31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Reference-Free Automatic Quality Assessment of Tracheoesophageal Speech

Andy Huang, Tiago H. Falk, Wai-Yip Chan
Department of Electrical and Computer Engineering
Queen's University, Kingston, Ontario, Canada

Vijay Parsa, Philip Doyle
Department of Electrical and Computer Engineering,
School of Communication Sciences and Speech Disorders
University of Western Ontario, London, Ontario, Canada

Abstract— **Evaluation of the quality of tracheoesophageal (TE) speech using machines instead of human experts can enhance the voice rehabilitation process for patients who have undergone total laryngectomy and voice restoration. Towards the goal of devising a reference-free TE speech quality estimation algorithm, we investigate the efficacy of speech signal features that are used in standard telephone-speech quality assessment algorithms, in conjunction with a recently introduced speech modulation spectrum measure. Tests performed on two TE speech databases demonstrate that the modulation spectral measure and a subset of features in the standard ITU-T P.563 algorithm estimate TE speech quality with better correlation (up to 0.9) than previously proposed features.**

I. Introduction

A variety of disease and medical complications can create abnormalities in voice quality. One particular voice abnormality is produced by total laryngectomy (the removal of the larynx), most frequently as the result of cancer. In this situation, several different methods to restore speech production are possible. No matter which "alaryngeal" voice mode is used, they all share one common element: reduction in voice quality and capacity, and consequently a need to evaluate the quality of a patient's voice during rehabilitation. The current benchmark for evaluation of voice quality is by *subjective* quality scores where a panel of listeners evaluate speech samples based on a set of pre-identified criteria (e.g. hoarseness, breathiness, roughness). However, due to the common necessity of expert evaluators for such an evaluation, this method is expensive in both labor and time. These costs severely limit the capacity to undertake these evaluations. A machine-based *objective* scoring system to evaluate the speech signal on subjective scales would greatly assist the rehabilitation process.

The voice restoration method that produces the greatest similarity to normal voice condition in frequency, intensity, and temporal domains is tracheoesophageal (TE) speech. A surgical puncture is made in the common anatomic wall between the trachea and esophagus and a one-way valve is inserted. The valve allows air to flow from the trachea into the esophagus, and induces vibrations in the upper esophagus/lower pharynx (the pharyngoesophageal segment) to produce voice/speech [1]. While TE speakers in some respects are able to approximate normal speech patterns, TE speech is characterized as highly aperiodic, rough, and noisy [2]. TE speakers are also restricted to a pitch range of 50-90 Hz for both genders in comparison to a normal pitch range of 50-400 Hz.

In this paper we pursue a machine-learning approach to objective quality assessment of TE speech. With this approach, the features extracted from the speech signal have a direct impact on performance. Speech quality estimation algorithms can be grouped into two categories: *reference-based* and *reference-free*. "Reference" refers to an input, distinct from the test signal, to the quality estimation algorithm to serve as a baseline (usually of "good" quality) for comparison. Reference-based algorithms rely on this reference in order to provide an estimate of speech signal quality. Reference-free algorithms, in turn, are not comparison based and the estimated speech quality score is dependent solely on features extracted from the test signal. Due to differences in evaluation strategy, reference-based and reference-free algorithms have distinct feature requirements. Prior work has been conducted utilizing both reference-based and reference-free methods [3][4][5][6].

Investigation into reference-based evaluation methods have utilized features from linear prediction analysis and auditory models [3], where a "good" quality speech signal produced by a separate speaker was used as a reference. The reference speech sample must be carefully selected in order to balance out various nuances and speech characteristics to ensure accurate evaluation. In addition, dynamic time warping is required to properly align the reference and test speech signals. In [4] an automatic speech recognizer was applied alongside additional prosodic features to predict subjective scores. The reference in this methodology is the transcript of the spoken speech, needed to calculate the speech recognition rate (SRR). While this method produced high correlation on the tested database, speech recognizer performance is problematic on atypical speech. Moreover, while SRR is strongly correlated with the "intelligibility" of speech, the "naturalness" of speech may not be reflected in the SRR. Although voice restoration is primary for those who lose their normal voice production, efforts that seek to facilitate the production of natural sounding speech is an important goal for rehabilitation of all alaryngeal speakers, but is of particular importance for those who use TE speech.

Reference-free TE speech evaluation using features extracted by time-frequency analysis has been investigated in [5]. However, the time-frequency features showed poor cor-

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

relation with TE speech subjective scores. Prosodic features were investigated in [6] and demonstrated promising results. These features, however, were only examined on sustained phonemes from TE speakers and may not correlate well with the overall speech quality of the speaker.

In this paper we investigate the use of existing objective quality measurement algorithms - originally developed for narrowband telephone speech - for TE speech quality estimation. In particular, we explore using the International Telecommunication Union ITU-T recommendation P.563 algorithm [7] and American National Standard Institute ANSI ANIQUE+ algorithm [8]. P.563 inputs a narrowband test speech signal and outputs an estimate of the subjective mean opinion score (MOS) of the signal. Features extracted from the test signal have been shown to accurately distinguish between good and poor quality of narrowband telephone speech [9]. These features and the estimated MOS serve as a potential set of candidate features for use in TE speech quality evaluation. To augment this candidate set we also include the MOS estimated by the ANIQUE+ algorithm [8]. ANIQUE+ estimates the MOS based on a perceptual model that mimics the human auditory system [8]. In contrast to P.563, ANIQUE+ primarily relies on analysis of the test speech signal's modulation spectrum. The modulation spectrum captures the frequency content of a signal's temporal envelope. From the modulation spectrum we can extract information on the slower temporal behavior exhibited by a signal. An additional modulation spectral feature, the reverberation to signal modulation ratio (RSMR) [10], is also included. This modulation spectral feature employs knowledge of natural clean speech's modulation spectral characteristics to create a reference-free comparison between TE and normal speech. Specifically, temporal variation at typical syllabic and phonemic rates of speech signals establishes a nominal modulation spectral characteristic for discrimination between normal and TE speech.

We propose the use of sequential forward feature selection with support vector regression (SVR) in order to estimate the TE speech subjective quality scores. The remainder of this paper is organized as follows. Candidate features are described in Section II, the feature selection algorithm is described in Section III, and experimental results are reported in Section IV. Lastly, conclusions are reported in Section V.

II. Feature Extraction

In this section we outline previously proposed features for TE speech quality assessment. The proposed RSMR feature is then described.

A. Adaptive Time-Frequency Analysis

From time frequency analysis a set of four features are extracted: the energy capture rate (ECR), frequency ratio, Ocmean, and Ocmax. A full description of the features can be found in [5].

B. Standardized Objective Quality Analysis

P.563 utilizes a total of 43 different features to produce an estimate of the MOS, including features based on vocal tract, noise, and continuity analysis. A full list of P.563 features can be found in [7]. Together with the estimated MOS, P.563 contributes 44 candidate features. The estimated MOS from ANIQUE+, a competitor to the ITU-T P.563 algorithm, adds one additional feature to the candidate pool resulting in 45 features from standardized objective quality analysis.

C. RSMR

Introduced in Falk et al. [10], the reverberation-to-signal modulation energy ratio (RSMR) is an adaptive feature that exploits the modulation spectral characteristics of clean speech to compare the modulation energy between the signal and "room reverberation". Typical clean speech contains modulation frequencies approximately in the 2-20 Hz modulation frequency band with spectral peak at approximately 4 Hz [10]. From this clean speech characteristic we presume that spectral content with modulation frequencies greater than 20 Hz is due to "noise" embedded in the speech signal. In the context of the original feature definition, this additional noise is caused by room reverberation. For TE speech, these additional modulation frequencies are due to artifacts (e.g. gurgling, raspiness) present in the signal. We essentially consider the RSMR feature as a artifact-to-signal modulation energy ratio. We summarize below the process for extraction of the RSMR feature.

To calculate the feature, the speech signal is initially processed by a bank of 23 critical band gammatone filters. Hilbert transform is performed at the output of each filter j to calculate the temporal envelope $e_j(n)$. These temporal envelopes are multiplied by 256 ms Hamming windows with frame shift of 32 ms to calculate the temporal envelopes of each frame, $e_j(m)$. The modulation spectrum is calculated as $E_j(m;f) = |\mathcal{F}(e_j(m))|$ where $\mathcal{F}$ denotes the DFT operation and f indexes the modulation frequency bins. These modulation frequencies are grouped into $\mathcal{K}$ bands, where the where the energy for frame m is denoted by $\mathcal{E}_{j,k}(m)$, $k = 1, \ldots, \mathcal{K}$. The mean of the modulation energy for all N_{act} active frames is calculated to be:

$$\bar{\mathcal{E}}_{j,k} = \frac{1}{N_{act}} \sum_{i=1}^{N_{act}} \mathcal{E}_{j,k}^{act}(i) \tag{1}$$

The average modulation energy per modulation frequency band is:

$$\bar{\mathcal{E}}_k = \frac{1}{23} \sum_{j=1}^{23} \mathcal{E}_{j,k} \tag{2}$$

The RSMR is then:

$$\text{RSMR} = \frac{\sum_{k=5}^{K^*} \bar{\mathcal{E}}_k}{\sum_{k=1}^{4} \bar{\mathcal{E}}_k} \tag{3}$$

where K^* is adapted to the speech signal [10].

III. Sequential Feature Selection

Given a set A of currently selected good features, a candidate feature B_i from a set B of candidate features is added to A to form a candidate feature set C_i. Cross validation (CV) is performed using SVR to evaluate the performance of the candidate feature set C_i. The candidate feature set that produces the best correlation between the predicted scores $\hat{y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N]$ and actual scores $y = [y_1, y_2, \ldots, y_N]$ is added to A and deleted from B

The algorithm can be summarized as follows:

1) Given initial feature set A and candidate feature set $B = [B_1, B_2, \ldots, B_M]$,
2) For $i = 1 : |B|$, do
   $C = A \cup \{B_i\}$,
   Perform CV using SVR with feature set C and get predicted score $\hat{y}_i = [\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,N}]$,
3) Find $j = \underset{i}{\operatorname{argmax}} \left| \frac{\sum_k (y_k - \bar{y})\left(\hat{y}_{i,k} - \bar{\hat{y}}_i\right)}{\sqrt{\sum_k (y_k - \bar{y})^2 \sum_k \left(\hat{y}_{i,k} - \bar{\hat{y}}_i\right)^2}} \right|$ where $\bar{y}$ and $\bar{\hat{y}}_i$ are the mean of y_k and $\hat{y}_{i,k}$, respectively.
4) $A = A \cup \{B_j\}$, $B = B - \{B_j\}$. Re-index the features in B.
5) If termination criterion not met (i.e. $|A| < L$, (where $L \in \mathbb{Z}$ is a limit on the size of feature set A) is not true), return to step 2.

IV. Experiment

In this section, we describe the process of TE speech data collection and the methodology for obtaining our results.

A. Data

Speech samples were collected from 28 adult males of 45-65 years in age that had undergone total laryngectomy and TE puncture at least one year prior to the recording. All recordings were gathered in a sound-treated environment recorded at 44.1 kHz sampling rate with 16-bit quantization. Each speaker read the sentence *The rainbow is a division of white light into many beautiful colors*. Each speech file was subjectively scored based on the severity of the speech by 24 naive listeners on a scale of 1-100, with a lower score indicating less severe speech. The final subjective score for each speech file was the mean subjective score from the 24 listeners. We aim to find TE speech signal features that are effective for estimating the mean subjective score.

The speech signal was downsampled to 8 kHz before processing by the P.563 and ANIQUE+ algorithms to produce 45 different features. The RSMR feature was extracted to provide a total of 46 different features. All features were then normalized to zero mean and unit variance; features with no variation between speech files were removed to produce a set of 40 candidate features for TE speech quality evaluation. The removed features (e.g. robotization, multiplicative noise) had preconditions that TE speech could not satisfy, causing the P.563 algorithm to assign them default values for all TE speech signals. The sequential feature selection algorithm was executed on the 40 remaining features to select a subset of features for SVR. Leave-one-out cross validation (LOOCV) was performed to evaluate the SVR based on the magnitude of the correlation between predicted and actual subjective scores. Both a linear kernel and non-linear radial basis function (RBF) kernel were utilized for SVR.

TABLE I
Time-Frequency Analysis Features

Feature	Correlation
ECR	0.52
Frequency ratio	0.28
Ocmean	0.38
Ocmax	0.24

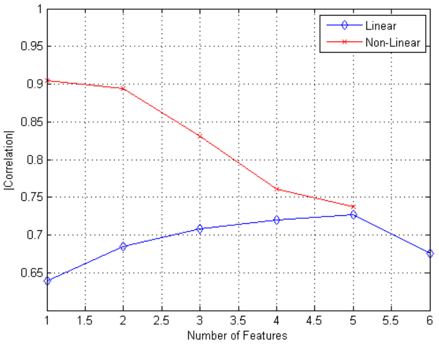

Fig. 1. Performance of SVR utilizing automatically selected features

B. Results

We first explore the correlations obtained between subjective TE speech scores and the features from time-frequency analysis. Results are reported in Table 1. As can be seen, the previously-proposed features show poor correlation with subjective scores. The best performing feature, ECR, exhibits a correlation of 0.52. Next, we examine the correlation obtained between subjective TE scores and our proposed set of 40 candidate features. Figure 1 shows the performance of the SVR with sequentially selected features. As observed, the best performance for linear SVR is obtained by utilizing five features. The inclusion of additional features beyond five features degrades performance due to overfitting. Non-linear SVR achieves best performance with the top one or two features. The inclusion of more than two features degrades estimation performance due to overfitting.

The ranking of the features for both linear and non-linear SVR are shown in Table II. Three of the five best features for linear SVR are extracted from the P.563 background noise analysis module: LocalBGNoise, LocalBGNoiseMean, and EstBGNoise. The first two local background noise features measure noise between phonemes. The LocalBGNoise feature is the percentage of samples classified as local background noise. The LocalBGNoiseMean is the mean energy of frames containing local background noise. The EstBGNoise feature is an estimate of the overall background

TABLE II
Ranking of Features

Rank	Feature	
	Linear kernel	RBF Kernel
1	RSMR	PitchCrossPower
2	LocalBGNoise	UnnaturalBeeps
3	LocalBGNoiseMean	FrameRepeatsTotEnergy
4	EstBGNoise	FrameRepeats
5	VTPMaxTubeSection	UnnaturalBeepsAffectedSamples

noise floor level in dBov (dB relative to overload). Low quality TE speech signals often contain artifacts. With the artifacts classified as noise, the presence of artifacts would be reflected in the extracted noise features. The final P.563 feature within the best five is VTPMaxTubeSection, which is a measure of the maximum glottis opening (for laryngeal speakers in the P.563 context) over the entire test signal and is used to evaluate the unnaturalness of the speech signal.

With the non-linear SVR, the best performance requires only a single pitch related feature - the PitchCrossPower. The PitchCrossPower is based on the cross power spectrum between consecutive pitch synchronous frames. The cross power spectrum is the Fourier transform of the cross correlation between the consecutive pitch synchronous frames. The impact of this feature on non-linear SVR performance reaffirms the importance of pitch related features in TE speech quality evaluation as reported in [6].

As a final comparison, we test the performance of our algorithm utilizing the TE database described in [3] [5]. The database described in [3] [5] (henceforth referred to as TE database two), has a different subjective scoring scale. The database described earlier in this paper rated the speakers on the severity of the speech on a scale from 1-100 while in TE database two the speakers are rated on “listener comfort” on a scale of 1-10. A full description of the compared features can be found in [3] and [5]. The performance of the linear and non-linear SVR along with the two best performing features, auditory model and ECR, from [3] and [5] respectively, are shown in Table III. Note that the auditory model is a reference-based evaluation since the D2 distance measure requires a reference signal. We can see that both the linear and the non-linear SVR outperform previous work conducted on this database. The linear SVR used five features for its best performance, and the non-linear SVR required three features for its best performance. Feature rankings are shown in Table IV. We also show the performance of the non-linear SVR on the database when all five features listed in Table IV are used. We can see that the overfitting due to the two additional features degrades the non-linear SVR performance to the same level as the auditory model.

On both databases the linear SVR requires five features for its best performance. Much like on the first database, background noise features play an important role as both LocalBGNoiseLog and the EstBGNoise feature are present in the best performing feature set. The non-linear SVR required three features to obtain its best performance on this database. CepCurt (kurtosis of the cepstral coefficients obtained from vocal tract analysis) is a statistic that serves as a measure of speech distortion. SpectralClarity measures the absence of spectral energy at frequencies in between the harmonic frequencies of voiced speech. SharpDeclines measures unnatural drops in temporal energy of the speech signal.

V. Conclusion

In this paper, we have assessed the efficacy of a set of candidate features for SVR based reference-free TE speech quality estimation. Including a modulation spectral feature (RSMR), the set of candidate features is extracted using P.563 and ANIQUE+, two standard algorithms for reference-free estimation of telephony speech quality. Test results for two TE speech databases demonstrate that RSMR and a subset of P.563 features estimate TE speech quality with better correlation than previously proposed features. Additional effort is under way to acquire more TE speech data in order to improve the robustness of the selected features.

TABLE III
Comparison of performance on the TE database described in [3] [5].

Feature	$\lvert$Correlation$\rvert$
ECR	0.63
Auditory Model (D2 distance measure)	0.73
5 feature linear SVR	0.86
3 feature non-linear SVR	0.77
5 feature non-linear SVR	0.73

TABLE IV
Ranking of Features on the TE database described in [3] [5]

Rank	Feature	
	Linear kernel	RBF Kernel
1	RSMR	RSMR
2	LocalBGNoiseLog	CepCurt
3	EstBGNoise	EstBGNoise
4	SharpDeclines	SNR
5	SpectralClarity	SharpDeclines

References

[1] M. Singer and E. Blom, “An endoscopic technique for restoration of voice after laryngectomy,” *Annals of Otology, Rhinology, and Laryngology*, vol. 45, pp. 202–210, 1980.

[2] J. Robbins, H. B. Fisher, E. C. Blom, and M. I. Singer, “A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production,” *J Speech Hear Disord*, vol. 49, no. 2, pp. 202–210, 1984.

[3] R. McDonald, V. Parsa, P. Doyle, and G. Chen, “On the prediction of speech quality ratings of tracheoesophageal speech using an auditory model,” *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4517–4520, 2008.

[4] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, “Peaks - a system for the automatic evaluation of voice and speech disorders,” *Speech Commun.*, vol. 51, no. 5, pp. 425–437, 2009.

[5] R. McDonald, V. Parsa, and P. Doyle, “Prediction of the quality ratings of tracheoespohageal speech using adaptive time-frequency representations,” *Proc. of Canadian Conference on Electrical and Computer Engineering*, pp. 1715–1718, May 2008.

[6] C. J. van As-Brooks, F. J. K. van Beinum, L. C. Pols, and F. J. Hilgers, “Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech,” *Journal of Voice*, vol. 20, no. 3, pp. 355 – 368, 2006.

[7] P.563, *Single-ended method for objective speech quality assessment in narrow-band telephony applications*, Intl. Telecom. Union Std., 2004.

[8] D.-S. Kim and A. Tarraf, “Anique+: A new american national standard for non-intrusive estimation of narrowband speech quality: Research articles,” *Bell Lab. Tech. J.*, vol. 12, no. 1, pp. 221–236, 2007.

[9] L. Malfait, J. Berger, and M. Kastner, “P.563 - the itu-t standard for single-ended speech quality assessment,” *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.

[10] T. H. Falk and W.-Y. Chan, “A non-intrusive quality measure of dereverberated speech,” *Proc. of Intl. Workshop for Acoustic Echo and Noise Control*, May 2008.