TESTING THE ABILITY OF SPEECH RECOGNIZERS TO MEASURE THE EFFECTIVENESS OF ENCODING ALGORITHMS FOR DIGITAL SPEECH TRANSMISSION

C. Michael Chernick, Stefan Leigh, Kevin L. Mills, and Robert Toense
National Institute of Standards and Technology
January 1999

ABSTRACT

Modern communication channels, such as digital cellular telephony, often convey human speech in a highly encoded form. Methods that rely on human subjects to evaluate the quality of such channels are too costly to deploy on a large scale; thus, automated methods are often used to model quality as perceived by humans. Traditional automated methods that use Signal to Noise Ratios (SNR) to judge the quality of channels do not model human perception well when applied to highly encoded speech. For this reason, researchers investigate alternative means to objectively measure the quality of such channels. In this paper we explore the feasibility and applicability of using automated speech recognition technology to model human perception of the quality of communication channels that carry highly encoded (compressed) human speech.

INTRODUCTION

Methods to measure the relative effectiveness of coding algorithms are necessary in order to compare competing approaches over a range of conditions. The most common method employs human listeners to grade perceived speech quality by assigning an opinion score from a subjective scale, typically consisting of five values from excellent to unsatisfactory [KOHL97, LI98]. While producing the desired comparisons, methods that depend on human subjects are too costly and time consuming to deploy on a large scale. For this reason, we seek new metrics for automatically evaluating the effectiveness of speech encoding algorithms. Such metrics must be objective, economical (in both time and money), and reflective of speech intelligibility as perceived by human listeners. This paper reports results from a preliminary investigation of the use of automated speech recognition technology as a means to evaluate coding algorithms for digital speech.

The paper is organized into seven sections. First, we discuss related work. Second, we present our motivation. Third, we describe our research methodology, and discuss the speech samples we used for the experiments. Fourth, we describe our experimental results with respect to both an automated speech recognizer and to human listeners. Fifth, we discuss the correlation between the performance of the speech recognizer and the perceptions of the human listeners. Sixth, we identify some future research related to our proposed evaluation method. Finally, we present our conclusions from the current experiments.

RELATED WORK

Traditional automated systems for measuring transmission channel quality employ signal-to-noise ratio (SNR) or segmental SNR (SEGSNR), or a frequency variant of SEGSNR [QUAC88]. While easy to measure and useful to assess selected encoding schemes, metrics based on SNR do not by themselves indicate the potential loss in recognition of compressed digitally encoded human voice signals. Quackenbush evaluated a wide range of objective measures that could possibly apply to vocoder-like systems. Most of these measures exhibited poor correlation with human perception.

Researchers continue to search for objective quality metrics that can be applied to vocoder-like systems. For example, Kubichek and others report results from investigating several such metrics proposed to the International Telecommunications Union (ITU) for standardization [KUBI91, KUBI92, BAYY96, LAM96, VORA95]. Other researchers investigate the possibility of measuring the quality of speech channels by transforming the channel input and output signals into an internal representation of the sound that a human would hear [BEER94, HANS97, PETE97, HAUE98]. While most proposed objective measures compare differences in input and output signals, Jin and Kubichek propose a metric based on comparing a quantized version of the output signal with a quantized version of a high-quality, reference signal [JIN96].

MOTIVATION

Most previous work on objective measures for speech quality seeks some easily measurable combination of parametric differences between channel input and output signals that can reliably predict how humans will perceive the quality of the output signal. In the vast majority of cases, subjective human perception is captured as a mean opinion score (MOS) that ranges from 5 (excellent) to 1 (unsatisfactory) [KOHL97, LI98]. As

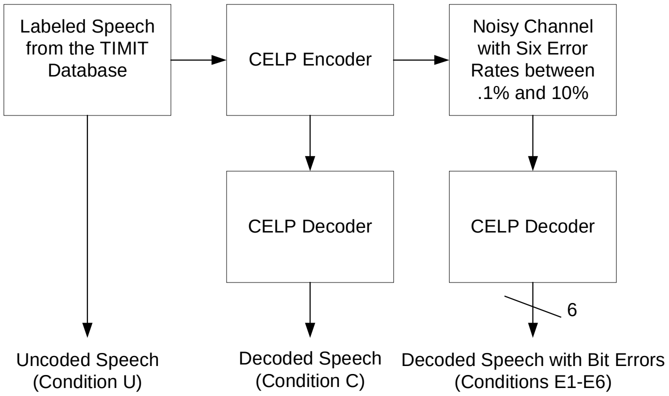
Labeled Speech from the TIMIT Database
CELP Encoder
Noisy Channel with Six Error Rates between .1% and 10%
CELP Decoder
CELP Decoder
6
Uncoded Speech (Condition U)
Decoded Speech (Condition C)
Decoded Speech with Bit Errors (Conditions E1-E6)

Figure 1. Method Used to Generate Speech Samples

discussed by Kubichek, the inherent variability of listeners and differing interpretations of the rating scale inhibits the reliability of MOS estimates [KUBI91]. These difficulties might compound as the number and granularity of scales to be scored increases. For example, Quackenbush measures subjective human perception for sixteen specific signal characteristics, where each characteristic can be distinguished on a 100-point scale [QUAC88]. This inherent variability might account for much of the variability reported by researchers who attempt to correlate objective measures with MOS. In thinking about this variability, we considered alternative approaches that might yield an objectively invariant result.

We found a test method that uses reference speech data and performance metrics to compare the performance of various speech recognizers [HTK97, GARF93]. Inverting this method, we wondered if speech recognizers might be an effective reference against which to measure speech quality on digital transmission channels. We foresee a quality score that might prove more reliable than MOS, yet still reflect differences in intelligibility as perceived by human listeners. Specifically, if a human subject were asked to transcribe the words from a speech segment and that transcription could be compared with the transcription generated by a speech recognizer for the same segment, then the correlation between human perception and automated objective measures could perhaps be made with greater reliability. If our ideas can be confirmed, then a new approach to objective measures for speech quality might prove feasible. Before proceeding to test our hypothesis, we decided to investigate how well a speech recognizer would perform as a predictor of MOS. This paper reports the findings of our initial investigation.

RESEARCH METHODOLOGY

Figure 1 illustrates the method used to generate speech samples for input to a speech recognizer, and for evaluation by human listeners. We selected nineteen speakers from the TIMIT database, a widely accepted database of labeled speech segments that has been used to evaluate speech recognizers, developed by researchers with funding from DARPA (Defense Advanced Research Projects Agency) [GARF93]. For our experiment we used the Code Excited Linear Prediction (CELP) algorithm to generate encoded speech samples [CELP]. Bit errors were generated using a Gaussian distribution.

Figure 2 depicts the general outline we used to score the speech samples, and then to assess the correlation between the recognizer results and human perceptions. The figure can be considered in three blocks: (1) automated scoring using a speech recognizer, (2) subjective scoring by human listeners, and (3) correlation analysis. We address each of these in turn.

Automated Scoring Using A Speech Recognizer. For the experiments reported here, we used a speech recognizer, HTK, readily available at NIST [HTK97]. The recognizer was included in the HTK Toolkit [RABI93]. In order to score the performance of the recognizer, we used a scoring package included in the toolkit. The scoring package generates several statistics, including correctly recognized phonemes, insertions, deletions, and substitutions. In this experiment, we used the number of phonemes correctly recognized, which we divided by the total number of phonemes in each speech sample in order to compute the percentage of phonemes recognized.

Subjective Scoring by Human Listeners. Since human listeners could not be asked realistically to identify phonemes in the speech samples, we recruited fourteen volunteers to listen to and then subjectively score the intelligibility of speech samples played through a loudspeaker. We ensured that two different volunteers listened to each sample; thus, the 98 input samples were doubled to give 196 test samples.

Correlation Analysis. We used correlation analysis to estimate how well the speech recognizer scores predicted the judgment of human listeners. We considered separately two classes of data: data based on speech samples from speakers one through eleven (used to train the recognizer) and data based on the other eight speech samples. Our results follow.

EXPERIMENTAL RESULTS

The first experiment applied a speech recognizer to the various speech samples generated. We were struck by the degree to which the speech recognition diminished for CELP-encoded speech samples, especially for speakers whose speech was used to train the recognizer. The performance of the speech recognizer differs significantly for the training speakers versus other

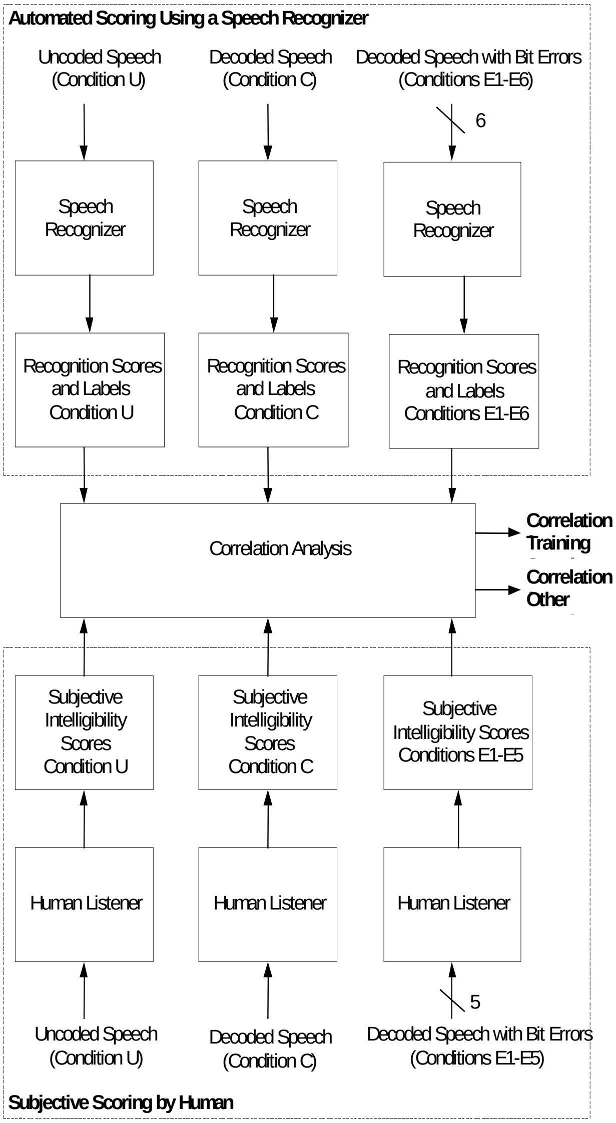
Automated Scoring Using a Speech Recognizer
Uncoded Speech (Condition U)
Decoded Speech (Condition C)
Decoded Speech with Bit Errors (Conditions E1-E6)
6
Speech Recognizer
Speech Recognizer
Speech Recognizer
Recognition Scores and Labels Condition U
Recognition Scores and Labels Condition C
Recognition Scores and Labels Conditions E1-E6
Correlation Analysis
Correlation Training
Correlation Other
Subjective Intelligibility Scores Condition U
Subjective Intelligibility Scores Condition C
Subjective Intelligibility Scores Conditions E1-E5
Human Listener
Human Listener
Human Listener
5
Uncoded Speech (Condition U)
Decoded Speech (Condition C)
Decoded Speech with Bit Errors (Conditions E1-E5)
Subjective Scoring by Human

Figure 2. Method of Scoring Speech Samples and Correlating Scores

speakers across all error rates. For this reason, we chose to separate these two classes of speakers for purposes of computing correlation with the human listeners.

The second experiment asked human listeners to evaluate a select, but substantial, subset of the speech samples generated. As expected, listeners judged the quality good to excellent for the unencoded speech and for CELP-encoded speech without errors. For CELP-encoded speech with .1% BER, the listeners found the speech understandable to good. As the BER reached .5%, listeners found the speech to be understandable but of poor listening quality. Understanding and quality dropped somewhat when the error rate rose to 1%. At the 2% error rate, listeners had difficulty understanding the speech samples. With a 5% BER listeners judged the speech to be practically unintelligible.

RESULTS FROM CORRELATION ANALYSES

The upper graph in Figure 3 plots human listener judgments against percent phoneme recognition from the speech recognizer for the eleven speakers used to train the speech recognizer. We computed the correlation at .816 ± .064 (2 stdevs). For the other eight speakers, as shown in the lower graph in Figure 3, we computed the correlation at .745 ± .074 (2 stdevs). The estimates of standard error were computed using resampling (bootstrap) [DIAC83]. To confirm these findings, we also computed correlation values using Spearman rank correlation [SACHS82]. The Spearman rank correlation for our training speakers is .789 and for the other speakers is .775. These values are the same whether the statistic is computed in the standard fashion as the correlation of the ranks of the scores, or whether an adjustment for ties in the scores is used. While the plots and correlation coefficients demonstrate clearly the monotone association between human and machine judgments of quality, the plots also display the inherent variation in the relation. Similar variation appears in other research the compares objective quality measures with human perception [BROO98].

FUTURE RESEARCH

Our next step is to compare the ability of commercial speech recognizers and human listeners to transcribe speech samples under the same conditions reported in this paper. While we expect human listeners to be superior to speech recognizers in all cases, if we can establish a relationship between the performance of human listeners and speech recognizers, then we can consider building and deploying a test system for automatically scoring speech coding algorithms. We foresee a system that enables developers to select speech samples from a database and to select from among a range of speech recognizers. The developer could also select from a range of error models and rates, including independent bit errors, alternating periods of good and bad channel signals, and various packet switching network properties. With such a test system, developers could explore the properties of proposed speech coding and decoding algorithms under a range of conditions.

Beyond the use of speech recognizers for automated scoring of network-based speech coding algorithms, we can imagine applying techniques emerging from image understanding research to develop similar test systems for image and video coding schemes used for network transmission. Of course, image understanding research is less well developed than speech recognition research. Still, edge-detection techniques and object-extraction techniques seem worth investigating for this purpose. The development of multi-media coding and transmission algorithms could be greatly accelerated by the ability to automatically score performance in a manner consistent with human perception.

CONCLUSIONS

We selected segments of speech from a widely accepted speech data base, and sent those segments through a speech recognizer under three conditions: (1) without encoding, (2) with encoding and decoding using a standard algorithm for speech compression, and (3) with encoding, transmission across a noisy channel, and then decoding. Speech recognition scores were computed for each speech segment under each condition. We then selected a subset of the speech segments, and asked human listeners to subjectively evaluate the intelligibility of the speech under the same conditions earlier input to the speech recognizer. We computed the correlation between the intelligibility of speech as evaluated by the automated recognizer and the human listeners. For unencoded speech segments used to train the recognizer, the correlation was .816 ± .064 (2 stdevs). For other unencoded speech segments, the correlation was .745 ± .074 (2 stdevs). Spearman rank correlation tests confirmed these numbers. These results are sufficient to encourage us to investigate the performance of commercial speech recognizers against human transcriptions. If the next phase of this research yields acceptable results, then construction of an automated evaluation system should be straightforward. Availability of an effective automated evaluation system will be useful to researchers and product engineers who are working toward advances in speech encoding algorithms for wireless communication channels and for Internet channels.

REFERENCES

[BAYY96] “Objective Measures for Speech Quality Assessment in Wireless Communications”, A. Bayya and M. Vis, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[BEER94] "A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation," J.G. Beerends and J.A. Stemerdink, *Journal of the Audio Engineering Society*, Vol. 42. No. 3, 1994.

[BROO98] "Getting the Message, Loud and Clear - Quantifying Call Clarity," S. Broom, P. Coackley, and P. Sheppard, *British Telecommunications Engineering*, Vol. 17, April 1998.

[CELP91] Federal Standard 1016, Telecommunications: Analog to Digital Conversion of Radio Voice by 4,800 Bit/Second Code Excited Linear Prediction (CELP), General Services Administration, Office of Information Resources Management, February 1991.

[DIAC83] "Computer-intensive Methods in Statistics", P. Diaconis and B. Efron, *Scientific American*, Vol. 248, pp. 116-130, 1983.

[GARF93] DARPA TIMIT Acoustic-Phonetic Continuous Corpus CD-ROM, John Garfolo, L.F. Lamel, William Fisher, John Fiscus, David Pallett, Nancy Dahlgren, NISTIR 4930, February 1993.

[HANS97] “Using a Quantitative Psycho-acoustical Signal Representation for Objective Speech Quality Measurement”, M. Hansen and B. Kollmeier, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997.

[HAUE98] “Application of Meddis’ Hair-Cell Model to the Prediction of Subjective Speech Quality”, M. Hauenstein, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998.

[HTK97] The HTK Book (for HTK Version 2.1), Steve Young, Julian Odell, Dave Ollason, Valthan Valtchev, Phil Woodland, Entropic Cambridge Research Laboratory Ltd., Compass House, 80-82 Newmarket Road, Cambridge CB5 8DZ, England, Tel: +44(0) 1223 302651 Fax: +44(0) 1223 324560, December 1997.

[JIN96] “Output-Based Objective Speech Quality Using Vector Quantization Techniques”, C. Jin and R. Kubichek, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[KOHL97] “A Comparison of the New 2400 bps MELP Federal Standard with Other Standard Coders”, M. A. Kohler, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1997.

[KUBI91] "Advances in Objective Voice Quality Assessment," R. Kubichek, et al, *IEEE Global Telecommunications Conference*, 1991.

[KUBI92] "Advances in Objective Voice Quality Assessment," R. Kubichek, et al, *IEEE 42nd Vehicular Technology Conference*, 1992.

[LAM96] “Objective Speech Quality Measure for Cellular Phone”, K.H. Lam, O.C. Au, C.C. Chan, K.F. Hui, and S.F. Lau, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[LI98] “Experimental Results on the Impact of Cell Delay Variation on Speech Quality in ATM Networks”, B. Li and X.R. Cao, *Proceedings of the IEEE International Conference on Communications*, 1998.

[PETE97] “Objective Speech Quality Assessment of Compounded Digital Telecommunication Systems”, K. T. Petersen, J. A. Sorensen, and S. D. Hansen, *Proceedings of the First Signal Processing Society Workshop on Multimedia Signal Processing*, 1997.

[QUAC88] Objective Measures of Speech Quality, Schuyler R. Quackenbush, Thomas P Barnwell, Mark A. Clements, Prentice-Hall, 1988.

[RABI93] Fundamentals of Speech Recognition, Lawrence Rabiner, Biing-Hwang Juang, Prentice-Hall, 1993.

[SACHS82] Applied Statistics: A Handbook Of Techniques, Lothar Sachs, Springer-Verlag, 1982.

[VORA95] "Perception-based Objective Estimators of Speech Quality," Stephen Voran, Connie Scholl, *Proceedings of the 1995 IEEE Workshop on Speech Coding for Telecommunications*, September 1995.

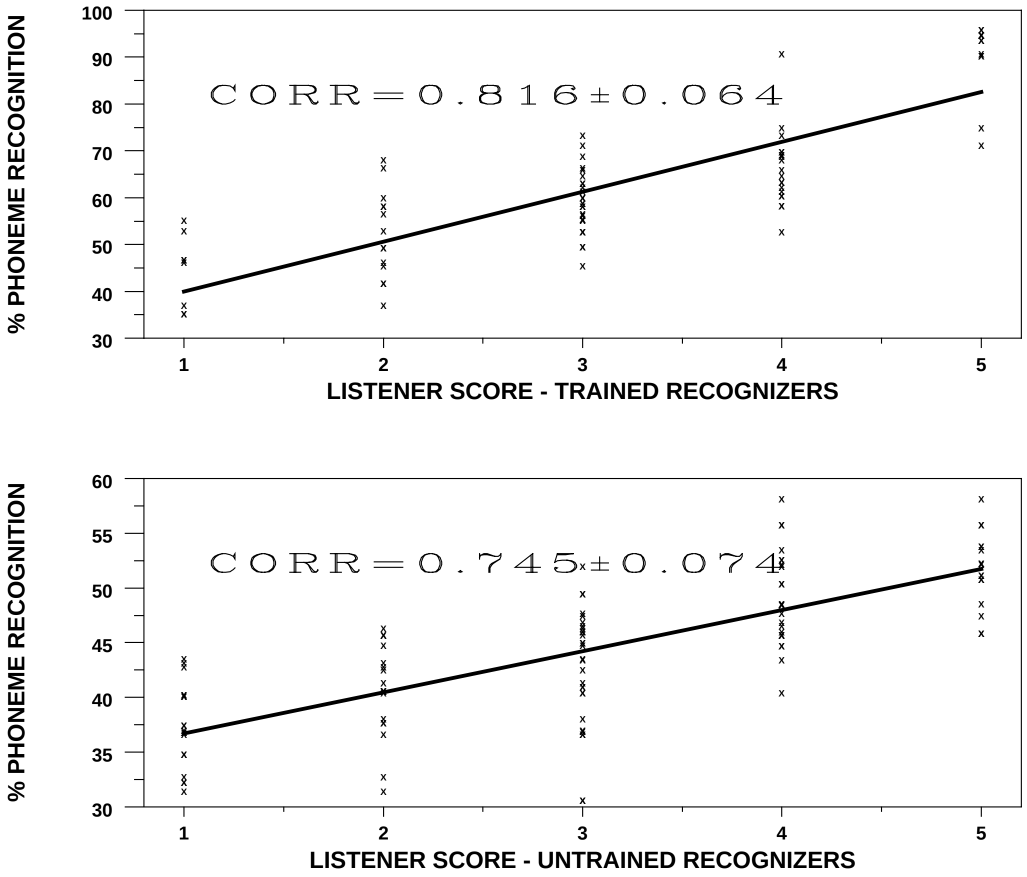

Figure 3. Correlation: Speech Recognizer and Human Listeners for Trained and Untrained Speakers