

**When Good Recognizers Go Bad:  
Identifying Prosodic Cues to  
Recognition Errors**

**Julia Hirschberg, Diane Litman**

***AT&T Labs -- Research***

**Mark Swerts**

***IPO***

**28 July 1999**

# What makes an utterance difficult to recognize?

- The speaker?

- female
- child
- non-native accent

- The words?

- OOV
- highly confusable with others in lexicon

- The speaking style?

- informal (Switchboard)
- hyperarticulate (corrections)

- Can we identify hard-to-recognize utterances as they are produced?

## **Related Research**

- **Sheep vs goats is a real distinction (Doddington et al 1998)**
- **Casual speaking style hurts ASR performance (Weintraub et al 1996)**
- **Modeling speaking mode effects can help identify acoustic modeling errors (Ostendorf et al 1997)**
- **Prosody can be used to rank recognition hypotheses (Wang & Hirschberg 1993; Veilleux 1994, Hirose 1997)**
- **Hyperarticulate speech often characterizes corrections and can hurt ASR performance (Wade et al 1992, Oviatt et al 1996, Levow 1998, Bell & Gustafson 1999)**

# Overall Goals

- Identify distinguishing prosodic characteristics of
  - misrecognized utterances
  - speaker corrections of recognition failures (misrecognitions and rejections)
  - speaker responses to evidence of ASR recognition failures
- Develop and test methods for using this information in spoken dialogue systems to recover from recognition failures in IVR systems
  - improve rejection decisions
  - inform decision to change dialogue strategy
  - change recognizer (acoustic models, dictionary, language mode) to deal with difficult utterances

# Identifying Misrecognitions

- Examine prosodic features of incorrectly recognized speaker turns to see if they differ significantly from features of correctly recognized turns
- Try to predict this distinction using automatically available prosodic features, alone and in combination with other automatically generated features (ASR confidence score, recognized string,...)
- Practical goals:
  - improve rejection strategy
  - coach speakers on how to get better performance from system
  - suggest additional requirements for ASR training materials

## Method

- Obtain corpus of recognized speech from an interactive voice response system
- Label speaker turns for recognition *concept accuracy* (semantic correctness)
- Obtain prosodic/acoustic features for each *speaker turn*
- Identify additional ASR-derived features for each turn
- Compare incorrectly vs. correctly recognized turns wrt all features to uncover significant differences
- Use machine learning techniques to derive automatic methods for identifying likely misrecognitions in next version of system

## **TOOT IVR Dialogues**

- Collected in 1998 to investigate differences in dialogue strategy wrt initiative (system, user, mixed) and type of confirmation provided (explicit, implicit, none)
- Domain: train information over the phone
- Subjects: 39 summer students -- 16/23 (F/M), 20/19 (native speaker/non)
- Task: find train information for 4 scenarios over the phone with speech i/o
- ASR: BLASR
  - 1152 context-dependent subword models trained on telephone speech
  - barge-in
  - different grammars at each dialogue state

- Corpus of 152 dialogues labeled by hand for concept accuracy based on number of database entities recognized correctly (e.g. target city, destination city, time, date)

- Original data: 2328 user turns (including 201 rejections, 546 semantic misrecognitions)

- Usable corpus: 2067 turns (including 170 rejections, 491 semantic misrecognitions)

- Concept accuracy treated as a binary feature

## **BLASR Version x.y**

- Acoustic models trained on telephone speech
  - 1152 context-dependent subword models
  - 11 whole-word digit models
  - multiple silence and noise models
- Grammars set in application, varied by dialogue state
- Produced 1-best match to a sentence in the grammar
- endpointing, no cepstral mean subtractions, vector quantization, smoothing of context-dependent phonemes

# Features Examined per Turn

- **Experimental conditions**
  - Mode of interaction: initiative type, style of confirmation
  - Subject
  - Task (first through fourth)
- **Prosodic/acoustic features**
  - f0 maximum and mean
  - rms maximum and mean
  - total duration
  - duration of preceding silence
  - amount of silence within turn
- **ASR-produced information**
  - confidence score
  - recognized string

## **Data Analysis**

- Calculated mean values for each feature by speaker for correctly vs. incorrectly recognized utterances (removing rejections)
- Paired t-test on means (e.g., for speaker 1, pairing mean values for correctly and incorrectly recognized turns)

# Results

•F0 and RMS maximum, RMS mean, overall duration, and length of prior pause show significant differences between correctly and incorrectly recognized turns

Feature	T stat	Mean Misrec - Mean Rec	P
<b>F0 max</b>	<b>5.40</b>	<b>28.87 Hz</b>	<b>0.000</b>
<b>RMS max</b>	<b>3.00</b>	<b>185.74</b>	<b>0.005</b>
<b>RMS mean</b>	<b>2.07</b>	<b>-32.43</b>	<b>0.050</b>
<b>Duration</b>	<b>10.42</b>	<b>2.34 sec</b>	<b>0</b>
<b>Prior pause</b>	<b>5.22</b>	<b>.038 sec</b>	<b>0</b>
Tempo	.45	.13 sps	.65
F0 mean	1.36	1.63 Hz	.18
% silence	.05	-.02%	.29

• Similar results for normalized features and when rejections included with misrecognitions (i.e. 'non-recognitions' vs correct recognitions -- although no information on correctness of rejections)

# **Are we just finding instances of hyperarticulation?**

- Hyperarticulated speech generally characterized as louder, slower in tempo and longer overall, with greater f0 excursion, internal pauses separating words and syllables, and exaggerated pronunciation
- Two authors previously hand labeled turns for evidence of hyperarticulation on 3 point scale (0:no hyperarticulation, 1:some part of turn hyperarticulated, 2:whole turn hyperarticulated, following Wade et al '92)
- Divided data into 0 labels by both labelers vs. rest
- Findings
  - Hyperarticulated turns recognized more poorly than non-hyperarticulated turns
  - Misrecognized turns have significantly higher hyperarticulation scores than correctly recognized turns

- But...when we remove hyperarticulated turns from the data set, we still find significantly higher f0 and rms maxima, and overall duration for misrecognized turns compared to recognized turns

- So, the differences are doing more than distinguishing hyperarticulated turns from others

# Sheep and Goats

- Previous findings speaker independent, but could the characteristics that distinguish misrecognitions from correct recognitions also help to distinguish goats from sheep? Do some speakers exhibit more extreme prosodic behavior than others?

- Divided speakers into goats (non-recognized/recognized turns > .5) and sheep

- of 18 goats, 7 were women, 9 were non-native, and 14 of 18 were either or both

- Chose one experimental condition with roughly equal numbers of each (mixed initiative, implicit confirmation), with 8 goats and 5 sheep

- of goats, 3 were non-native and 5 female (1 overlap)

- of sheep, 2 were non-native and 1 female (no overlap)

- Again calculated means for prosodic features on a per speaker basis, over all turns

- Findings:

- normalizing for F0, significant differences between sheep and goats only in speaking rate
- significant difference in raw F0 maxima (more females are goats than males)

## **Can we predict whether a particular turn will be correctly recognized or not?**

- Machine learning experiments to automatically induce rule sets for predicting semantic accuracy scores
- Ripper (Cohen '96) uses greedy search guided by a measure of information gain to produce ordered rule set for input vectors of feature values and observations of semantically accurate or not (excluding rejections)
- Feature set included all automatically obtainable
  - all prosodic/acoustic features
  - experimental condition features
  - ASR confidence scores and recognized string

# Ripper Results

- Baseline (predicting most frequent class): 26% error
- Best performance: ~11% error using prosodic features, ASR confidence, recognized string, system experimental conditions, but no significant difference using
  - prosody+ASR confidence+recognized string
  - prosody+recognized string
  - prosody+ASR confidence
  - ASR confidence+recognized string
- Rules for prosody+ASR confidence+recognized string
  - if (asr <= -3.33589) AND (dur >= 1.42496 ) AND (tempo <= 2.54345)  
then F
  - if (asr <= -3.80883) AND (dur >= 1.13314 ) AND (f0av <= 201.595)  
then F
  - if (asr <= -2.75271) AND (tempo <= 1.51501 ) AND (zeros <= 0.595294) then F
  - if (asr <= -2.74865) stringsvals ~ help then F

```
if (asr <= -5.04968) AND (dur $geq$ 0.925401) AND (zeros <=
0.453237) then F
if (asr <= -2.71435) AND (dur $geq$ 0.982058) AND (f0max <=
155.887) then F
if (asr <= -3.61652) AND stringsvals ~ '8' then F
if (dur >= 1.39754) AND (tempo $leq$ 2.25751) AND (zeros <=
0.626667) then F
if tempo >= 7.25392 then F
else T
```

- Recognized string is best single features (14.41% error), but not significantly different from all prosodic features (16.15% error)
- Both better than ASR confidence measures alone (18.3% error)
- Conclusion: misrecognitions can be predicted better than chance and better than simple rejection strategy (nb: rejections excluded from above, so this is additional predictive power)

## Discussion

• Semantic misrecognitions are different from correct recognitions in terms of their prosodic features:

- F0 maximum (higher)
- RMS maximum (louder)
- mean RMS (lower)
- turn duration (longer)
- preceding pause (longer)

• Effect holds up across speakers and when hyperarticulated turns are excluded and appears independent of native/non-native and gender distinctions

• Prosodic features can be used with other automatically available features to identify semantic misrecognitions with ~11% error

## **Future Research**

- Do results hold up across recognizers?
- How could this information really be useful in recognition? in designing better IVR systems?
- What makes a goat?
- Are user corrections of recognition failures also prosodically distinct? User 'awareness sites'?