INTERACTION OF SPEECH DISORDERS WITH SPEECH CODERS: EFFECTS ON SPEECH INTELLIGIBILITY

D.G. Jamieson[1], L. Deng[2], M. Price[1] , Vijay Parsa[1] and J. Till[3]

[1]Hearing Health Care Research Unit, The University of Western Ontario, London, Ontario, Canada N6G 1H1; [2]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1; [3]Audiology and Speech Pathology Service, VA Medical Center, Long Beach, California 90822 USA

1. INTRODUCTION

Modern speech coding schemes have been developed to address the demand for economical spoken language telecommunication of acceptable quality. A variety of speech coding algorithms have been described, which compress speech to facilitate efficient transmission of spoken language over communication networks [2,3,4]. Most such speech coding algorithms are lossy in the sense that the "processed"[1] speech is not identical to the original speech. As a result, some distortion is invariably introduced with any lossy speech coding strategy. For this reason, candidate coders undergo detailed evaluation to ensure that the associated speech output is of acceptable quality [1].

Potential limitations of previous evaluations of the performance and acceptability of coding techniques are the restriction of the speech dataset used in the evaluation to normal speech and the restriction of the listeners to persons who have normal or superior hearing. Assumptions regarding the properties of the speech input and of the listener may be violated when the speaker has some form of speech or voice disorder and/or when the listener has some form of hearing loss. In such circumstances, coder evaluations based on studies with normal speech and/or hearing may not generalize to the performance achieved when the talker / listener possesses some form of communication disorder. Specifically, disordered speech input may interact with the coding scheme to reduce the intelligibility and/or quality of the speech at the output of the transmission system. The present study examined the interaction of several alternative coding schemes, using a database derived from a corpus of speech data collected from persons with a range of speech production disorders, as well as from normal talkers.

Objective measures of speech quality are intended to quantify this distortion. Finding a good objective speech quality measure that correlates well with the subjective judgement of speech quality is important in speech coding applications for a variety of reasons: (a) it could become an integral part of a speech coding strategy providing the criteria for selecting an optimal coder, and (b) it could be used in place of subjective quality tests which are costly and very time consuming [1,2]. Compared to the subjective quality measures, objective measures are more reproducible and less expensive to administer.

[1] The word "processed" speech refers to the speech signal which has gone through encoding and subsequent decoding processes.

Objective measures of speech quality are commonly based on the Signal-to-Noise Ratio (SNR) parameter or some metric distance between the original speech and the "processed" speech in terms of spectral, Linear Predictive Coding (LPC), cepstral coefficients.

The "processed" speech is first time-aligned and subtracted from the original speech waveform. For the traditional measure of SNR, the energies of the "signal" and the residual "noise" components are computed over the whole speech waveform. To compute the segmental SNR (SSNR), the speech data are segmented into frames and the frame-by-frame SNR is computed and is averaged over the number of frames. Another measure, the Frequency-Weighted SSNR (FWSSNR), is calculated by first weighting the speech spectrum in each frame with a set of perceptual weights and then computing the resulting signal and noise energies. The FWSSNR tends to be the best predictor of perceptual ratings, followed by SSNR measures [1]. However, SNR measures are meaningful only for coding schemes where the "noise" is deemed to be additive. These measures are thus useful with subband coders and transform coders, but not with vocoders and vocoder-like systems which introduce speech-correlated noise components [1,2]. In addition, these coders focus only on the magnitude of the speech spectrum, as human auditory system is relatively insensitive to phase distortion [2,3]. Thus the "processed" speech can be quite different from the original yet still be perceived well.

For the latter type of speech coders, objective measures based on the spectral and LPC parameters are more valid. A simple frame-by-frame L_2 norm spectral measure can be calculated as the square root of the mean squared difference, across M frequencies and n frames, of $S(n,\theta)$, the original speech spectrum, and $\overline{S}(n,\theta)$, the "processed" speech spectrum. This linear spectral distance measure has been reported to correlate .38 with subjective quality judgements [1].

Instead of the linear spectral distance, one can calculate the same L_2 norm measure using the log spectra. In addition, spectral weighting to match ear's critical bands and L_p norm can also be applied. A slight variation of the linear spectral distance measure, involves applying an exponential weighting to the spectral estimates; using an exponent, δ, having a value of 0.2, was found to correlate 0.61 with subjective results [1] while the log spectral distance correlated 0.60.

Similar L_2 norm metric measures can be calculated using the LPC and cepstral coefficients. The basic linear L_2 norm LPC measure can be calculated as the mean squared difference of the LPC coefficients extracted from the n*th* frame of the original and "processed" speech respectively. In a similar manner, distance measures based on

PARCOR (or reflection) coefficients and cepstral coefficients can be calculated. Two important measures derived from the LPC coefficients are the Log Area Ratio (LAR) measure and the Itakura-Saito measure. The LAR measure is defined as the square root of the mean square, across frames, of 20 log the ratio of the area functions of the original and processed speech in each frame. The Itakura-Saito measure is a widely used speech quality measure which is the ratio of the residual energies produced by the original speech when inverse filtered using the LPC coefficients derived from original and "processed" speech. The Itakura-Saito measure is very sensitive to the spectral mismatch in formant locations and is less affected by the mismatch in spectral valleys [2]. This is desirable as the human auditory system is more sensitive to errors in formant location and bandwidth.

Quackenbush *et al.* [1] investigated the correlation between the abovementioned objective measures and the subjective quality assessments. Among LPC based measures, the LAR measure exhibited the highest correlation with subjective judgements (r=0.62) followed by the Itakura-Saito measure (r = 0.59). The linear LPC measure was the least reliable predictor of subjective judgements (r= 0.06).

A clear potential concern is that the literature examining the relation between subjective and objective measures of speech quality has heretofore been restricted to speech data gathered from normal talkers. The present study was designed to investigate the potential interactions between digital coding algorithms and talker, by including speech samples from talkers with common speech disorders in our data set.

Three different coding algorithms were investigated relative to unprocessed speech: the Codebook Excited Linear Prediction (CELP), the Global System for Mobile communications (GSM) algorithm which is a standardized speech coding algorithm in Europe, and the Linear Predictive Coding (LPC) algorithm. The specific coding schemes evaluated were MatLab implementations of NSA FS-1015 LPC-10e; NSA FS-1016 CELP-v3.2; and ETSI GSM [5]. One of the goals of this study was to quantify the coding distortion using the objective measures described above and to correlate these measures with speech intelligibility and subjective quality data, in the hope of identifying one or more measures that can predict the subjective results.

2. METHODS

2.1 Speech Database

Speech samples were recorded from clients of the Department of Speech Pathology, Veterans Administration Hospital, Long Beach, California. Our signal database included continuous speech, 21 consonant targets in a syllable-medial, fixed-vowel environment, and 15 vowel targets in /hVd/ and /bVd/ environments. The database included talkers with a range of speech production disorders, including dysarthria, apraxia, aphasia, hypernasality, and breathiness.

Signals were recorded as 16-bit samples at 48 kHz, on Digital Audio Tape (DAT). Consonant targets were the 21 English-language consonants (/b, tʃ, d, g, h, ,j, k, l, m, n, p, r, s, ʃ, t, θ, ð, v, w, y, z/) spoken within an /aCil/ context, following the UWODFD format [6]. Vowels were the 15 English vowels (i, *i*, ɛi, ɛ, æ, a, u, ɔ, o, ɷ, ʌ, e^r, æu, ai, oi), spoken in each of the /hVd/ and /bVd/ contexts (e.g., heed, hid, hayed, head, had, hod, hawed, hode, hood, who'd, hud, heard, howd, hide, hoyd). The continuous speech target was the Rainbow Passage.

During the recording of consonant and vowel targets, the patient was asked to repeat each word after the experimenter. If the client had particular difficulty saying a certain word, recording progressed to other items, with the difficult item reintroduced at a later time. In some cases, the client was unable to produce one or two of the words. For the Rainbow Passage, the client was asked to read the passage out loud. If reading the passage proved too difficult, the experimenter read the passage aloud phrase by phrase allowing the patient to repeat each phrase.

Each recorded DAT file was converted to a 16 bit/48 kHz .WAV format file and stored on computer disk using a Zefiro Acoustics Model ZA1 Sony/Philips Digital Interface Card and software together with a Sony PCM-2000 digital audio recorder and a Gravis UltraSound 16-bit sound card. The 16 bit 48 kHz .WAV files were then digitally edited to isolate the target utterance. Each of the edited .WAV files was then input to each of the speech coders, with the output stored as a 16 bit/8 kHz file in the CSRE [7] .ADF format.

2.2 Behavioral Data Collection Methods

A variety of perceptual experiments were conducted with these data samples using normal-hearing listeners:

1. a closed-set consonant identification task using English-language consonants in the /aCil/ environment;
2. a closed-set vowel identification task using English-language vowels in both the /hVd/ and /bVd/ environments;
3. a paired-comparison quality rating task using various disordered speech samples, processed with different coding schemes;
4. a quality rating task using a multi-factor rating scale.

The experiment generator from CSRE 4.5 [7] was used to generate listening tests to evaluate the intelligibility of the utterances processed through the speech coders, in relation to the intelligibility of the unprocessed words. Data are reported here on three tests for a series of talker, one test for the consonants in /a_il/ context, and one each for the vowels in /h_d/ and /b_d/ context, respectively.

Listeners were ten young, native English-speaking adults with normal hearing. Testing was conducted individually in a double-walled sound-attenuating chamber. After hearing each test item listeners used the mouse to select the word they thought had been presented, from the list displayed on the computer screen.

Listeners were trained with each task by presenting the full set of unprocessed utterances once to familiarize them with the nonsense words being used and to orient them to the layout of the response items on the screen in relation to target items.

Each listening task consisted of a series of test sessions with the set of target items and talker fixed within session. Each session consisted of four blocks, one each for the unprocessed speech and for the three processed-speech conditions. Within each session, these blocks were presented in a randomly determined order for each listener.

3. RESULTS

Listeners with normal hearing have significantly more difficulty understanding disordered speech when this has been processed using certain coding schemes and they judge such speech to be of lesser quality than the original disordered speech. Our listeners consistently performed best with the original unprocessed utterances, with the GSM coder yielding the highest intelligibility among the three coders. LPC consistently yielded the lowest intelligibility for all listeners and all talkers. However, the magnitude of the differences observed between processing techniques is different for different talkers. Figures 1 and 2 summarize the results obtained for a sample of talkers, both normal and with disordered speech, after various types of speech processing.

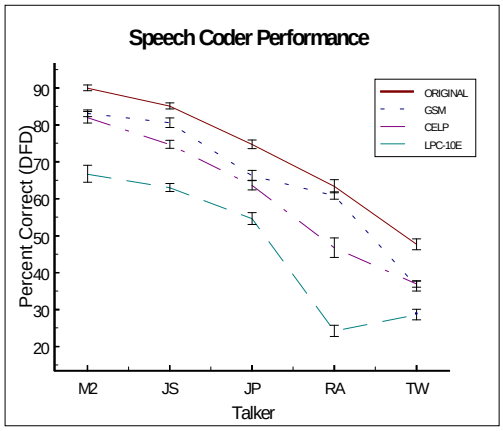

Figure 1. Intelligibility of syllable-medial consonant targets for each talker, as a function of the coding scheme applied to the speech.

For normal male talkers, unprocessed speech yielded identification levels averaging 87 to 90% correct performance for consonant targets. Accuracy declined to approximately 83% correct for the GSM coder, falling further to between 74% and 82% for CELP, and then to 51 to 67% for LPC.

For talkers with disordered speech, there was a clear interaction between talker and coding technique. For talker JS (hypernasal and breathy) consonant intelligibility was between 85% correct for the unprocessed samples, falling to 81% for the GSM, 75% for CELP and just 63% for LPC. For talker RA, (hypernasal and breathy) consonant intelligibility was between 63% correct for the unprocessed samples and 61% for the GSM, but fell to 47% for CELP and all the way to 24% for LPC speech.

For talker TW (dysarthria), consonant intelligibility was just 48% for unprocessed speech, falling to 36% for GSM and CELP and to just 29% for LPC speech. For male talker JP (aphasia and dysarthria) consonant identification accuracy was 75% correct for unprocessed speech, 66% for GSM processed speech, and 64% for CELP, falling to 55% for LPC speech.

These differences in both overall level of performance across talkers and in the interaction of talker with type of coder show striking differences for listeners having the same disordered speech classification. For example, notice the substantial decline in accuracy with LPC speech for talker RA vs JS, both of whom were classified as being hypernasal, with breathy speech.

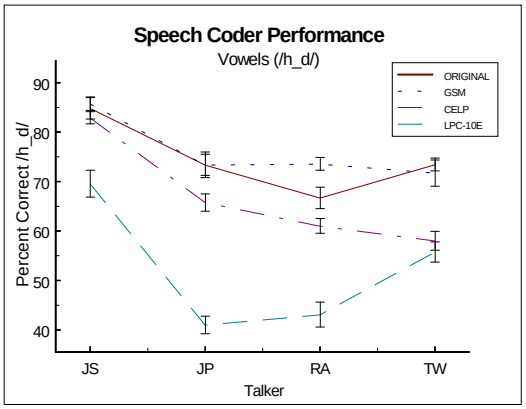

Figure 2. Intelligibility of vowel targets for each talker, as a function of the coding scheme applied to the speech.

Different objective measures as explained in the beginning of the document were computed and one particular measure, *viz.* the Itakura-Saito (IS) measure for the consonant set is shown in Figure 3. The "processed" speech at the output of each coding algorithm is time-aligned with the original speech waveform and the frame-by-frame IS measure was calculated using the LPC coefficients. This measure is then averaged over all the frames to obtain a single IS measure for each of the coders.

A few interesting conclusions can be drawn from the comparison of objective and subjective results:

1. The rank ordering of coders obtained by applying the IS measure matches that obtained with subjective measures: GSM is better than CELP, which in turn is better than the LPC. Note that the objective measures compute a **distance** measure representing the distance between the "processed" speech and the original speech.
2. The strong, sloping trend obtained in the subjective results (Fig. 1) is not reflected by the IS measure,

nor in any of the other objective measures we have studied to date.

3 Fig. 1 shows that the intelligibility of original and GSM processed speech are highly similar for talker RA. This is also reflected in the IS measure, where GSM-processed speech for talker RA has the least IS measure.

4. LPC-processed speech from talker RA is substantially less intelligible than LPC processed speech from talker TW (Fig. 1). For GSM and CELP coders, the situation is reversed. This interaction is also captured by the IS measure.

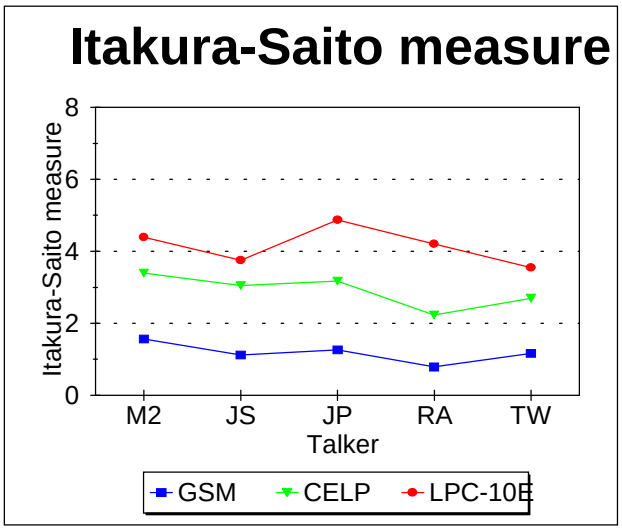

Figure 3. Predicted speech quality for each coding scheme, based on the Itakura-Saito (IS) measure.

4. DISCUSSION

These results confirm that the selection of coder has a substantial impact on the intelligibility of the resulting speech signals. For our normal talkers, the intelligibility of essentially open-set consonant identification in a fixed syllable-medial context varied from 90% correct with unprocessed speech to just 51% correct with LPC speech. For all our listeners, the rank ordering of performance for speech from our normal talkers was as follows: unprocessed speech, GSM-coded speech, CELP-coded speech, and LPC-speech.

For speech from our sample of talkers having a speech disorder, consonant intelligibility varied from a high of 85% correct for unprocessed speech from talker JS to just 24% for LPC-processed speech from talker RA (hypernasal and breathy speech). The rank ordering of the mean consonant intelligibility across our listeners was unprocessed speech > GSM speech > CELP speech > LPC speech, for all talkers.. However, there are indications that the type of coder used did interact with the details of speakers voices, for at least some of our disordered talkers. For example, Figure 1 shows that for talker RA, LPC speech was substantially less intelligible than CELP speech, while for the other disordered talkers, LPC speech was only marginally less intelligible. The IS objective measure correlates quite well with the subjective data in terms of the ranking the speech coding systems. However, this and the other objective measures fail to capture the detailed features of the intelligibility data, probably due to the fact that the objective measures fail to consider the intelligibility and /or quality of the original speech input signal. Future research may usefully involve the incorporation of a parameter reflecting this factor, into the computation of the objective measures.

5. REFERENCES

1. Quackenbush, S.R., Barnwell T.P., III, and Clements, M.A. *Objective Measures of Speech Quality.* Prentice Hall Advanced Reference Series, Englewood Cliffs, NJ, 1988.

2. Deller Jr., J.R., Proakis, J.G. and Hansen., J.H.L. *Discrete-time Processing of Speech Signals.* Macmillan Publishing Company, NY, 1993.

3. Papamichalis., P.E. *Practical Approaches to Speech Coding.*, Prentice Hall Inc., Englewood Cliffs, NJ, 1987.

4. O'Shaughnessy, D. *Speech Communication: Human and Machine.*, Addison-Wesley Publishing Company, NY, 1987.

5. Spanias, A. MatLab implementations of the GSM, NSA CELP3.2, and LPC-55 C Coders. Arizona State University, 1995.

6. Cheesman, M.F. and Jamieson, D.G. "Development, evaluation and scoring of a nonsense word test suitable for use with speakers of Canadian English", *Canadian Acoustics, Vol., 24, 1996, p 3-11.*

7. *Computerized Speech Research Environment (CSRE) Version 4.5 (1996)*, Avaaz Innovations Inc., London, Canada, 1996.

Acknowledgements: This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.