METHODS FOR OBJECTIVE DETERMINATION OF TELEPHONE SPEECH QUALITY FOR THE HEARING IMPAIRED

J. van Zyl and J.J. Hanekom
Department of Electrical, Electronic and Computer Engineering,
University of Pretoria, 0002 PRETORIA, South Africa
joey@tuks.co.za || j.hanekom@postino.up.ac.za

***Abstract*—Perceived speech quality is most directly measured by subjective listening tests. These tests are often time-consuming and expensive, and numerous attempts have been made to supplement them with objective estimators of perceived speech quality. PESQ is the objective speech evaluation method that performs the best when compared to subjective listening tests for narrow band telephone systems. The PESQ algorithm is the industry standard after becoming the ITU-T P.862 recommendation for end-to-end speech quality evaluation of narrowband telecommunication systems. No extensive research has been done to improve on these methods to include the hearing impaired. This paper discusses what needs to be done to extend the PESQ algorithm so that it can accurately predict how well an individual with a cochlear implant will hear on a telephone system.**

***Key Words*—Auditory System, Speech Quality Evaluation, Perceptual Speech Quality, Transmission Quality, Hearing Impaired, Cochlear Implants**

I. INTRODUCTION

In order to measure speech transmission quality in a telephone network on a regular basis, it is necessary to avoid the complicated and expensive procedure of subjective determination, and to use objective systems instead. Today there are several systems and methods in use to evaluate perceived sound quality without the need for human listeners. The last 20 years have seen speech evaluation methods become more standardized in transformations they implement and the accuracy they deliver. Earlier methods could only predict the perceived sound quality of speech codecs. Now standards such as the ITU-T P.862 are available for the objective determination of end-to-end speech quality of a telecommunication system. These standards however encompass the hearing of the general population and lack adequate accuracy regarding the hearing impaired.

Individuals fitted with cochlear implants have extremely limited speech perception because of the damage done to their inner ear. Because of this, these individuals can hardly comprehend any speech when using a standard telephone. If their hearing could be predicted on a telephone system, the development for proper technology to improve their telephone use would be less time consuming and expensive.

Section 2 of this paper gives an overview of current subjective and objective testing methods. Section 3 will look at the proposed extension that needs to be made to improve on the current testing methods to include the hearing impaired.

II. OVERVIEW OF EXISTING OBJECTIVE SPEECH EVALUATION SYSTEMS

A. Subjective Measurements

In subjective testing, speech materials are played to a group of listeners, who are asked to rate the speech they just heard. The ratings are then gathered and averaged to yield the final score. The ITU-T Recommendation P.800 defines procedures for the subjective determination of transmission quality of a telephone system [1]. The most popular test, Absolute Category Rating (ACR) is where listeners listen to simple sentences over a telecommunications system and rate the quality of the speech they just heard without hearing the original signal. The answers must be given in a scale from 1 to 5, 1 being bad, 2 - poor, 3 - fair, 4 - good, and 5 representing excellent quality. The answers given by the listeners are defined as their Mean Opinion Score (MOS) [1]. Since these tests are subjective they vary greatly from person to person. Therefore to obtain a satisfactory estimation of the sound quality of a telecommunication system a large number of people need to be utilized which can become time consuming and inefficient.

B. Objective Measurements

On the other hand are objective methods which are easier to implement, less time consuming, and less expensive [2]. Numerous objective speech quality methods have been proposed and used for evaluations of speech coding.

Because the human hearing and recognition system is highly non-linear and by far not completely understood today, we cannot analytically predict the human perception of the quality of a speech signal being transmitted through a network. However, it is clear that there are objective (physically measurable) factors as well as inter- and intra-individual aspects that we can use in objective speech algorithms. Therefore, a quantitative expression of speech quality will always be a statistical mean value. The averaging is not limited to the objectively measurable factors but also includes a "mean physiological and psychological sensitivity" of human beings [3].

The signal processing within objective methods based on the comparison of speech samples can be structured into three major steps: pre-processing (time-alignment), psycho-acoustic (perceptual) modeling, and speech quality estimation (cognitive) model [4].

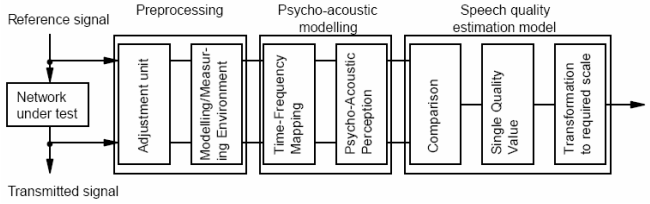

Figure 1. Building Blocks for Speech Quality Measurement Systems. [3]

The best sound evaluation methods are intrusive perception methods that compare the original signal with the degraded signal. (Non-intrusive systems are also available but not as accurate.) The basic common building blocks for intrusive methods can be seen in figure 1.

In the pre-processing stage the two signals are time aligned, intensity aligned, and filtered [5].

In the psycho-acoustic modeling stage the human auditory system is modeled. This mapping to the perceptual domain involves time-frequency mapping and psychoacoustic alterations to the signal.

Masking represents one of the most basic effects in human psychoacoustics according to Fastl in [6]. In the ear if two sounds are very close to each other in the time domain or in the frequency domain, either the second or the softer of the two sounds is masked out by the other sound. The way masking is tested is by determining the audibility of pure tones in the presence of masking sounds. (See [9,10] for equations used for frequency and temporal masking.)

Loudness is a dominant feature for sound-quality evaluation, and one used in almost all perceived sound quality features. This stems from the fact that humans perceive loudness differently than what would be expected from sound intensity levels. To be precise for the same intensity of sound, traditionally measured in decibels, across a range of frequencies the perceived loudness of the sound will change. Zwicker's law is used most commonly for the loudness transformation; this is shown in equation 1.

$$LX(f)_n = S_1\left[\frac{P_0(f)}{0.5}\right] \times\left\{\left[0.5+0.5\frac{PPX'_{WIRSS}(f)_n}{P_0(f)}\right]^{Y}-1\right\} \quad (1)$$

Perceptual Evaluation of Speech Quality (PESQ) is the method that is recommended by the ITU-T in P.862 [7]. This algorithm implements the above transforms in its auditory model to change the input sound to simulate how the human ear interprets it. PESQ currently has the highest correlation to subjective tests for normal hearing individuals. In 38 speech evaluation tests in mobile, fixed-line, and VoIP and multi-networks PESQ had above 90% correlation with the subjective tests [8]. What makes the PESQ algorithm better than the rest is its accurate time-delay compensation [9] and it's well-defined auditory transform (psychoacoustic model) [10]. After these two steps a cognitive model was added which evaluates the audible errors in the output signal by counting the noise disturbance for individual time-frequency cells. These two values are combined to produce a predicted MOS score between 1.0 and 5.0. These scores are meant to correlate with the subjective scores given by listeners in an ACR test. Figure 2 shows the basic outline of the PESQ method.

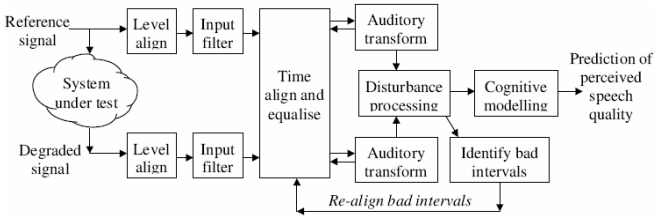

Figure 2. Basic layout of PESQ Algorithm[7]

C. Background of Cochlear Implants

In normal hearing, the outer ear picks up acoustic pressure waves. The middle ear then converts these waves to mechanical vibrations; a number of small bones are responsible for this conversion. The mechanical vibrations are transformed to vibrations in fluid in the inner ear, the cochlea. When the fluids of the cochlea undergo pressure variations, the basilar membrane undergo displacements which contain information about the frequency of the acoustic signal. Hair cells, which are attached to the basilar membrane, are bent related to the displacement of the membrane. An electrochemical substance is released when the hair cells are bended; this substance causes neurons to fire, indicating that there is excitation in the inner ear at a specific place. Information about the acoustic signal is transmitted to the brain via the central nervous system which is connected with the auditory nerves.

The auditory system has no way of transforming acoustic pressure waves (sound) to neural impulses when the hair cells are damaged. This is one of the causes of hearing impairment. The hair cells can be damaged in various ways, including diseases such as meningitis and Meniere's disease, congenital disorders and some drug treatments. The auditory neurons can degenerate as a result of the damaged hair cells. If a person has a large number of damaged auditory nerves or hair cells in the cochlea, he/she is diagnosed as profoundly deaf [11].

Cochlear implants are meant to stimulate the auditory nerve electrically to bypass the defects in the patient's ear. A cochlear implant consists of an array of electrodes that is planted into the cochlear duct of a patient. The patient has an external processing unit that transforms incoming sounds to acceptable frequency bands that can be interpreted by the

brain. The cochlea works like a spectrum analyzer that picks up different frequencies along the cochlear duct. The sound information goes through band-pass filters in the external processing unit, and each of these bands is sent to one of the electrodes in the array in the cochlea.

III. Research Gap

From current literature it is apparent that their does not exist an objective perceived sound quality evaluation method that gives an indication of how individuals that are hearing impaired hear on telephones. Governments require that telephone network companies to make its services available to as many people as possible including disabled individuals. Although the ITU-T has various specifications on equipment, they do not give any tools to help the development of new technology to make telephone systems more effective for hearing impaired people. The contribution of this research project will give network developers and other media developers a means of reducing the cost on testing with respect to hearing impaired persons. This will serve as an addition to the various objective speech evaluation methods already in use for the development of technology for normal hearing persons and benefit the development of tools for individuals who are hearing impaired.

IV. Objectives of Current Research

The overall objective of this study is to extend on these speech quality methods with regard to the hearing impaired. Techniques will be developed to assess the speech quality of the information received by hearing impaired listeners and psychoacoustic experimental work will be conducted with these listeners. In order to achieve these goals a model of speech reception in the auditory system, that incorporates hearing impairment, will be developed and incorporated with current speech evaluation knowledge to form a new objective speech measurement system. Stemming from this, research will also be done on what improvements must be made at the output of telephones to improve the transmission quality of telephones to benefit individuals with cochlear implants.

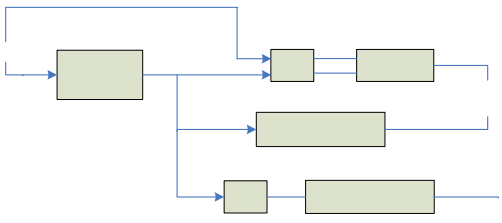

Figure 3. Block diagram of Research.

The system in development to predict the MOS score of a hearing impaired is shown in figure 3. This system will use the existing PESQ algorithm and do speech processing on its input (Transformation T in figure 3.) The new *model T* is placed between the original and degraded inputs of PESQ and the output of the telephone channel. The model T does an auditory transformation that simulates the transformation that takes place in the cochlear implants.

V. Cochlear Implant Processing

The two signal processing strategies that are most widely used in cochlear implants currently is the SPEAK strategy and the CIS strategy. These two strategies will be implemented in the transformation T to imitate the sound conversion that takes place inside the cochlear implant. The degraded input that comes out of the telephone system to the PESQ algorithm will therefore be processed to sound same as what an individual with a cochlear implant will hear. The strategies implemented will be discussed next.

A. CIS Strategy

The CIS strategy was originally designed to address the problem of channel interactions when stimulating all the electrodes simultaneously. The CIS approach uses pulses which are non-simultaneous and interleaved to stimulate nerve cells. The nerve cells are stimulated by sending biphasic pulse trains to the electrodes, i.e. only one electrode is stimulated at any given time. The pre-emphasis filter is used to attenuate the low frequencies such that the frequency band has equal loudness across the whole speech spectrum [12]. The signal is now passed through a bank of band-pass filters, the number of filters depend on the number of channels used. To extract the envelopes of the filtered waveforms, full-wave rectification and low-pass filtering is used, the typical cut-off frequency for the low-pass filter is 200 or 400 Hz. The outputs of the filters are compressed and used to modulate biphasic pulses [11]. The logarithmic compression of the signal depends on the particular patient's dynamic range of electrically evoked hearing. The amplitudes of the trains of balanced biphasic pulses are proportional to the envelopes of the processed waveforms. The pulses are then delivered to the electrodes at a constant rate and in a sequential manner.

The rate of stimulation has a significant influence on speech recognition. The number of pulses per second (pps) that is used varies from patient to patient. Some patients achieve optimum performance with 833 pps while others achieve optimum performance with 1365 pps. Pulse rates vary from as low as 100 pps to as high as 2500 pps. [13] reports stimulation rates that vary between 900 and 2400 pps.

B. SPEAK Strategy

The SPEAK strategy is an "*n-of-m*" strategy, where the speech signal is filtered into m frequency bands and the processor selects the n ($n < m$) outputs with the largest energy in the envelope. Only the n electrodes corresponding to these selected outputs are then stimulated.

A pre-emphasis filter is used similar to that of the CIS strategy to produce speech that is equal in loudness across the frequency spectrum. The speech signal is then filtered into 20 frequency bands with center frequencies ranging from 250 Hz to 10 kHz. The outputs of the filters are

rectified and low-passed filtered (cut-off frequency of 200 Hz). The SPEAK processor now selects a number of maxima at 4 ms intervals to modulate the amplitude of the stimulating pulse train. The number of maxima varies between 5 and 10, with an average of 6. "Maxima" does not necessarily refer to the spectral peaks within the signal, but to the largest amplitudes of the filtered waveforms.

The electrodes are organized according to the tonotopic order within the cochlea, each output of the bandpass filters is allocated to a specific electrode. For example, the most apical electrode corresponds to the output of the filter with the lowest center frequency. The stimulation rate of the electrodes varies between 180 pps and 300 pps [11], a rate of 250 pps is reported in [13].

The rate of stimulation depends on the number of selected maxima as well as the particular patient's parameters. When more maxima are selected for broader spectra, the stimulation rate needs to be reduced. Temporal information is increased when the spectral content is reduced and the stimulation rate increased. Note that there is a trade off between spectral content and temporal information.

These transforms will be implemented in Matlab and fed into the PESQ algorithm which is written in C, and freely available from the ITU-T.

VI. Future Work

With a speech evaluation algorithm that actively predicts how well an individual using a cochlear implant can recognize speech on a telephone system, further work can be done to make telephones accessible to individuals who find it hard to hear on a telephone. Functional block V (in figure 3) will be looked at that will improve the quality of hearing for these individuals. This might be a different coding scheme on the cochlear implant so that the signal processing in the cochlear implant will correlate better with that of a telephone system. It could also be a telephone handset sold to these individuals that will boost certain frequency components for better quality for these people.

VII. Conclusion

By modeling the sound transformations made by a cochlear implant and feeding this as an input to an objective speech evaluation algorithm will allow a more accurate estimation of how well the hearing impaired hear on a telephone. The goal is to create a new algorithm that can be added to the P.862 recommendation for objective speech evaluation for the hearing impaired. This will hopefully inspire telecommunication manufacturers to include the hearing impaired when developing new equipment to make telecommunications available to more people.

VIII. References

[1] ITU-T Recommendation P.800, "Methods for Subjective Determination of Transmission Quality," International Telecommunications Union, Geneva, Switzerland,Aug.1996.

[2] B. Somek, J. Herceg, and M. Maletic, "Speech quality assessment," Proceedings of the 46th International Symposium of Electronics in Marine, IEEE, 2004, pp. 307-312.

[3] ETSI EG 201 377 v1.2.1, "Specification and measurement of speech transmission quality. Part 1: Introduction to objective comparison measurement methods for one-way speech quality across networks," European Telecommunications Standards Institute, Dec.2002.

[4] L. Thorpe and Y. Wonho, "Performance of current perceptual objective speech quality measures," Proceedings of the IEEE Workshop on Speech Coding, 1999, pp. 144-146.

[5] ITU-T Recommendation P.861, "Objective quality measurement of telephone-band (300–3400 Hz) speech codecs," International Telecommunications Union, Geneva, Switzerland,1996.

[6] H. Fastl, "The Psychoacoustics of Sound Quality Evaluation," *Acta Acustica*, vol. 83, 1997, pp. 754-764.

[7] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," International Telecommunications Union, Geneva, Switzerland,2001.

[8] T. Goldstein and A. W. Rix, "Perceptual speech quality assessment in acoustic and binaural applications," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2004, Vol 3, pp. 3-7.

[9] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment. Part I - Time-delay compensation," *AES: Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 755-764, Oct.2002.

[10] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment. Part II - Psychoacoustic model," *AES: Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 765-778, Oct.2002.

[11] P. C. Loizou, "Signal processing techniques for cochlear implants," *IEEE Engineering in Medicine and Biology*, vol. 18, no. 3, pp. 34-46, 1999.

[12] W. M. Hartmann, *Signals, Sound, and Sensation*, Springer Science+Business Media ed. New York: 1998.

[13] M. W. Skinner, P. L. Arndt, and J. S. Staller, "Nucleus 24 advanced encoder conversion study: performance versus preference," *Ear and Hearing*, vol. 23, no. 1, pp. 2S-17S, Feb.2002.

Joe van Zyl obtained his B.Eng degree in Computer Engineering from the University of Pretoria in 2003. He is currently busy with his Masters degree in Bio-Electronic Engineering at the University of Pretoria. His interests include sound quality and speech processing, and he aims continue specializing in the human auditory system for improving sound quality for individuals with extreme hearing impairment.

CONTACT DETAILS:

Name:	Joe van Zyl
Affiliation	M.Eng Student, University of Pretoria
Address:	94 Jacobson Str, Lynnwood Ridge, Pta, 0081.
Email:	joey@tuks.co.za
Telephone:	012-420 2647 (W)
	082-494 7496 (C)
	012-352 6000 (F)