

Objective Measures for Evaluating the Quality of Digital Speech Transmission

Kevin L. Mills

Advanced Networking Division

Stefan Leigh

Statistical Engineering Division

C. Michael Chernick

Distributed Computing & Information Services Division



What we are trying to do, and Why

Goal: Develop an objective measure for speech quality that correlates with the quality perceived by human listeners and that can be effectively automated.

Motivation

- To provide a means for developers of voice coding algorithms to repeatedly, automatically, and inexpensively test their algorithms under a range of networking conditions, emphasizing wireless and Internet voice traffic
- To provide a reliable means of correlating an objective measure of speech quality with speech quality as perceived by human listeners



How Others Approach the Measurement of Speech Quality

SUBJECTIVE

Use Human Opinion

- Mean Opinion Score (5, excellent, to 1, poor) is by far the most common approach
- Diagnostic Acceptability Measure (16 specific characteristics each on a 100-point scale)

OBJECTIVE

Compute Difference Between Channel Input and Output Signals for Some Property

- Signal-Noise Ratio (SNR), Segmental SNR, or a frequency variant of Segmental SNR (appropriate only for waveform coders)
- Cepstral Distance, Coherence Function, Information Index, Spectral Distance (intended for vocoder-like systems)
- Combinations of Single Distance Measures

Apply Psycho-Acoustical Transform to Channel Input and Output Signals for Some Property

- Irregularity, Rasping, Hissing, Crackling
- Gamma Tone Filtering, Hair Cell Transduction, and Adaptation Looping
- Loudness-Time-Pitch model using four transformations in series

Measure Distance Between Channel Output and a Reference

- Quantized Channel Output vs. Quantized Entries in a Reference Codebook



Limitations of Current Approaches

- Subjective approaches: (1) are expensive, (2) are not easily repeatable, (3) possess inherent variability in interpretation of the rating scale by humans, and (4) for large numbers of characteristics with fine-grained scoring ask humans to make finer judgments than might be reasonable
- Designers of objective measures typically attempt to show reasonable statistical correlation between their measure and subjective human judgments, which themselves are highly variable
- Literature exhibits a wide variation in correlation with subjective human judgments, even for the same objective measures - this variability is often due to differences in the test data, coding algorithm used, error conditions assumed, and subjective scales employed (i.e., there exists no standard means for evaluating objective measures for speech quality)
- Even if a standard means of evaluating objective measures existed, objective measures meant to predict subjective measures are limited by the variability inherent in the subjective measures



What's New in the Approach Here?

USE SPEECH RECOGNITION TECHNOLOGY TO MEASURE SPEECH QUALITY

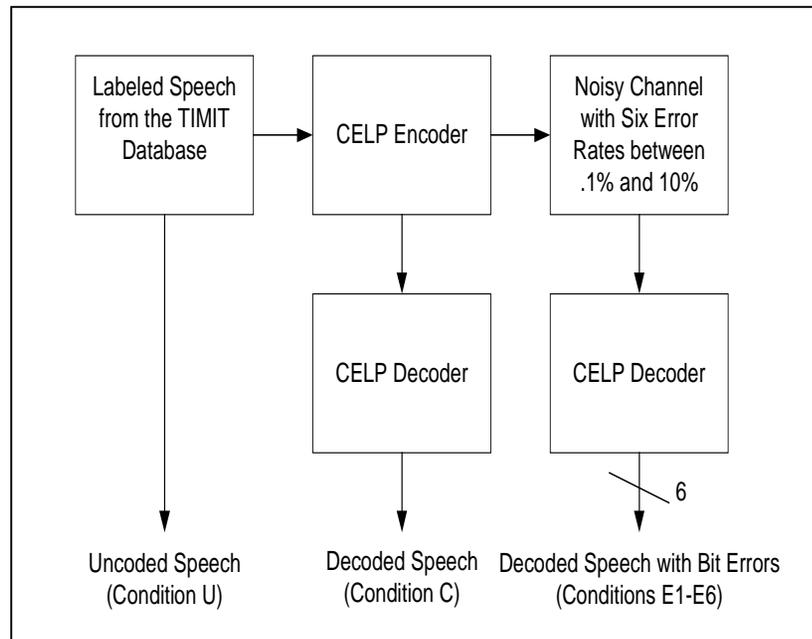
- Use a speech recognizer to generate a transcription of the voice signal output from the channel and then compute the word-error rate against a reference transcription.
- An alternative might be to have the speech recognizer generate a transcription of the signal input to and output from the channel and then compute the word-error rate between the two transcriptions.

Why would this be an improvement?

- Metric validation would be conducted against humans who also create transcriptions, which should provide a more reliable correlation with an objective measure because the transcription task yields an objective result.

Initially, we investigated the ability of a speech recognizer to predict human subjective perception, measured as a mean opinion score on a five-unit ordinal scale.

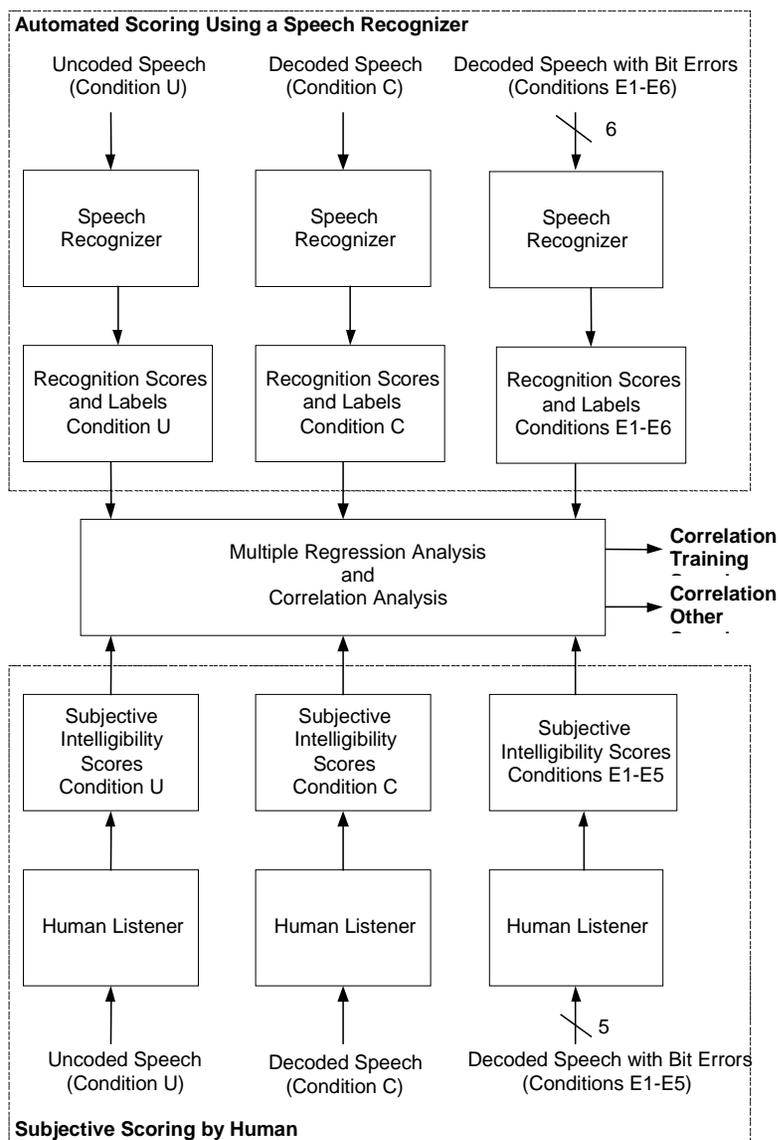
Generating the Speech Samples



- **19 speakers selected from ITL TIMIT speech database**
- **Combinations of coding and bit-error injection produced 152 speech samples for use in scoring**
- **For the human listening tests, we used a subset of only 14 speakers, and we discarded all samples with 10% bit errors (too noisy for humans)**



Scoring Speech Recognizer vs. Human Perception

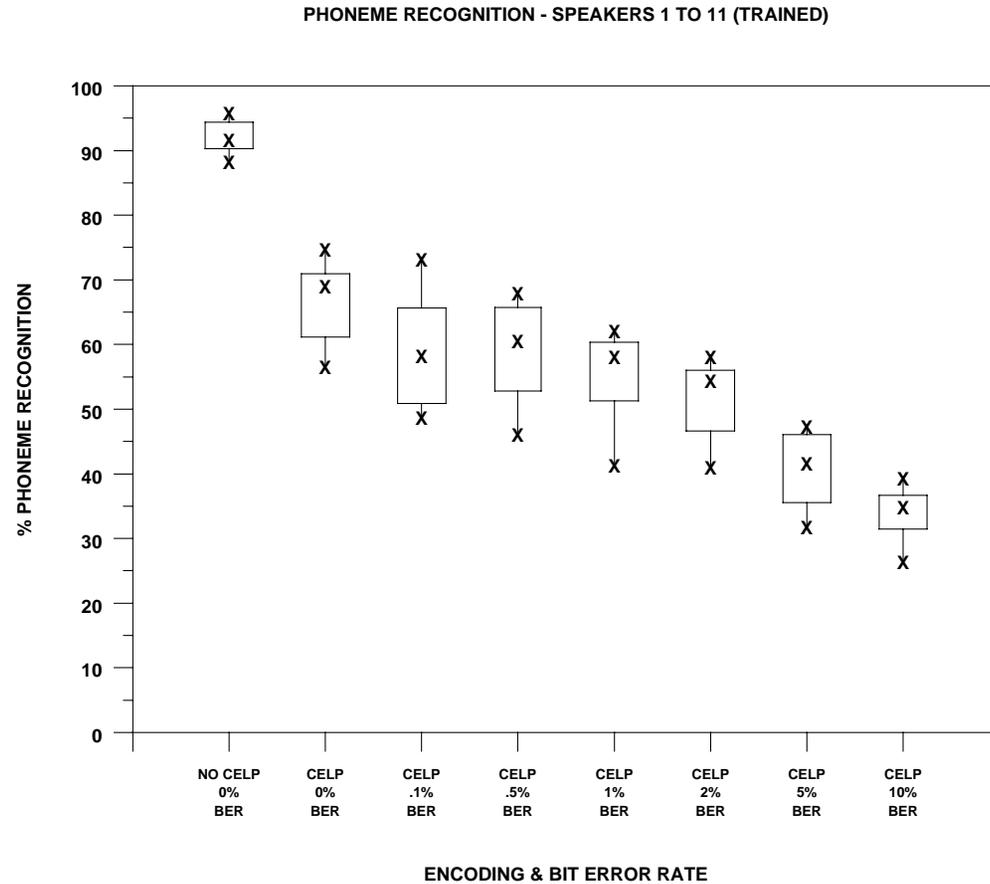


Measuring Performance

- **Used the HTK speech recognizer from Cambridge University**
- **11 of the 19 speakers were used to train the speech recognizer**
- **All 152 speech samples were offered to the speech recognizer and phoneme recognition scores were computed**
- **14 human listeners each evaluated 14 speech samples, and assigned a mean opinion score**
- **We computed correlation between the results of the speech recognizer and the human listeners**

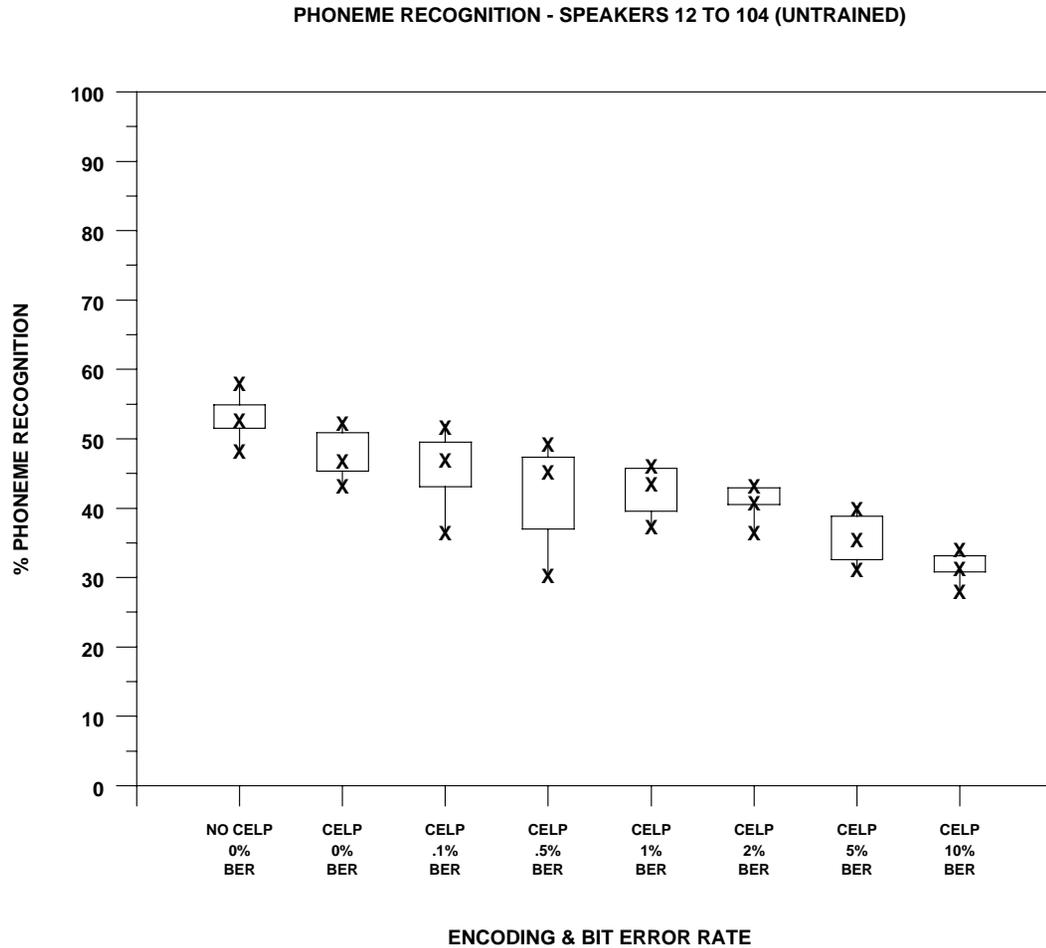


Performance of Speech Recognizer - Trained Speakers





Performance of Speech Recognizer - Untrained Speakers





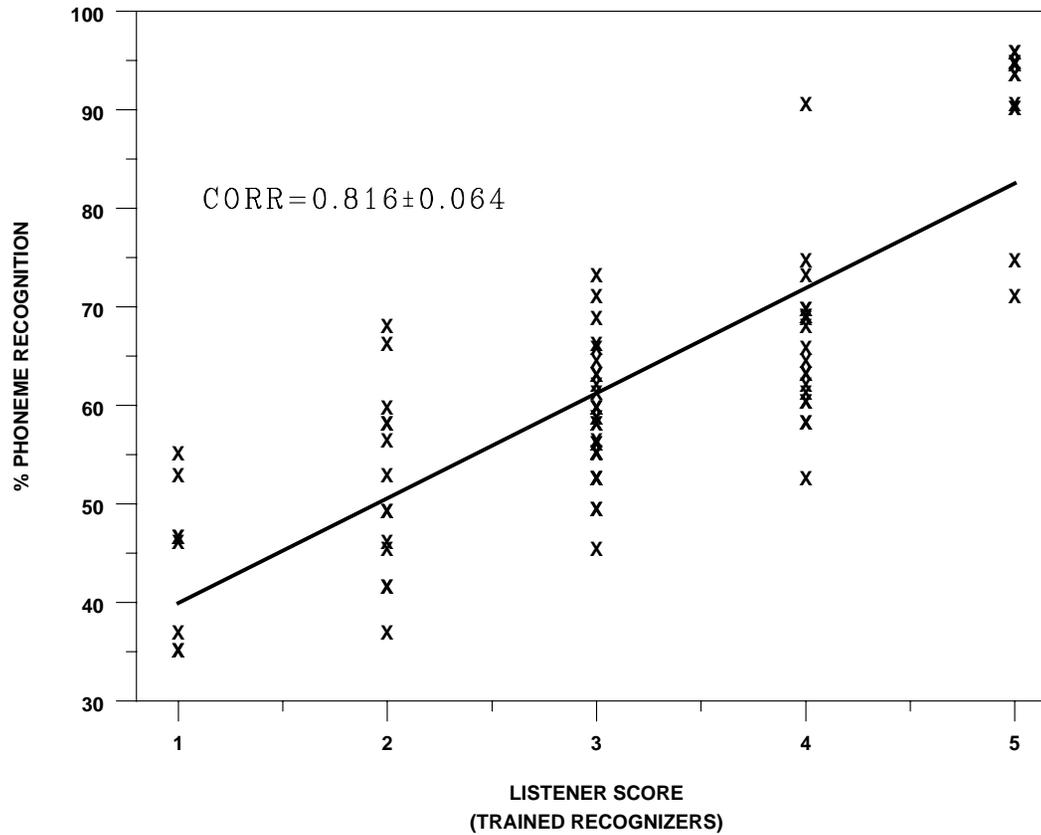
Mean Opinion Scores Assigned by Human Listeners

Speaker Number	No CELP 0% BER	CELP 0% BER	CELP .1% BER	CELP .5% BER	CELP 1% BER	CELP 2% BER	CELP 5% BER
1	5.0	4.0	4.0	3.5	3.0	2.5	2.5
3	5.0	3.0	3.0	3.5	3.0	1.5	1.5
4	4.5	3.5	4.0	3.0	2.5	3.0	2.0
7	5.0	4.0	3.5	2.5	3.0	2.0	1.0
10	5.0	4.0	3.5	3.5	3.0	2.0	1.0
11	5.0	4.5	4.0	3.0	4.0	2.0	1.5
12	4.5	4.0	3.5	3.0	2.5	2.5	1.0
13	5.0	5.0	3.5	3.0	3.0	2.5	1.5
16	4.5	4.0	4.0	3.5	3.5	2.0	1.0
17	5.0	4.0	3.5	4.0	3.0	2.5	1.5
101	4.5	3.5	3.0	2.0	2.0	1.5	1.0
102	5.0	3.5	4.0	3.0	2.5	2.0	1.0
103	4.5	3.5	3.0	3.0	1.5	1.5	1.0
104	5.0	5.0	4.0	3.0	2.5	2.0	1.0
All	4.82	3.96	3.61	3.11	2.79	2.11	1.32



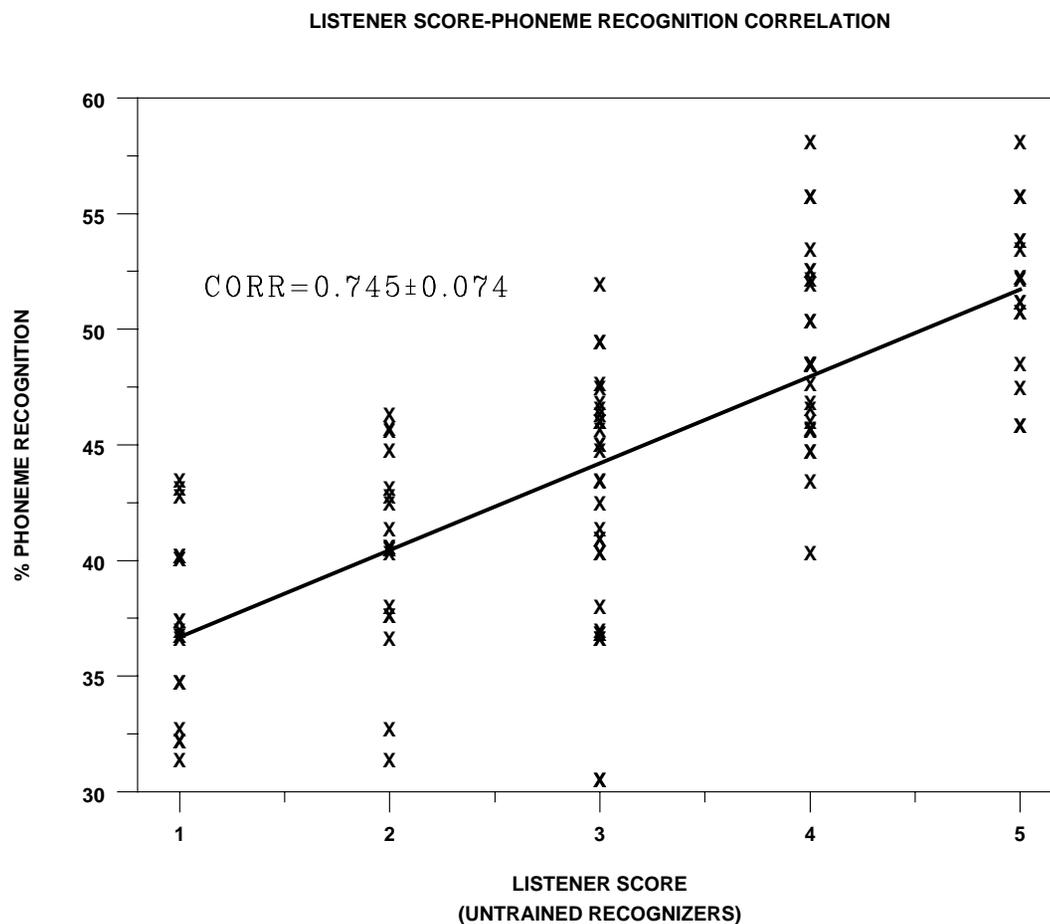
Speech Recognizer vs. Human Listener - Trained Speakers

LISTENER SCORE-PHONEME RECOGNITION CORRELATION





Speech Recognizer vs. Human Listener - Untrained Speakers





Results Encouraging - Next Steps?

- Repeat our experiment but evaluate the performance of commercial speech recognizers against human listeners on a transcription task
- If results are successful, construct a prototype test system that can select among various commercial recognizers, error models, and speech samples.
- Use the prototype test system to evaluate our approach against a variety of coding algorithms
- If results are successful, install our test system in a Web-accessible form, and make the software available for downloading and use by designers of coding algorithms and implementers of voice transmission products (e.g., cell-phone developers and Voice-Over-IP developers)