance
- P is GMM likelihood, L is number of frames

Modeling Overview



Linear logistic regression (LR)

- Weights multiple scores
- Calibrates for prior probabilities

$$p = \frac{1}{1 + e^{-(a_0 + \sum_i a_i x_i)}}$$

Features and Models

(Shriberg, Stolcke, Ravuri, 2013)

Model types: ngrams **-ng** , Gaussian **-g**, boosting **-b**

- Lexical
  - **asr-ng**: recognized words, modeled by trigram LMs
  - **ref-ng**: reference words, modeled by trigram LMs – **for comparison only**
- Prosodic-stylistic (rhythmicity)
  - **encon-g**: DCT of mean-normed energy in 200ms window, modeled by GMMs
  - **encon-b**: same, but using boosting classifier
- Prosodic-stylistic (vocal effort)
  - **tilt-b**: spectral tilt measures; boosting classifier
  - **devo-b**: delta energy at voicing transitions, boosting classifier
  - **tilt-g, devo-g**: same, using GMMs
- Standard features (used in prior work)
  - utterance length, speech and voicing pattern statistics, utterance-level energy statistics, energy in voiced regions, speaking rate measures

Metrics

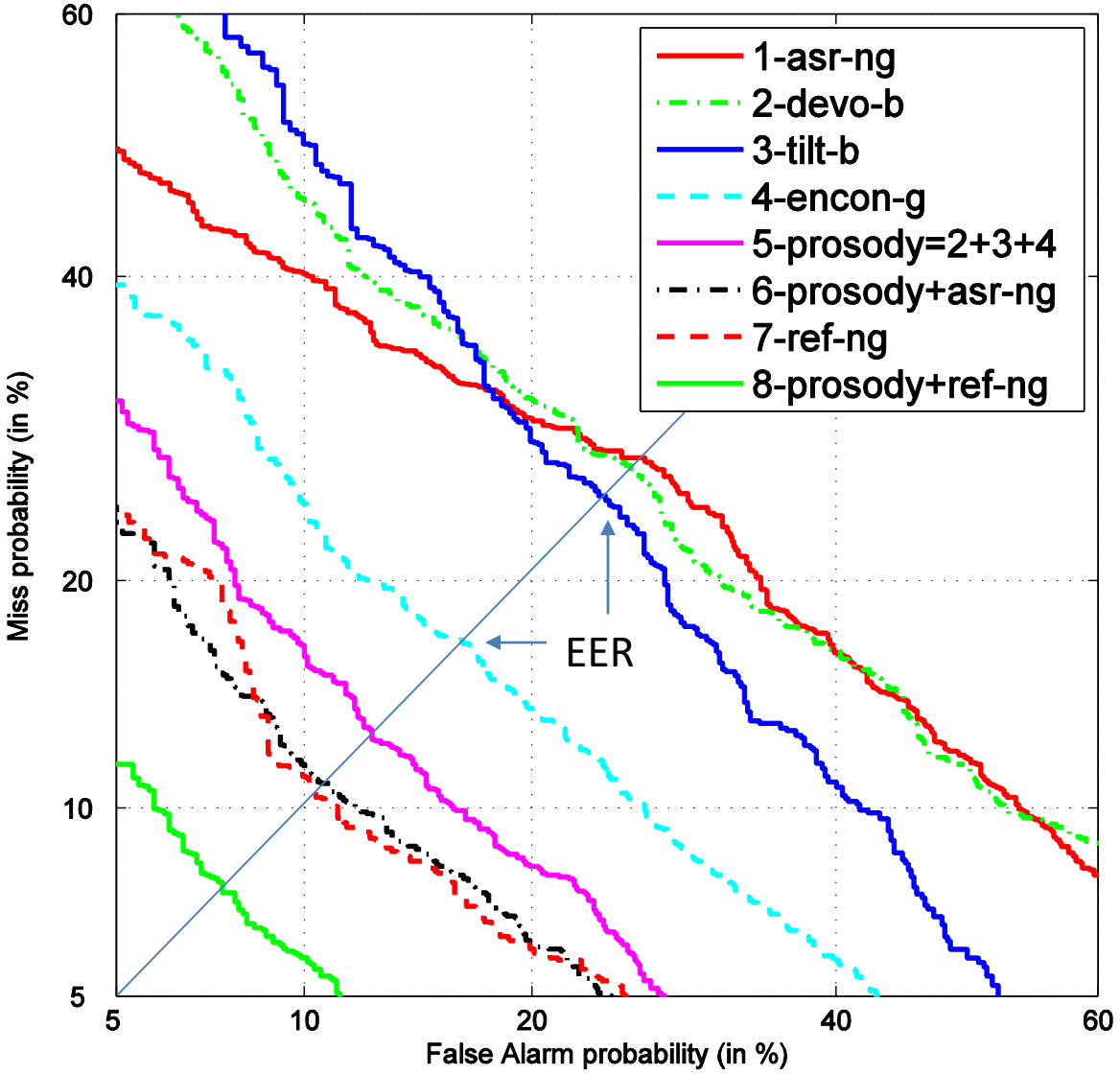
- Evaluate AD as a detection task
  - Detect speech H-C utterances / ignore H-H utterances
- Equal error rate (EER)
  - Error rate when decision threshold chosen so that P(false alarm) = P(miss)
  - Independent of class priors
  - Chance performance = 50%
- Decision Error Tradeoff (DET) plot
  - False alarm vs. miss probability on normal deviate scale

EER Performance

	Model	EER %
Lexical	asr-ng	27.0
	ref-ng	*10.4*
Prosody - Rhythmicity	encon-g	16.8
	encon-b	17.8
	encon-(b+g)	15.4
Prosody – Vocal effort	devo-b	26.2
	tilt-(b+g)	21.7
Combined	All prosody	12.5
	All prosody + standard features	12.6
	All prosody + asr-ng	10.9
	All prosody + ref-ng	*7.5*

DET Plot

for different feature types and their combinations



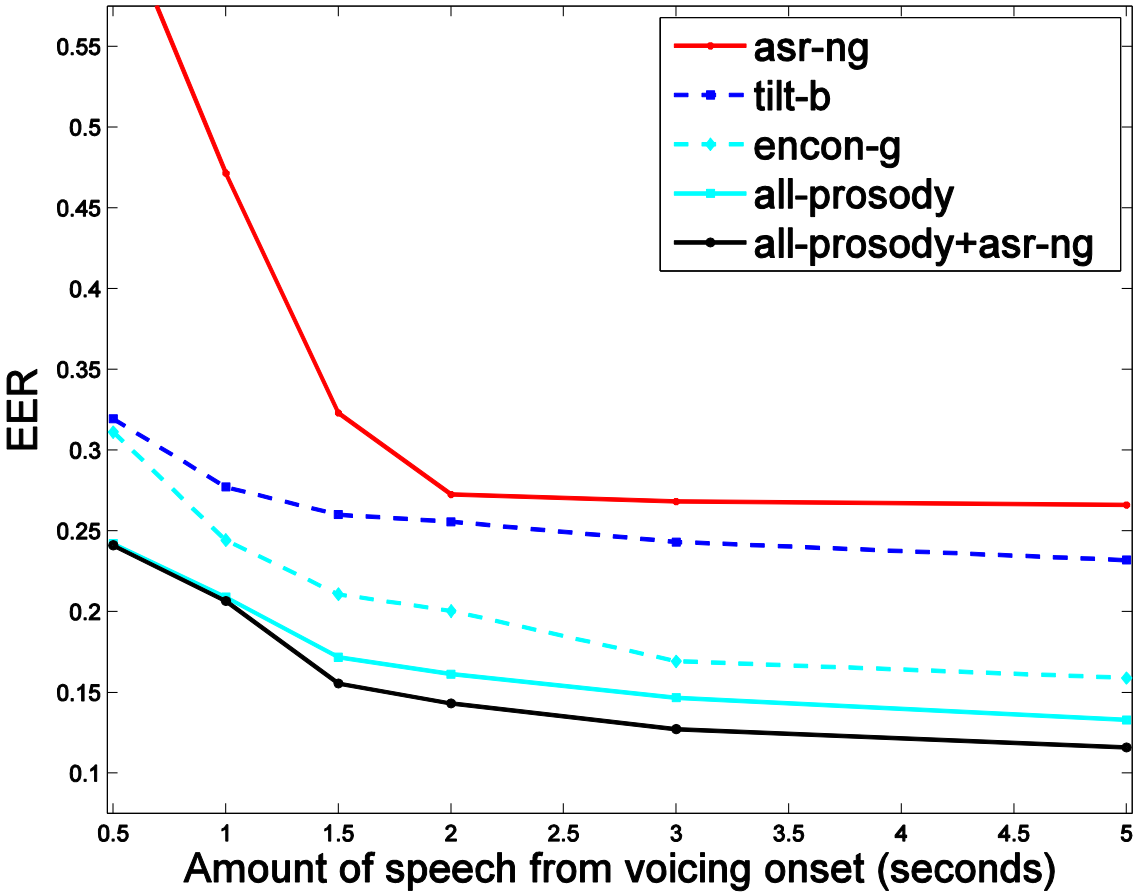
Conclusions (so far)

- Energy contours capture rhythmicity of C-directed speech
  - Single best source of information
- Spectral modeling (capturing vocal effort) are not as effective, but complement temporal modeling
- Combined prosodic features are better than ASR-based lexical features
- Big gains from lexical + prosodic combination
  - even if words were perfectly recognized
- Standard features used in prior work did not add to those prosodic models

Online Addressee Detection

How early in an utterance can we decide on addressee?

- Equal error rate as a function of duration of speech from first voicing
- All prosodic systems show earlier detection than lexical system
- Most cues available by 2 seconds – partly due to presence of short commands



Domain Generalization

(Lee, Stolcke & Shriberg 2013; Shriberg, Stolcke & Ravuri 2013)

- Collecting H-H-C data is expensive – Do we need to do it for each new dialog domain?
- Do people use similar lexical and stylistic features for addressee classes across corpora?
- If so, can H-H-C addressee models be trained on
  - data from other domains?
  - data from H-C and H-H scenarios?

Generalization Experiment

- Test data is always **in-domain**
  - CB - Conversational Browser
  - Multi-user (H-H-C)
- Training data is **out-of-domain**
  - For H-C utterance models: CB single-user, DARPA Communicator, ATIS
  - For H-H utterance models: Fisher telephone conversations, ICSI Meetings
- Best prosody model: ATIS (H-C) vs. ICSI (H-H)
- Best lexical model: CB *single*-user (H-C) vs. ICSI (H-H)

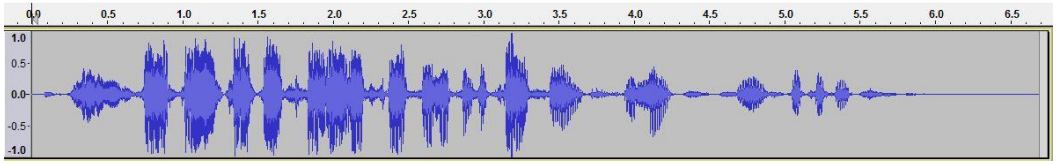
Generalization Results (% EER)

System	Train on In-Domain	Train on Best Out-of-Domain
encon-g	16.8	22.8
tilt-b	24.5	40.6
encon-g + tilt-b	14.0	20.3
asr-ng	27.0	26.4
prosody + asr-ng	10.9	15.9

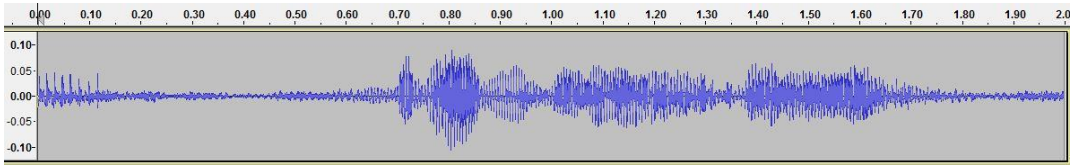
- Degradation relative to in-domain training is about 35%
- Best generalization from rhythmicity features; spectral capture of raised voice less robust
- Still, prosodic features combine with each other when trained on outside data
- Prosodic and lexical models also combine effectively with outside data training

But does it work on Real Data?

- Cortana, Windows Phone
  - C-directed, sample 1
  - C-directed, sample 2



  - H-directed



- ~ 4% human-directed utterances (“off-talk”)
- 2 ASR confidence scores (on-device, on server)

Cortana Results

Model	EER (%)
ASR 4grams	29.6
Energy contours	21.9
ASR 4grams + Energy contours	18.7
ASR confidence	13.0
ASR conf. + 4grams + Energy contours	9.2

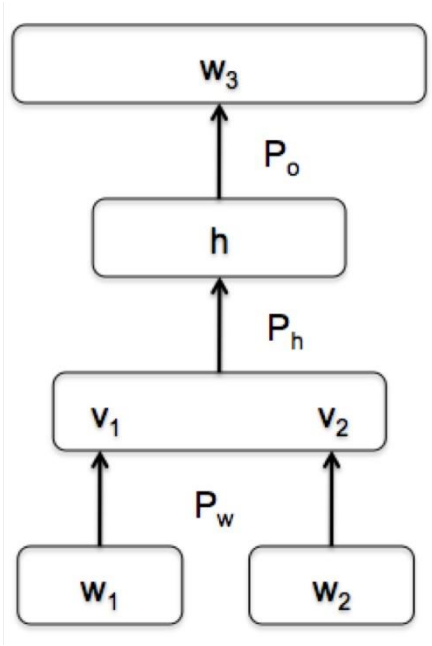
- ASR confidence very effective
- Prosody and ngrams still complementary
- ASRconf + prosody + lexical highly effective

Lexical AD Modeling with Neural Nets

(Ravuri & Stolcke, 2014-2015)

- Improve lexical generalization through continuous vector space embedding
- Feedforward neural net LMs (NN-LM) [Bengio]
  - Feed-forward net
  - Predicts next word or utterance class from N-1 previous word vectors
  - Learns continuous vector space word representations
- Recurrent neural net LM (RNN-LM) [Mikolov]
  - Predicts next work from hidden state vector
  - Modified to predict utterance class (averaged over all words)
  - Potentially infinite history
- Recurrent Long-short-term Memory (LSTM) [Hochreiter & Schmidhuber]
  - Gating mechanism to control forgetting

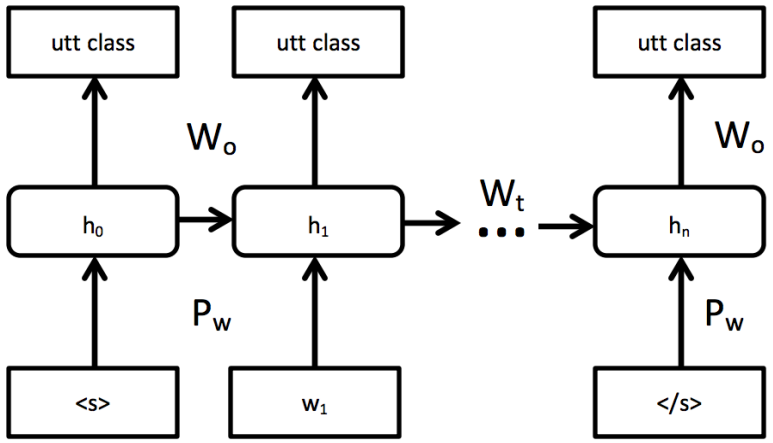
NN Models (1)



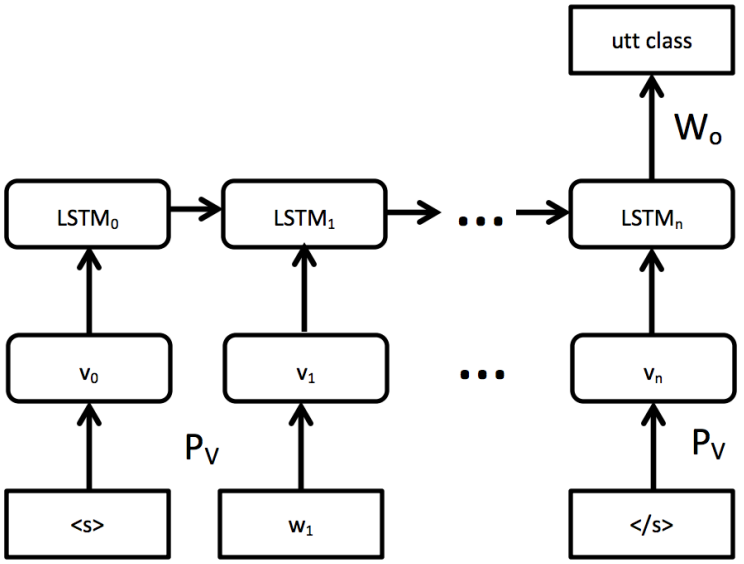
- **NNLM-ngram**: predict words
  - Drop-in for class ngram LMs
  - Form likelihood ratios as before
- **NNLM-addressee**: predict utterance class
  - Average predictions
  - Trained to discriminate

NN Models (2)

- RNN
  - Predict utterance class *at each word*
  - Average predictions



- LSTM
  - Predicts utterance class *at end of utterance*

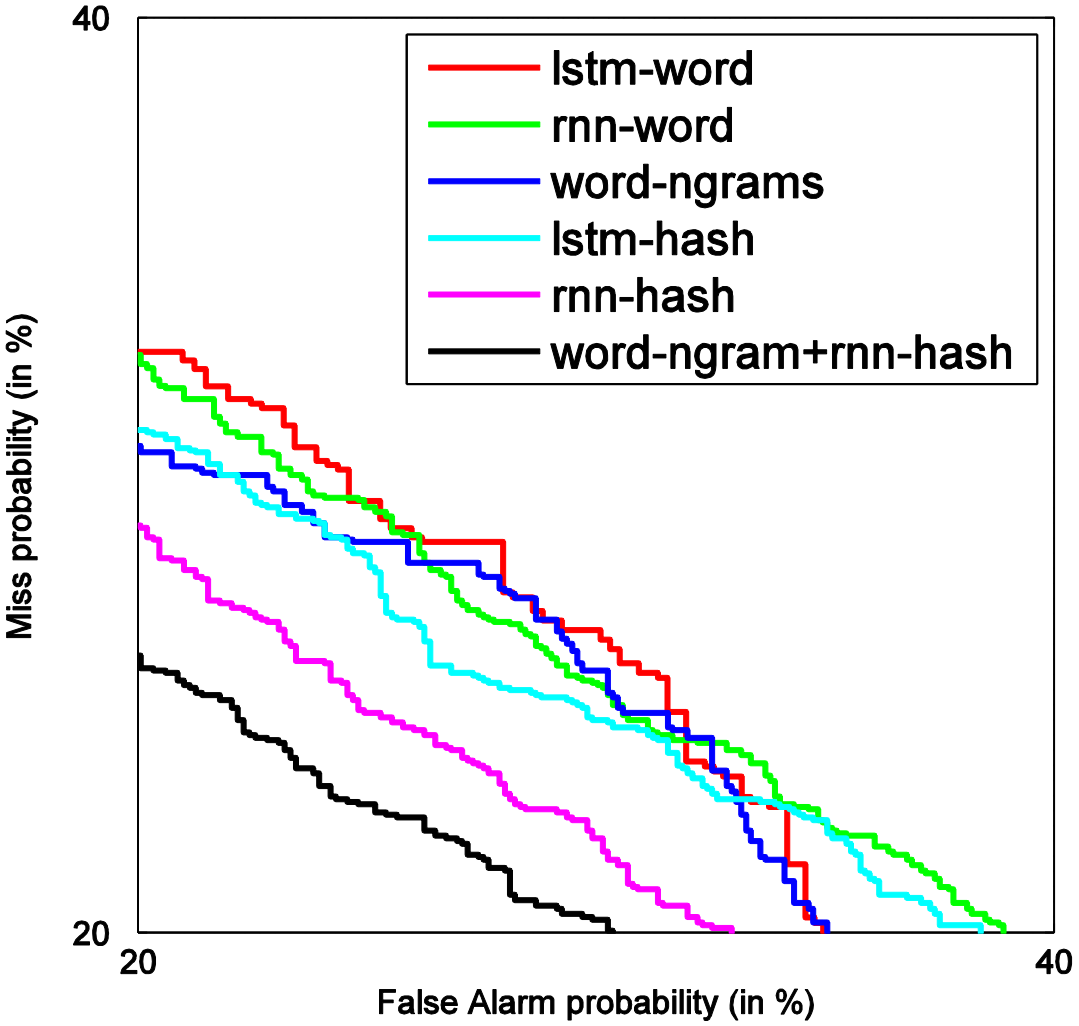


NN-based Lexical AD: Results

- Unary vs. character ngram hash word encoding
- LR combination with word ngram model

Model	Encoding	EER %	EER% w/word-ngram
Word-ngram	word	**27.0**	
NNLM-addressee	word	30.0	
NNLM-ngram	word	29.5	
RNN	word	26.4	24.7
RNN	char	**24.3**	**23.1**
LSTM	word	27.2	25.7
LSTM	char	26.7	25.3

NN Lexical AD: DET Plot



- 14% relative EER reduction over word ngram

Multimodal Addressee Detection

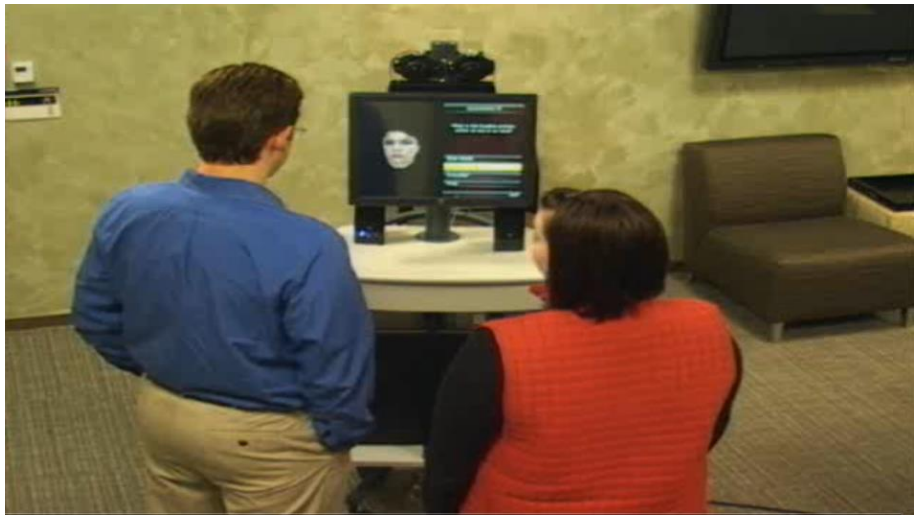
(Tsai, Stolcke & Slaney, 2015)

- So far only using information derived from audio
  - Acoustic features
  - Automatic speech recognition
- … and only information contained in utterance
- What can other modalities, information sources contribute to AD?
  - Visual
  - Spatial
  - Dialog system state

Monaco Dataset

(D. Bohus & E. Horvitz, Facilitating Multiparty Dialog with Gaze, Gesture, and Speech, ICMI 2010)

- 2-3 participants playing trivia game with agent
- Data available:
  - Audio, video
  - Microphone array beam
  - Dialog system state
  - Annotations
  - ASR, face detection, face pose estimation



- 2001 training, 1952 test utterances

Modalities and Features

Modality	Family	Intuition	#Features
Acoustic	Energy	People tend to speak more loudly to C	21
	Eg. change	People pause while waiting for C response	24
	Eg. contour	People speak more slowly and rhythmically to C	2
Visual	Movement	People more stationary when interacting with C	12
	Face angle	Where a person is looking predicts addressee	11
	Distance b/w speakers	Social signal indicating comfort level, distant means less likely to have discussions	18
System	Various	System behavior predicts user behavior	6
Beam	Various	Spread indicates level of discussion among actors	16
Reco	N-grams	Speakers say different things to H vs. C	2
	Various	Recognizer performs better on C-directed speech	5

Features in Detail

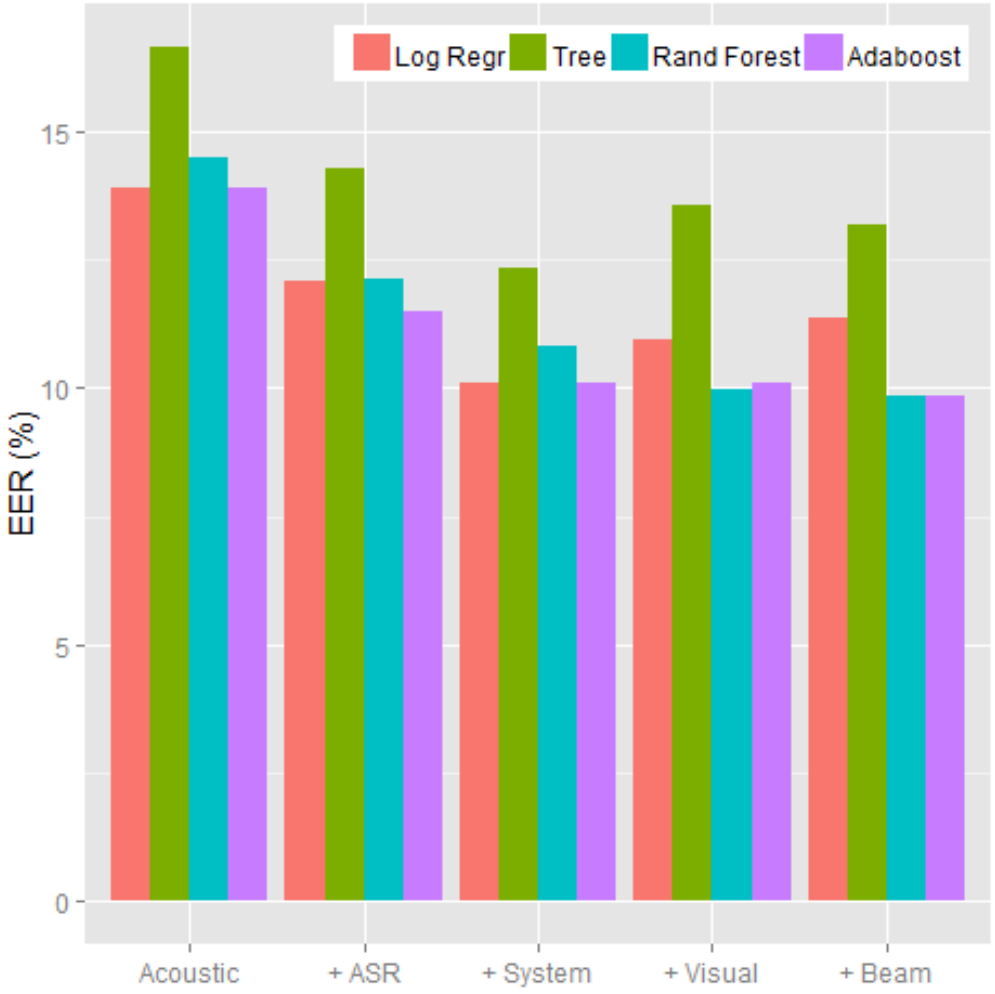
- Acoustic
  - Energy, energy change, energy contour (encon – see earlier work)
- Visual
  - Movement
  - Face pose
  - Distance between participants
- System
  - Number of participants
  - Text prompt?
  - Time elapsed since agent speech
  - Content type of agent speech
- Beam
  - Spread of beam angles
- Recognition
  - Lexical ngram likelihood ratio (as before)
  - Duration, confidence, no. hyps, no. words in hyps

Modeling

- Tried multiple classifiers
  - Logistic Regression
  - Regression Tree
  - Random Forest
  - Boosting
- Likelihood ratio models estimated on training set, yield real-valued feature for classifier
  - encon-g, asr-ng as before
  - Both with and without length-normalization

Results by Classifier

- Adaboost best
- Diminishing returns as more modalities are added
- Some overfitting
  - More modalities can hurt



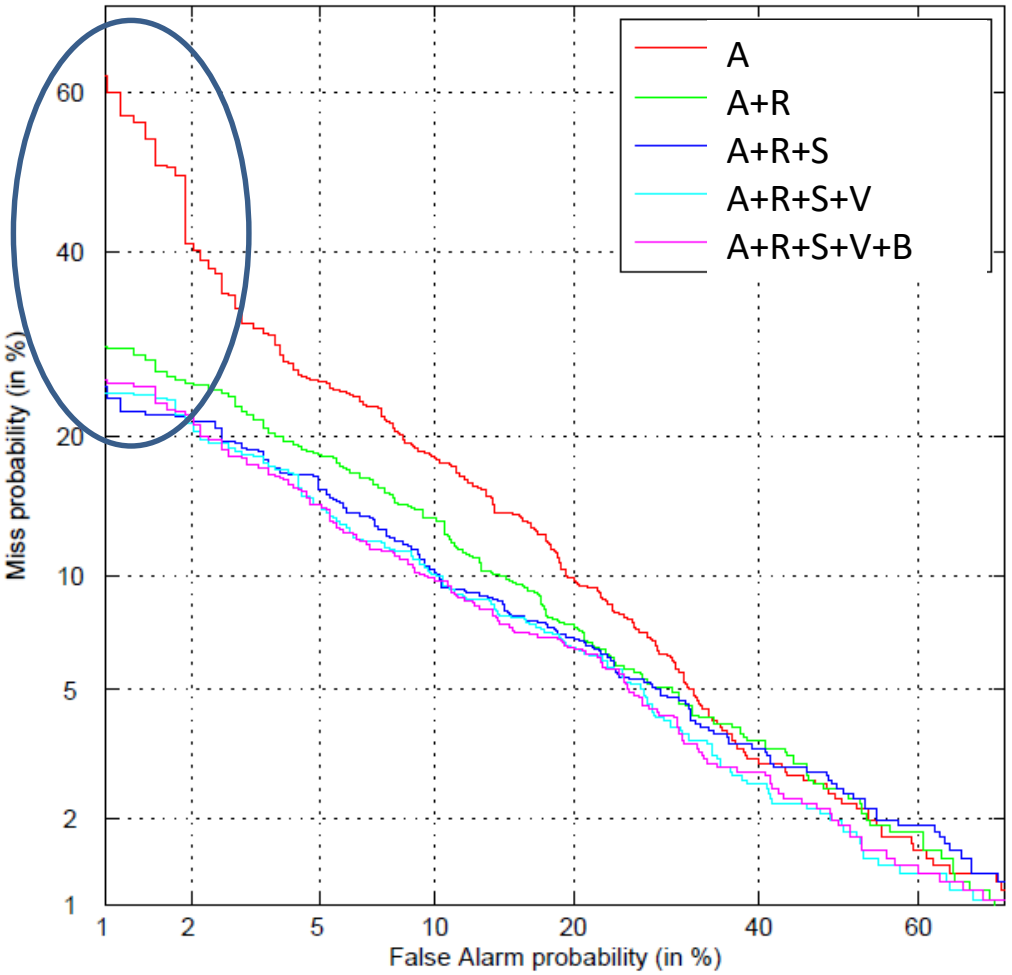
Performance by Modality

- Adaboost classifier
- Ranking modalities by
  - using only one modality at a time
  - leaving out one modality at a time

Modality	Leave-in EER %	Leave-out EER %
Acoustic	13.9	16.4
Reco	19.0	11.3
System	30.0	10.6
Visual	32.7	10.0
Beam	41.0	9.9
ALL	**9.8**	

DET Plot by Modality

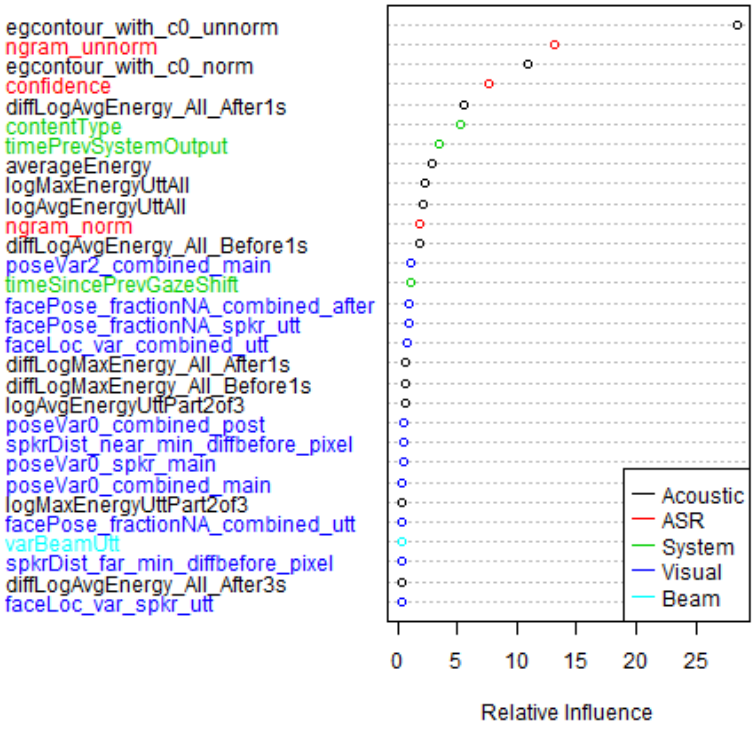
- Adaboost classifiers
- Modalities added in order of usefulness
- Multiple modalities help in low-false alarm region



Feature-level Relative Influence

- Based on adaboost classifier
- Rank features by loss reduction loss from single feature

Acoustic Reco System Visual Beam



Top 10

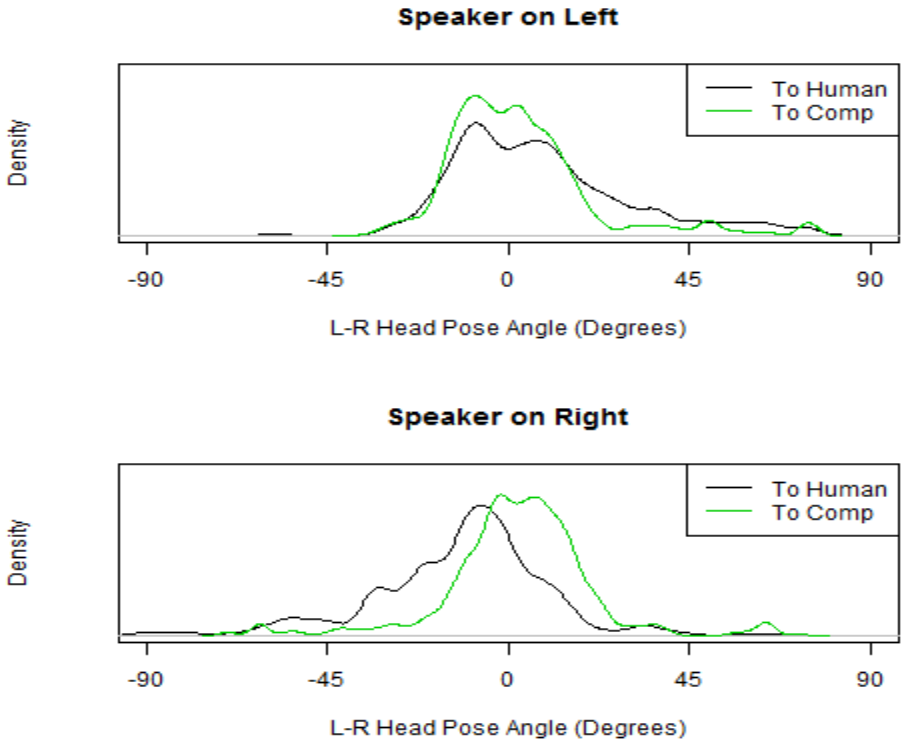
1. **Likelihood ratio: energy contour GMMs**
2. **Likelihood ratio: LMs**
3. **Likelihood ratio: energy contour GMMs (norm by length)**
4. ASR confidence
5. Energy change between utt and 1 sec interval after utt
6. Content type of agent speech
7. Time elapsed since agent speech
8. Log average energy during utt
9. Average energy during utt
10. Likelihood ratio: LMs (norm by number of words)

No Absolute Energy

- Energy is a very scenario-dependent feature, hard to calibrate
- What happens if we remove absolute energy info?
  - No individual abs energy features
  - Remove first DCT coefficient based on c0 in encon model
- Results
  - Overall EER 9.8% → 14.0%
  - **R**eco now more important than **A**coustic
  - **V**isual, **B**eam become somewhat more important
  - encon EER = 19% very similar to CB result (EER = 17%)

Why Don't Visual Features Help?

- Face pose does correlate with addressee



- But correlation is weak, redundant given other feats
- System screen serves as "situational attractor" even during H-H speech

In Summary …

- High-accuracy addressee detection in H-H-C scenarios is possible, using audio-based processing only
- Stylistic features:
  - speakers modify rhythmicity and vocal effort features when addressing a dialog system
  - Consistent across different corpora, domains
- Recognition-based lexical modeling
  - N-grams are effective
  - NN models improve over standard LMs based on continuous word embeddings
  - H-H speech can be modeled using general conversational corpora
  - H-C speech can modeled using single-user data
- Most information occurs early in an utterance (first 2 seconds)
- Effective on Cortana data

Summary, cont.

- Other modalities are somewhat useful
- Most useful: dialog system state and context
- Little gains from visual features
  - Screen is a situational attractor
- Microphone beam information not (yet) useful

Open Questions

- Can NN models be leveraged more thoroughly?
  - for prosodic features
  - combination of modalities (especially lexical + prosodic)
- Can multi-modal features benefit in general (despite negative results in our study)?
- How will algorithms hold up in Echo / Home style settings? – We need data!

Credits

- Co-authors

  Dilek Hakkani-Tür, Larry Heck, Heeyoung Lee, Suman Ravuri, Malcolm Slaney, Elizabeth Shriberg, TJ Tsai

- Acknowledgments

  Ashley Fidler, Dan Bohus, Gokhan Tur, Lisa Stifelman, Madhu Chinthakunta, Oriol Vinyals

Thank You!

References

- E. Shriberg, A. Stolcke, D. Hakkani-Tür, & L. Heck, Learning When to Listen: Detecting System-Addressed Speech in Human-Human-Computer Dialog, Proc. Interspeech, pp. 334-337, 2012
- H. Lee, A. Stolcke, & Elizabeth Shriberg, Using Out-of-Domain Data for Lexical Addressee Detection in Human-Human-Computer Dialog, Proc. NAACL/HLT, pp. 221-229, 2013
- E. Shriberg, A. Stolcke, & S. Ravuri, Addressee Detection for Dialog Systems Using Temporal and Spectral Dimensions of Speaking Style, Proc. Interspeech, pp. 2559-2563, 2013
- S. Ravuri & Andreas Stolcke, Neural Network Models for Lexical Addressee Detection, Proc. Interspeech, pp. 298-302, 2014
- S. Ravuri & Andreas Stolcke, Recurrent Neural Network and LSTM Models for Lexical Utterance Classification, Proc. Interspeech, pp. 135-139, 2015
- T. J. Tsai, A. Stolcke, & M. Slaney, A Study of Multimodal Addressee Detection in Human-Human-Computer Interaction, *IEEE Transactions on Multimedia* 17(9), 1550-1561, 2015