Predicting Automatic Speech Recognition Performance Using Prosodic Cues

Diane Litman, Julia Hirschberg

AT&T Labs -- Research

Mark Swerts, IPO

NAACL 2000

Systems Must Recognize their Own Mistakes

- Users find it difficult to correct system misunderstandings (Krahmer et al '99)

  User: I want to go to Boston.

  System: What day do you want to go to Baltimore?

- Easier to correct explicit requests for confirmation or ASR rejections:

  System: Did you say you want to go to Baltimore?

  ...or...

  System: I'm sorry. I didn't understand you. Where do you want to go?

- But (unnecessary) explicit confirmation or rejection makes the dialogue longer and more tedious

Can We Identify Mistakes Automatically Using Prosodic Information?

- 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,...)

TOOT Dialogues

- Collected in 1998 (Litman & P an '99) to investigate differences in dialogue strategy wrt adaptability , type of initiative (system, user , mixed) and type of confirmation provided (explicit, implicit, none)
- Platform: StarTrek SGI version (Kamm et al '97)
- Subjects: 39 summer students -- 16/23 (F/M), 20/19 (native speaker/non)
- Task: find train information for 4 scenarios over the phone with speec h i/o
- Corpus of 152 dialogues: 1994 speaker turns (1975 with ASR hypothesis) used for current study
- Overall word error rate (WER): 49%; overall concept accuracy(C A): 71%

A Successful Dialogue

S: Hi, this is AT&T Amtrak schedule system. This is TOOT. How may I help you?

U: I would like a train to New York City from Philadelphia on Sunday at ten thirty P M

S: I heard you say " go to New York City from Philadelphia on Sunday approximately at ten thirty p.m. " Do you want me to find the trains from Philadelphia to New York City on Sunday approximately at ten thirty p. m. now?

U: Yes.

S: I am going to get the train schedule for you....

An Unsuccessful Dialogue

S: Hi, this is AT&T Amtrak schedule system. This is TOOT. How may I help you?

U: I'd like to find a train from Ch- to Chicago from Baltimore.

S: Sorry, I can't understand you. Please repeat your utterance.

U: I'd like to find a train from Baltimore to Chicago.

S: I heard you say " leave today around 5 thirty in the morning. " Which city do you want to go to?

U: I would like to leave Saturday at eight in the morning.

S: You can say the arrival city name, such as "New York City."

U: Baltimore to Chicago.

S: You can say the arrival city name, such as “New York City.”

U: Baltimore to Chicago. (I hate it.)

U: Train from Baltimore ...

S: ...I heard you s ...

U: to Chicago .... hello.

S: You can say the name of your departure city, such as “New York City.”

U: Baltimore.

U: Bal-ti-more ....

S: You can say...

U: Baltimore.

S: Do you want to continue and find a new set of trains?

U: Yes.

S: You can say “yes” or “I am done here” to find a new set of trains.

U: Yes.

S: Thank you for using AT&T Amtrak train time table system. See you next time.

U: I nev-

Distinguishing Correct from Incorrect Recognitions

- Correct recognitions defined as WER=0 or CA=1
- For each speaker, for each feature, calculated mean values for each correctly recognized *speaker turn* and for each incorrectly recognized turn
- Performed paired t-tests on speaker means (e.g., for each speaker, pairing mean values for correctly and incorrectly recognized turns)

Prosodic Features Examined per Turn

- Raw prosodic/acoustic features
  - f0 maximum and mean
  - rms maximum and mean
  - total duration
  - duration of preceding silence
  - amount of silence within turn
  - speaking rate (syllables of recognized string per second)
- Normalized versions of each feature (compared to first turn in task, to mean of task, Z scores)

WER Results

- F0 and RMS maximum, F0 mean, overall duration, length of prior pause , and % silence in turn show signifi cant differences between correctly and incorrectly recognized turns

Feature	T stat	Mean Misrec - Mean Rec	P
F0 max	**7.83**	**30.31 Hz**	**0.000**
F0 mean	**3.66**	**4.12 Hz**	**0.000**
RMS max	**5.70**	**235.93**	**0.000**
RMS mean	-.57	-8.50	0.570
Duration	**10.30**	**2.20 sec**	**0.000**
Prior pause	**5.55**	**0.35 sec**	**0.000**
Tempo	-.05	.15 sps	0.130
% silence	**-5.15**	**-.06 %**	**0.000**

Are we just finding instances of hyperarticulation?

- Corrections are often hyperarticulated (louder, slower, more internal pauses, exaggerated pronunciation), leading to more ASR error (Wade et al '92, Oviatt et al '96, Swerts & Ostendorf '97, Levow '98, Bell & Gustafson '99)

- 34.8% of turns hyperarticulated

- Hyperarticulated turns are recognized more poorly (81% WER) than non-hyperarticulated turns (41%)

- More misrecognized turns are hyperarticulated (41%) than correctly recognized turns (10.1%)

- But.... removing hyperarticulated turns from the data set does not change prosodic differences found

Using Machine Learning to Predict Misrecognized Turns

- Performed ML (Ripper) experiments to automatically induce rule sets for predicting WER and CA scores
- Feature set:
  - all prosodic features, raw and normalized
  - features related to the original experimental conditions (adaptability of system, initiative type, style of confirmation, subject, task)
  - gender, native/non-native status
  - ASR recognized string, grammar and acoustic confidence score

ML WER Results

- Baseline (predicting most frequent c lass): 48.66% error

Feature Sets	Error (SE)
Prosody, ASR Conf, String, Grammar	6.53% (.63)
ASR Conf, String, Grammar	9.01% (.70)
Prosody	12.76% (.79)
ASR String	15.24% (1.11)
Duration	17.42% (.88)
ASR Conf	22.23% (1.16)

Best WER Rule-Set

if (duration <= 0.90) ^ (conf >= -2.63) ^ (string contains `yes’) then T

if (duration <= 1.04) ^ (conf >= -2.70) ^ (string contains `no’) then T

if (duration <= 0.98) ^ (conf>= -2.00) ^ (tempo >= 3.11) then T

...

else F

Current Research

- Effect holds up for new corpus (W99) and different ASR engine (AT&T Watson)
- Corrections and ‘aware’ sites are also prosodically distinct in the TOOT corpus
- ‘Goats’ are prosodically different from ‘sheep’ in speaking rate and F0 maxima (at least)
- Future application: train a speech recognizer on corrections and run in parallel with standard recognizer on turns predicted to be misrecognized to improve performance

Conclusions

- Misrecognitions are different from correct recognitions in terms of their prosodic features:
  - F0 maximum (higher)
  - RMS maximum (louder)
  - turn duration (longer)
  - preceding pause (longer)
- Effect holds up across speakers and when hyperarticulated turns are excluded
- Prosodic features plus other automatically available features (acoustic confidence scores, recognized string . grammar) predict WER-based misrecognitions with ~6.5% error -- better than ~22% of acoustic confidence measures alone

Related Research

- Casual speaking style hurts ASR performance (Weintraub et al ‘96)
- Modeling speaking mode effects can help identify acoustic modeling errors (Ostendorf et al ‘97)
- Prosody can be used to rank recognition hypotheses (Wang & Hirschberg ‘93; Veilleux ’94; Hirose ‘97)
- Corrections are often hyperarticulated, leading to more ASR error (Wade et al ‘92, Oviatt et al ‘96, Swerts & Ostendorf ‘97, Levow ‘98, Bell & Gustafson ‘99)
- Promising results in predicting poor recognition performance at the dialogue level, using a wide variety of automatically generated features (Litman, Walker & Kearns ‘99)