ON THE PREDICTION OF SPEECH QUALITY RATINGS OF TRACHEOESOPHAGEAL SPEECH USING AN AUDITORY MODEL

Rob McDonald, Vijay Parsa, Philip Doyle

School of Communication Sciences and Disorders, and Department of Electrical and Computer Engineering. University of Western Ontario, London, Canada

Guo Chen

School of Information Technology and Engineering. University of Ottawa, Ottawa, Canada

ABSTRACT

Total laryngectomy is often the treatment of choice for many patients suffering from laryngeal cancer. This procedure alters the speech production mechanism, and tracheoesophageal (TE) speech is an alternative where the pulmonary air is forced through the esophagus. TE speech is often characterized by poor intelligibility and voice quality. Acoustic analysis of TE speech has the potential of quantifying the voice quality and assisting the speech pathologist in determining and monitoring the therapy process. In this paper, we apply two different methods for predicting the voice quality ratings of TE speakers by naive listeners: (a) conventional spectral and linear prediction measurements that were investigated in earlier studies, and (b) a methodology based on a perceptual auditory model that attempts to mimic the speech quality perception by a normal hearing listener. Experimental results with a database of 35 TE speakers showed that the auditory-model based approach significantly outperforms the traditional methods.

Index Terms— vocal system, speech analysis, linear prediction coding, auditory system

I. INTRODUCTION

Measurements of voice and speech quality are important during the assessment, treatment, and monitoring of talkers with abnormal voices. Speech quality measurements can be subjective or objective. *Subjective* analysis involves having a group of listeners rate the quality of the speech sample based on how natural it sounds, or how much effort is required for it to be understood. Subjective ratings are well known to be the gold standard for speech quality, however they suffer from being time consuming and expensive. *Objective* measures are often desired, as they use mathematical equations and physical models to predict the speech quality.

A speech abnormality can be the result of any number of factors such as disease or injury. The focus of this paper is on a particular type of pathology known as tracheoesophageal speech. Tracheoesophageal (TE) speech is a surgical-prosthetic method of voice restoration following surgical removal of the larynx. Voice restoration through TE puncture involves surgical creation of mid-line puncture in the common wall between the trachea and the esophagus [1]. This puncture is then stented with a small one-way valved prosthesis. When the airway is sealed, the TE puncture prosthesis permits pulmonary air to flow from the airway to the esophageal reservoir. Once air fills the esophageal reservoir, it vibrates muscular tissue of the upper esophagus and lower pharynx and this intrinsic, alaryngeal voice source is transmitted into the vocal tract where it is articulated into speech. TE speech is characterized by a generally lowered frequency, near normal intensity, and because of access to the large volume of pulmonary air, generally normal temporal features when compared to normal speakers [2]. However, the overall sound quality of TE speech is best described as highly aperiodic, rough, and noisy. Additionally, considerable variability across TE speakers does exist [3]. Therefore measurements of TE speech quality are often useful in TE speech rehabilitation process.

In this paper, we investigated different methods that predict the subjective speech quality ratings of TE speech. A novel contribution of this paper is the incorporation of an auditory model into the prediction process. A validated psychoacoustic model based on Moore and Glasberg's work [4] was used, and our results showed that this approach provides significantly improved correlation with subjective quality ratings of TE speech.

II. LINEAR PREDICTION ANALYSIS

Previous studies on TE speech analysis have looked at several different temporal and spectral parameters. One of the most powerful speech analysis techniques is the linear predictive (LP) modeling for its accurate estimates of speech parameters. Several studies [5], [6] have shown that LP-based metrics provide very good classification accuracy and correlate well with subjective ratings for normal speech. Parsa *et al.* [5] demonstrated that measures based on LP modeling of vowel samples were superior to other glottal measures in classifying pathological voices. Perhaps not

directly related, but Grancharov et al. [6] have shown that statistical quantities derived from linear prediction spectrum analysis are useful in non-intrusive estimation of the speech coder quality. Thus, our initial study concentrated on the analysis of continuous TE speech samples using the global statistical properties of the per-frame feature vector[6].

The per-frame power spectrum is initially calculated based on the LP coefficients a_k

$$P(\omega) = \frac{1}{\mid 1 + \sum_{k=1}^{p} a_k e^{-j\omega k} \mid^2} \tag{1}$$

from which three features are calculated, *viz.* Spectral Flatness $\Phi_1(n)$, Spectral Dynamics $\Phi_2(n)$, and Spectral Centroid $\Phi_3(n)$. The Spectral Flatness Ratio (SFR)

$$\Phi(n) = \frac{exp(\frac{1}{2\pi}\int_{-\pi}^{\pi} log(P_n(\omega))d\omega)}{\frac{1}{2\pi}\int_{-\pi}^{\pi} P_n(\omega)d\omega} \tag{2}$$

measures the distribution of frequencies in the spectrum. A 0 dB SFR results from a flat spectrum consisting mainly of noise. The output moves further away from 0 dB when the spectrum contains peaks and valleys [7]. The second feature used was the Spectral Centroid (SC)

$$\Phi_3(n) = \frac{\int_{-\pi}^{\pi} \omega log(P_n(\omega))d\omega}{\int_{-\pi}^{\pi} log(P_n(\omega))d\omega} \tag{3}$$

which defines the frequency region containing most of the signal energy. The final feature studied was the Spectral Dynamics(SD)

$$\Phi_2(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} (logP_n(\omega) - logP_{n-1}(\omega))^2 d\omega \tag{4}$$

which captures the frame-to-frame difference in the spectral density values.

III. AUDITORY MODEL

In order to better predict the subjective ratings, a validated psychoacoustic model was applied. The psychoacoustic model provides a way to link the physical acoustical data to the perceptual data by extracting perceptually relevant features [8]. Several models for auditory perception do exist but the Moore-Glasberg (MG) model [11], [10], [4], has been recently shown to be a more accurate revision of the earlier auditory models. The M-G model allows the computation of perceived loudness patterns which can be assimilated to form indices of voice quality. The computation of the loudness patterns is achieved through a series of steps and these are briefly explained below.

The level normalization step involves setting the Sound Pressure Level (SPL) of the input signal to a fixed level of 79dB [8], which represents a level to which most listeners have the best quality response. The next step in the process is converting the normalized signal into the time-frequency domain using the Short-Term Fourier Transform (STFT). The speech signal was divided into frames of $32ms$ length with 50% overlap. The power spectrum of each frame was calculated by taking the STFT and then squaring the real and imaginary components. Once the power spectrum is calculated the weighted power spectrum can be obtained as

$$P_w(i,k) = H(k) * P(i,k) \tag{5}$$

where i is the frame number and k represents the frequency scale. The frequency dependent weighting function which models the outer and middle ear is [12]

$$\begin{aligned} H(k) \quad = \quad & 10\frac{6.5exp(-0.6(\frac{f(k)}{1000} - 3.3)^2)}{20} \\ & -10^{-3}\frac{(\frac{f(k)}{1000})^{3.6}}{20} + 10\frac{-2.184(\frac{f(k)}{1000})^{-0.8}}{20} \end{aligned} \tag{6}$$

Here $f(k) = k\frac{8000}{256}$. The weighted power spectrum was used to calculate the excitation pattern

$$E(f_c) = \int_0^{\infty} \phi(f, f_c, P_w) P_w(f) df \tag{7}$$

$\phi(f, f_c, P_w)$ is the auditory filter. The excitation pattern was used to represent the output level of each auditory filter as a function of each respective center frequency [8]. The excitation pattern was transformed into the associated loudness pattern.The loudness pattern is more closely related to the subjective perception of the speech. Using the MG model the loudness patterns can be calculated for three different cases [4], [8]:

$$\begin{aligned} & Case1 : IF(10^9 \geq E(f_c) \geq E_{TH}(f_c)) \\ & N(f_c) = C[(G(f_c)E(f_c)E(f_c) + A(f_c))^{\alpha(f_c)} \\ & -A(f_c)^{\alpha(f_c)}] \end{aligned} \tag{8}$$

$$\begin{aligned} & Case2 : IF(E(f_c) > 10^9) \\ & N(f_c) = C[\frac{E(f_c)}{1.115}]^{0.2} \end{aligned} \tag{9}$$

$$\begin{aligned} & Case3 : IF(E(f_c) < E_{TH}(f_c)) \\ & N(f_c) = C(\frac{2E(f_c)}{E(f_c) + E_{TH}(f_c)})^{1.5} \\ & [(G(f_c)E(f_c) + A(f_c))^{\alpha(f_c)} - A(f_c)^{\alpha(f_c)}] \end{aligned} \tag{10}$$

The variables used in the loudness pattern calculations are shown in Table I. The loudness patterns are summed across frequency to form the overall loudness pattern of each speech frame. Eight separate distance metrics were computed from the loudness pattern differences, each with a unique

Table I. LPD variables

Variable	Meaning
E	excitation patterns produced by signal
E_{TH}	excitation patterns in quiet
G(f)	low level gain of cochlear amplifier
A(f)	constant to determine the level dependence of loudness compression
C	scaling constant = 0.047
$\alpha(f)$	compression exponent

frequency weighting scheme [8]. It has been shown, using a database of speech coder quality ratings, that the loudness pattern distortion (LPD) between the loudness patterns of the input (reference signal) and output (test signal) of a speech coder is a good predictor of the speech coder quality [8]. In this paper, we applied the LPD measurement procedure to the TE speech samples, by selecting the speech sample with the best subjective rating as the reference and all the other speech samples as test samples.

IV. RESULTS

The speech samples were gathered from thirty-five adult males between the ages of 45-65 years. All had undergone total laryngectomy and TE puncture at least one year prior to their participation. All recordings were gathered in a sound-treated environment using stereo recordings at 44.1kHz sampling rate with 16-bit quantization. The sentence *The rainbow is a division of white light into many beautiful colors* was recorded from all the speakers and used for acoustic and perceptual measurements.

The TE speech samples were played back to a group of 37 naive listeners who had no prior exposure to TE speech. The signals were played back in a random order and the listeners were instructed to rate the overall perceived quality on a scale of 1 to 10. The average of listener ratings was then used to determine the speech sample with the best perceptual rating and in the computation of correlation coefficients between objective and subjective ratings.

For the acoustic analysis portion, the speech signals were down-sampled to 8 kHz, and both the linear prediction spectral measures and the LPD values based on the Moore-Glasberg model were computed. Figures 1 and 2 show the loudness patterns for the reference signal (high quality subjective rating) and one of the test signals respectively. A substantial difference between the two patterns is quite evident in these figures.

Table II displays the correlation coefficients between the LPC spectral metrics and the subjective quality ratings. Only those features that demonstrated significant correlations were reported in this table. Even then, the magnitude of the correlation coefficients is quite low indicating a poor predictability performance by spectral metrics. Table III displays the correlation coefficients of the auditory model based distance metrics with the subjective data.Two distance

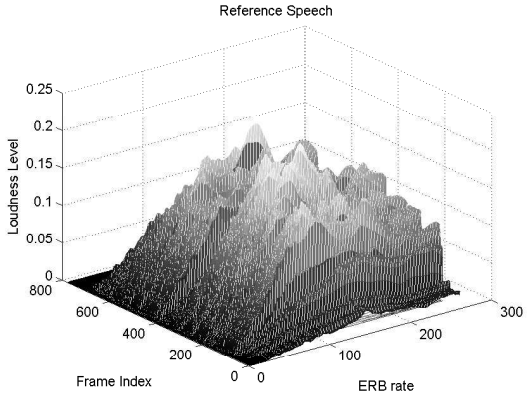

Fig. 1. Loudness patterns of reference speech signal

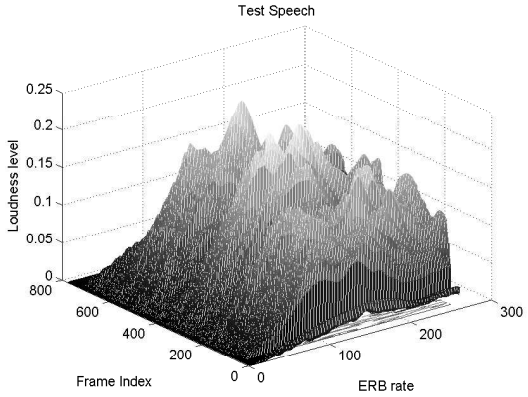

Fig. 2. Loudness patterns of test speech signal

metrics were calculated. One based on the overall difference between the loudness patterns of each frame and the other based on the specific loudness of each frame.

$$D_1 = \frac{1}{N}\sum_{i=1}^{N}\left[L_x(i) - L_y(i)\right] + L_{offset} \tag{11}$$

$$D_2 = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{u=1}^{M}\left[N_x(i,u) - N_y(i,u)\right]^2\right)}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{u=1}^{M}\left[N_x(i,u)\right]^2\right)}} \tag{12}$$

L_{offset} is an offset constant and N is the number of speech frames.The results in table III are for two versions of the speech samples. The "original" row reports the data obtained by processing the unmodified samples, while the "processed" row reports the data when the reference and test signals have been subject to an energy threshold based Voice Activity Detector (VAD). The best correlations were observed on the first two distance metrics of the processed speech signal. For the "original" case however, a correlation of 0.79 was

achieved by a linear regression analysis using $D1$ and $D2$ distance metrics.

Table II. Significant acoustic feature results

Feature	Statistic	Correlation
SF	Mean	0.23
	Variance	0.28
SD	Mean	0.23
	Skewness	0.28
SC	Mean	0.20
	Kurtosis	0.26

Table III. Significant auditory model correlations

Signal	Distance measure	Correlation value
Original Signal	D1	-0.69
	D2	-0.18
Processed Signal	D1	-0.73
	D2	-0.73

Figure 3 depicts the scatter plot between the objective (D1) and subjective ratings of the quality of TE speech. It can be seen from the plot that the higher the perceptual quality score, the lower the distance measure. This is to be expected as more distorted the test signal becomes, the lower its quality will be, making its distance further away from the high quality reference signal.

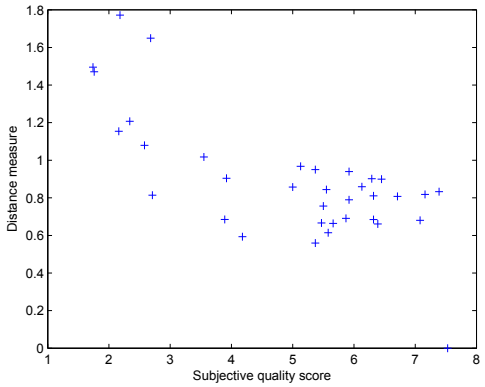

Fig. 3. Scatterplot of subjective rating vs. distance

V. CONCLUSIONS

Objective quality measurements are a valuable tool to speech pathologists in guiding patients to improve their speech. This paper has demonstrated the fact that typical spectral metrics extracted from running speech samples do not correlate well with perceptual subjective data of TE speech samples. Psychoacoustic models however, are better suited for modeling perceptually rated data. The proposed method uses the Moore-Glasberg loudness model to compute the LPD distances from the feature matrix. We showed that good correlation results could be obtained using this model for the TE database. Since we have not explicitly defined the distortion model, the algorithm can be extended towards future quality assessment of TE speakers.

VI. ACKNOWLEDGMENT

Financial support from the NSERC Canada and the Oticon Foundation, Denmark is gratefully acknowledged. We thank all the speakers and the listeners who participated in this study.

VII. REFERENCES

[1] Singer, M.I. & Blom, E.D. (1980). An endoscopic technique for restoration of voice after laryngectomy. *Annals of Otology, Rhinology, and Laryngology*, 89, 529-533.

[2] Robbins, J., Fisher, H.B., Blom, E.D., & Singer, M.I. (1984). A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production. *Journal of Speech and Hearing Disorders*, 49, 202-210.

[3] Eadie,T.L.,& Doyle, P.C. (2002). Direct magnitude estimation and interval scaling of naturalness and severity in tracheoesophageal (TE) speakers. *Journal of Speech-Language-Hearing Research*, 45, 1088-96.

[4] B.C.J.Moore and B.R.Glasberg, "A revised model of loudness perception applied to cochlear hearing loss", *Hearing research*,vol.188,pp.70-88, 2004.

[5] V.Parsa and D.G.Jamieson, "Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech," *J. Speech, Language, Hear. Res.*, vol.44,pp.327-339, 2001.

[6] V.Grancharov, D.Y.Zhao,J.Lindblom, and W.B.Kleijn, "Low-Complexity, Nonintrusive Speech Quality Assessment", *IEEE Trans.,Audio,Speech, and Language*,vol.14,pp.1948-1956,2006.

[7] V.Parsa and d.G.Jamieson, "Identifcation of pathological voices based on glottal noise measures," *J. Speech Hear. Res.*,vol.43,pp.469-485, 2000.

[8] G.Chen,"Statistical Model-Based Objective Measures of Speech Quality",PhD. Thesis, University of Western Ontario, London, ON, Canada, 2007.

[9] H.Fletcher,"Auditory Patterns",*Reviews of modern physics*,vol.12, pp.47-65, 1940.

[10] B.R.Glasberg, B.C.J.Moore and T.Baer, "A model for the prediction of thresholds, loudness and partial loudness", *Journal of the audio engineering society*,vol.45,no.4,pp.224-239, 1994.

[11] B.R.Glasberg and B.C.J.Moore,"Derivation of auditory filter shapes from notched-noise data",*Hearing Research*,vol.47,pp.103-138,1990.

[12] ITU, "Method for objective measurements of perceived quality",*ITU-R Recommendation BS.1387-1*, 2001.

[13] E.Zwicker and H.Fastl, *Psychoacoustics:Facts and Models*. Berlin, Germany:Springer-Verlag, 1990.