3pSC3. Automatic scoring of responses to the Hearing In Noise Test using utterance verification based on hidden Markov models

Jason Lilley[a], Justin Aronoff[b], Sigfrid Soli[b], H. Timothy Bunnell[c], Ivan Pal[d]

[a]Department of Linguistics & Cognitive Science, University of Delaware, Newark DE; [b]House Ear Institute, Los Angeles CA;

[c]Nemours Biomedical Research, Alfred I. duPont Hospital for Children, Wilmington DE; [d]Compreval Inc., Grayslake IL

BACKGROUND

The **Hearing In Noise Test** (HINT) (Nilsson et al. 1994) adaptively measures *speech recognition thresholds* (SRTs): the *signal-to-noise ratio* (SNR) at which a listener has difficulty recognizing sentences.

- Listener hears 20 sentences embedded in noise and must repeat each nearly exactly (certain deviations are allowed, e.g. "a" for "the")
- The SNR of each sentence changes adaptively according to the previous response (e.g. the SNR is decreased if the listener answers correctly)
- So each response must be scored manually before the next sentence can be presented

Goal: Automate HINT scoring with an **utterance verification** (UV) engine

- Eliminates the necessity and subjectivity of human scoring

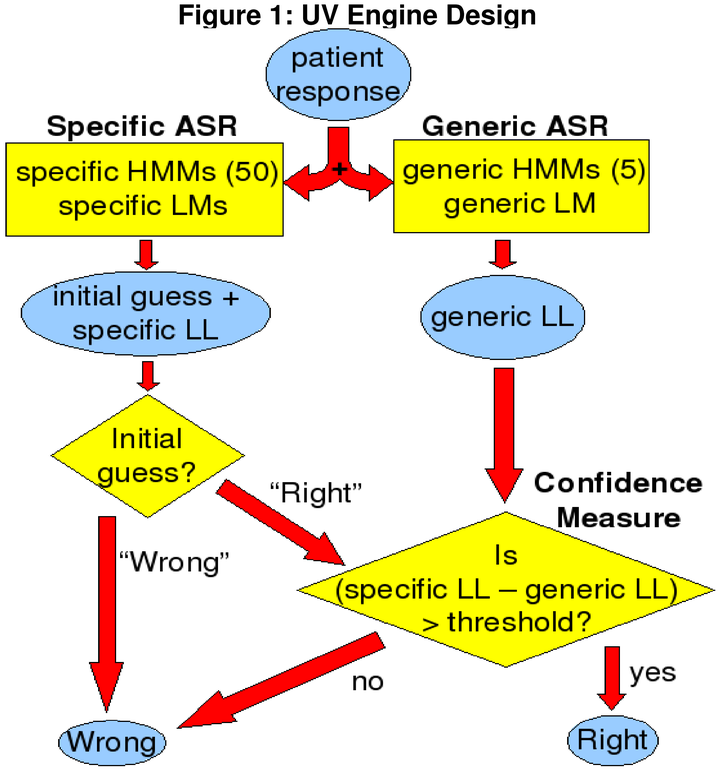
PROTOTYPE UV ENGINE CONSTRUCTION

HINT Seed Corpus

- 25 normal hearing AE speakers * 80 HINT sentences = 2000 utterances
- Collected under simulated HINT conditions
- SNRs varied under a preset schedule to maximize unique errors

UV Engine Design and Training

Figure 1: UV Engine Design



Specific ASR components (Young, 1993):

1. Acoustic Models

- 10-ms frame rate, 25-ms window size
- 13 MFCCs * 3 = 39D speech vectors
- 50-phoneme monophone HMM set trained on TIMIT (Garofolo et al. 1986)
- Converted to 3-Gaussian triphones, then trained on HINT seed corpus

UV Engine Design and Training (cont.)

2. Language Models

- One language model for each HINT stimulus sentence
- Consists of a list of actual responses to that sentence from seed corpus
- Each response is labeled "correct" or "incorrect"

Generic ASR components (Young, 1993):

1. Acoustic Models

- 1-Gaussian monophone models trained on TIMIT
- 5 generic models:
  - Generic consonant model trained on all TIMIT consonants
  - Generic vowel model trained on all TIMIT vowels
  - 3 generic silence models (Start, Mid, End)

2. Language Model

- Simply aligns the 5 generic HMMs to the utterance in any order

Confidence Measure (CM):

Initial CM **threshold** determined from cross-validation study of seed corpus

- Threshold chosen to minimize errors of UVE when run on seed corpus
- Improved initial accuracy from 91.15% to 93.00%

EVALUATION STUDY

We evaluated the UV engine in a real-world setting, in which the UV engine controlled the SNR of HINT stimuli to **25** new normal-hearing AE speakers

METHOD:

Each subject's Speech Recognition Threshold was measured 4 times with the old manual software, and 4 times with the new software incorporating the UV engine

The UV components were updated after every block of 5 subjects:

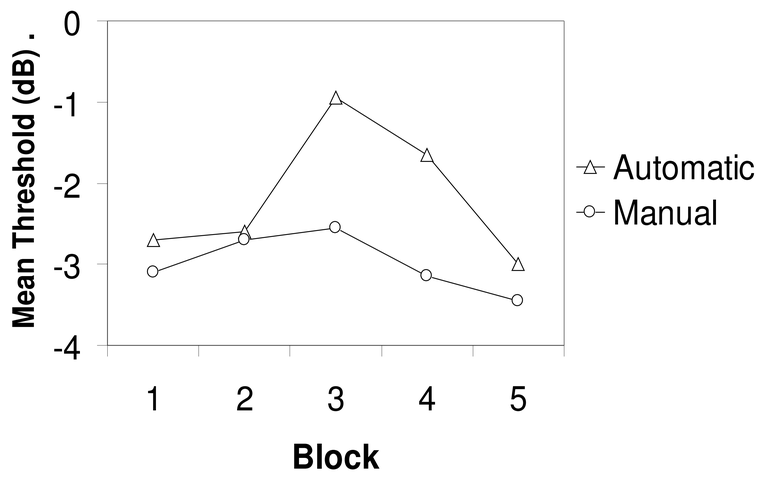
- Novel incorrect responses added to language models
- Specific HMMs retrained on all subjects
- Confidence Measure threshold re-estimated from all subjects

RESULTS:

Overall diff in threshold SNRs between human and ASR: **0.775 dB**

- **This is about half the test-retest reliability of human scorers: 1.5 dB** (Vermiglio 2008)

Figure 2: Mean Measured Threshold SNRs by Block and Method



RESULTS (cont.):

Mixed model ANOVA:

- Block (1-5) as between-subjects factor, Method (auto vs. manual) as within-subjects repeated measure
- Main effect of Block: **non-significant** (F[4,20]=1.35, p=0.286)
- Main effect of Method: **non-significant but marginal** (F[1,20]=3.88, p=0.063)
- Interaction of Block and Method: **non-significant** (F[4,20]=0.57, p=0.69)

Most of the difference is due to 1 subject in block 3, and 2 subjects in block 4

- Most of this difference is due to the Confidence Measure incorrectly rejecting correct utterances

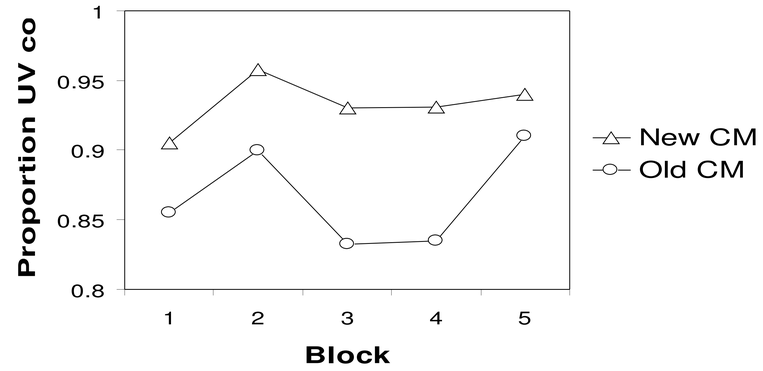
POST-HOC IMPROVEMENT OF CONFIDENCE MEASURE

METHOD:

- Added bigram phoneme recognizer to CM algorithm
  - bigram probabilities from Brown corpus (Kucera and Francis, 1967)
- A new log-likelihood is calculated from the bigram phoneme recognizer
- All LLs and associated parameters used in a logistic regression model
- New cross-validation study run on evaluation study data

RESULTS:

Figure 3: Old and new UV sentence classification accuracies



- Overall improvement across all blocks
- The decrease in accuracy in blocks 3 and 4 is greatly reduced

CONCLUSIONS

- Our study shows the feasibility of using speech recognition to automate HINT.
- Human/ASR score difference (0.775 dB) is well within human test-retest difference (1.5 dB)
- Improvement of accuracy is possible by fine-tuning the parameters of the UV engine.

Acknowledgements: Work supported by NIDCD grant #R43DC008212.

References

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., and Zue, V. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus.* Linguistic Data Consortium, Philadelphia. http://www.ldc.upenn.edu/Catalog/LDC93S1.html.

Kucera, H., and Francis, W.N. (1967). *Computational Analysis of Present-Day American English.* Providence, R. I.: Brown U. Press.

Nilsson, M., Soli, S.D., and Sullivan, J.A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustic Society of America* 95: 1085-1099.

Vermiglio, A.J. (2008). The American English Hearing in Noise Test. *International Journal of Audiology*, 1708-8186, Issue 6, pp. 386-7.

Young, S.J. (1993). *The HTK Hidden Markov Toolkit: Design and Philosophy*, in *Technical Report.* Department of Engineering, Cambridge University.

[Presented on April 21, 2010, at the 159th Meeting of the ASA, Baltimore, MD]