J. King Saud Univ., Vol. 14, *Eng. Sci.* (1), pp. 29-46, Riyadh (1423/2002)

Rating of Speech Quality in Mobile Phone Networks

Khalid A. Al-Mashouq
Electrical Engineering Department, King Saud University
PO Box 800, Riyadh 11421, Saudi Arabia
Email: mashouq@ksu.edu.sa

(Received 05 June, 2001; accepted for publication 29 December, 2001)

Abstract. Mobile phone users rely only on their intuition to evaluate the quality of service provided by their network operator. In this paper, we propose an output-based objective method to rate the speech quality in mobile phones. This rating, when averaged over time, will be useful in comparing between different mobile phone networks. Moreover, the rating could be incorporated into the calls tariff of mobile phones.

Here, A time-delay multilayer neural network is used to rate the speech quality after a proper training stage. The training input set consists of speech features (such as linear predictive coefficients LPC and per-frame energy). The training target is a customized per-frame speech quality measure. We performed extensive simulations using the proposed method. We use speech samples from simulated channels, real GSM channel and from the more elaborate CTIMIT database. These simulations show that the proposed method can predict the speech quality within reasonable accuracy.

Keywords: *Speech quality, mobile network, CTIMIT database, time-delay neural network.*

Introduction

Today's communications market is witnessing unprecedented high penetration rate of mobile phones. This has energized the competition between mobile network operators. Seeking customer satisfaction is a top priority to reduce customer "churn". Now more than ever, customers are looking for better quality of services.

On the top of his list, the customer is demanding good speech quality. Many customers believe that they are being overcharged for calls with speech quality that is less than acceptable. It is desirable that a customer has an objective means to measure the speech quality. A "smart' handset that can give rating for the speech quality will be essential in comparing between the different operators. Ultimately, it is conceivable that a call is charged based not only on airtime but also on speech quality assessed by the

handset. Speech quality tests are also essential for the network operator for fault detection and network optimization.

Traditionally, speech quality in mobile phones is measured through the perception of an expert listener (subjective test). This method, however, is expensive and of limited use. The alternative method is an automated one. A reference speech signal is transmitted through the network to a mobile unit connected to a computer for objective speech quality analysis. There have been many objective test algorithms [1,2]. Recently, S. Voran [3] proposed a new simple objective test which has high correlation (in the range of 0.84-0.98) with the mean opinion score (MOS) subjective test.

We wish that the user got the privilege of obtaining a quantitative quality measure of the telephone calls; traditionally this is restricted to the mobile network engineers. Here we propose a new method for speech quality rating without the need for a reference signal. This is called output-based speech quality assessment. Due to its difficulty, this problem has been tackled by only a limited number of researchers. Jin and Kubichek [4] presented a vector quantization method to assess the speech quality, and obtained correlation with MOS in the range of 0.68-0.9. They conjectured that 0.9 will be the limit for any output-based algorithm.

We agree partially with these expectations, especially if one does not use complicated high order Markovian models for the speech. This, however, does not imply that the 0.9 correlation is a sharp end. From a mobile user point view, a continuous quality scale may not be of utmost necessity; one with discrete number of levels (e.g. excellent, good, fair, poor and bad) may be sufficient.

Our approach here is to use a time-delay multilayer neural network [5] and train it with features extracted from noisy speech samples. The target of the network is an objective quality score, formulated in such a way to fit into the neural network framework. The obtained results are promising, which suggest more elaborate experimentation to have better generalizations.

The organization of this paper is as follows. The next section is a review of the main approaches to speech quality assessment. Then we describes our approach to output-based speech rating. The experimental work section contains simulation results obtained for different speech samples in different environments, including the CTIMIT database, which is obtained from the standard TIMIT database as transmitted through a cellular network. Main conclusions and future work are outlined in the last section.

Speech Quality Assessment

There are two approaches to measure the speech quality [1], subjective and objective approach. In the subjective approach, a group of people listens to speech

samples and each individual is asked to give quality scores. The average score represents the mean opinion of the listeners. A widely acceptable method is the mean opinion score (MOS). ITU-T [2] recommends five levels of quality, "excellent", "good", "fair", "poor" and "bad" and explains a procedure to determine MOS by listening tests under laboratory conditions. This method is usually applied in the evaluation of different codecs. A more sophisticated measure is the diagnostic acceptability measure (DAM), which is used for evaluating medium to high quality speech. All subjective tests are, however, not practical for the continuous evaluation of speech in mobile units.

The alternative is to use objective approaches [2]. These approaches range from simply measuring the bit error rate to the more complex schemes of modeling the human hearing system. Currently the objective approach is widely used by many network operators. Generally, in the objective approach, a known reference speech sample is transmitted through the network. The received speech sample along with the "clean" copy is preprocessed to ease the comparison. The preprocessing, or feature extraction, includes filtering, time-shift adjustment, normalization and framing. In the literature, different features have been proposed. Linear predictive coding (LPC) and its variants, such as the log-are ratio (LAR), and short-term DFT are examples of these features. Comparison is made between the features of the received signal and the clean copy. The comparison result is then mapped into MOS score.

In [3] S. Voran proposed a log-spectral-error-based method along with frame-energy plane partitioning for the objective test. His method is relatively simple and reaches 0.74 up to 0.98 correlation with the MOS. For online assessment speech quality by the mobile handset, any reference-based method is difficult to use since it requires the repeated transmission of a known speech sample.

Output-based Approach

Basic approach

Our approach is based on the idea that a human can judge the quality of impaired speech sample, within certain precision, without listening to the original speech sample. Therefore, we propose a time-delay neural network technique that can find the speech quality score by learning. The learning phase consists of presenting impaired speech samples (after pre-processing) along with their quality score. To make the input to the neural network of reasonable size, we divide the speech signal into short adjacent (possibly overlapping) frames. We extract certain features from each frame and arrange them in a vector, which we call a training vector. We should associate a speech quality score (or target) with each training vector. For obvious reasons, we cannot find a subjective quality score for an individual frame. In our work, therefore, we rely on the objective quality score, which can be calculated per frame. The overall speech quality score is the sum (or average) of the frames' score.

We train a time-delay neural network with the feature vectors. Thus if X(n) is the feature vector of the n^{th} frame, the n^{th} input to the network, $\chi(n)$, is

$$\chi(n)=\{X(n-m), X(n-m+1), ..., X(n), ..., X(n+m)\}$$

Information about the adjacent frames is needed to exploit the dependency between them. In addition to the present vector X(n), $\chi(n)$ contains m previous vectors and m future vectors. In simulated channels, errors are injected randomly in some frames. We use LPC coefficients or short-term DFT feature extraction methods. The training target will be simply the number of errors in the present frame.

Training for the digital cellular systems

The digital cellular systems are susceptible to different types of impairments, such as:

- Noise injection
- Echo
- Speech muting
- Robotic voice
- Ping Pong

These types are the major causes of degrading the speech quality in mobile environments. Some of these impairments, such as echo and noise injection, are common in almost all speech communication channels. Other impairments are due to the digital nature of wireless system. The speech muting could result from a damaged frame or due to an outage in the transcoder channels. Robotic voice is caused by erased frames. The decoder substitutes the erased frame with previous frames, causing the robotic-like voice. In severe channels, more frames are damaged causing the ping pong or bottle smashing sound.

For a real cellular channel, we use the log-area ratio (LAR) parameters as derived from the LPC coefficients. It is known that LAR-based objective measure has the highest correlation with the MOS among all LPC- parameters variations [1]. Let T(n) be the target associated with the input $\chi(n)$. This target should reflect some kind of per-frame quality measure of the n^{th} frame. Initially, we define T(n) as the Euclidean distance between the feature vector (LAR parameters) of the input noisy frame, X(n), and the clean frame, $X_c(n)$, i.e.

$$T(n)=\|X(n)-X_c(n)\| \quad (1)$$

Frame-Energy Plane Partitioning. Inspired by the work of Voran [3], the target, T(n) of the neural network is modified from (1) according to the loudness of the input/output frames. We first find the energy of the reference frame and received frame, $E_c(n)$ and E(n), respectively. The energy of signal *x* is defined as

$$E_x(\text{dB}) = 10.\log_{10}\left(\sum_{i=2}^{N} X_i\right) \quad (2)$$

where x_i is the frequency-domain sample of the speech frame for the N point FFT. The target is calculated according to the following "frame-energy plane portioning" (as shown in Fig. 1).

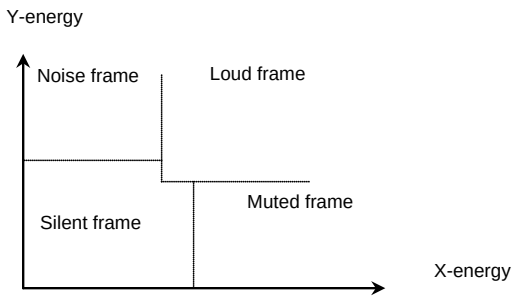

Fig. 1. Frame-energy plane partitioning.

- T(n) is as in (1) if E(n) and $E_c(n)$> 17.4 dB (loud frames).
- T(n)=35 if E(n)> 19 dB and $E_c(n)$ <17.4 dB (noise frame), or if $E_c(n)$> 19 dB and E(n) <17.4 dB (muted frame).
- T(n) =0 otherwise (silent frames)

This partitioning serves as a rough approximation of the human speech perception. More sophisticated partitioning is also possible.

We use different speech samples to extract the training pairs (χ(n), T(n)) from all the frames. A time-delay multilayer neural network is then trained with these pairs. After the training phase is completed, we test the performance of the trained network using new speech samples. Sometimes it is useful to post-process the network's output before obtaining the objective quality test as discussed previously.

Experimental Work

In our experiments, we used three types of channels. The first one is a simulated simple noise injection channel. The second channel is a typical GSM channel. The third

channel is a real cellular phone network as described by [6]. The speech data collected from this third channel is called CTIMIT database.

Simulated channel

We recorded several 5-second speech signals as uttered by different male speakers. Then we sampled each speech signal at a rate of 8000 sample/second. A speech signal is divided into 100 frames of length 50 ms each. The noise is injected in three arbitrarily chosen "damaged" regions. These regions correspond to the following frames: frame 3-10, frame 35-44, and frame 77-86. A frame contains 400 speech samples. For a damaged frame, we select at random N out of the 400 samples to be injected with (added to) additive Gaussian noise. The noise has zero mean and unity variance. N is called the noise level and it will take the values 10, 40, 100 and 200. The target, T(n), of the n^{th} frame represents the number of injected errors relative to the number of samples in the frame.

Training with LPC coefficients. In the first part of this experiment, we extract the LPC coefficients of the 11^{th} order from each frame. We arrange these coefficients in a vector X(n) for training the neural network as discussed in the previous section. The clean and the noisy speech samples, along with their targets, are used to train the network. Different 2-layer time-delay neural networks are tried. The number of taps used is 3 and 5. Fig. 2 shows the output of the 5-taps network with 7 hidden nodes due to noisy speech input. We notice that the general trend of the output moves in parallel with the desired output. However, for the noise-free regions we observe deviation from this trend. One can explain that by the inherent overlap between the noise and speech. Thus the network will sometimes confuse a clean speech with noise.

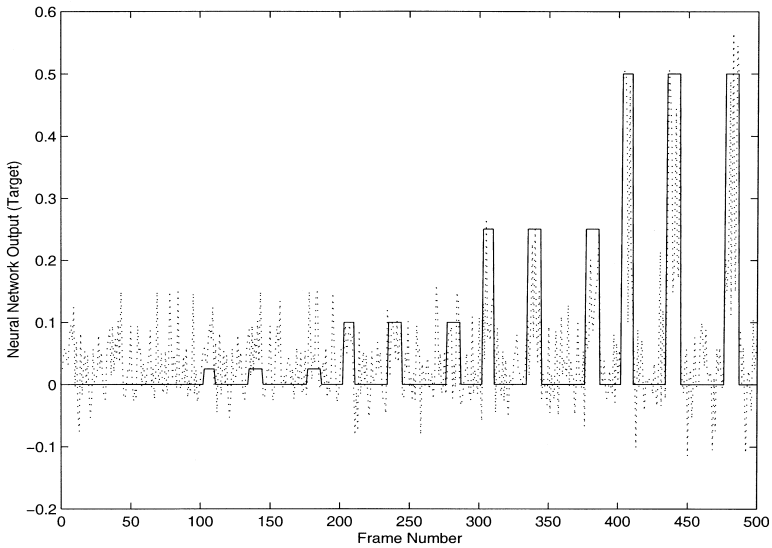

Fig. 2. Per-frame quality score (solid line) and its prediction by the neural network (dashed line). LPC coefficients are used as the network input.

Training with short-term DFT. In the second part of this experiment, we use the magnitude of the short-term DFT as a feature vector. A 50-ms frame size is considered with 50% overlapping. Hanning window is used to taper the frame. The resolution in the frequency domain is 300 points.

Again we performed several training passes with the magnitude of the short-term DFT as the input vector and T(n) as the target or desired output. As we did with the LPC coefficient case, we employed different 2-layer time-delay neural networks with different number of taps. Figure 3 shows the output of a 2-layer neural network with 5 taps in the input and 9 hidden nodes. Here we notice more agreement with the target value. The problem of the noise-free region is mitigated to some extent.

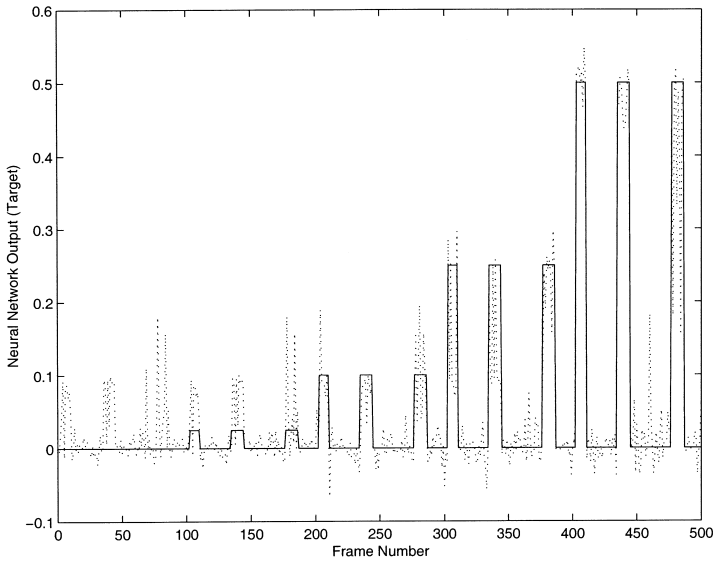

Fig. 3. Per-frame quality score (solid line) and its prediction by the neural network (dashed line). The Short-term DFT is the network input.

For comparison purpose between the LPC and short-term DFT features, we show Table 1. In this table, we compare the training MSE error with the testing error. The MSE error is defined as

$$\text{MSE} = \sum_{n=1}^{L} (T(n) - Y(n))^2 \qquad (3)$$

where Y(n) is the neural network output due to the n^{th} input $\chi(n)$, and L=100, is the number of frames in a speech signal.

Table 1 shows that both the LPC and short-term DFT features result in comparable performance with more preference to the short-term DFT. Moreover, increasing the number of taps in the time delay network can improve its performance.

Table 1. MSE of different networks

		LPC Coefficients		Short-Term DFT	
	No. hidden nodes	Train MSE	Actual MSE	Train MSE	Actual MSE
3-taps	7	.0057	.0098	4.0×10^{-5}	.0053
	9	.0074	.0074	1.8×10^{-4}	.0393
	12	.0065	.0065	1.9×10^{-5}	.0135
	15	.0051	.0052	2.1×10^{-5}	.0092
5-taps	7	.0078	.0079	4.7×10^{-4}	.0128
	9	.0076	.0076	7.3×10^{-5}	.0051
	12	.0057	.0057	1.9×10^{-4}	.0046
	15	.0068	.0068	3.4×10^{-4}	.0064

GSM-network channel

In this part, we use collections of 5-second speech samples prepared by Ascom [7]. These speech signals correspond to one sentence, which is phonetically balanced and uttered by a male and female. The speech samples have different MOS quality levels, ranging from excellent to bad. Their distribution based on the quality is shown in Table 2.

Table 2. Quality distribution of the training and testing speech samples

Quality rating	Training set	Testing set
Excellent	1	1
Good	4	5
Fair	4	5
Poor	4	5
Bad	3	4

The preprocessing stage of the speech signal includes normalizing the long-term mean to zero and the variance to one. Simple correlation is used to align the received noisy speech with the original one.

The sampling rate is 8000 sample/second and the frame size is 32 ms. Hanning window is used for each frame with 50% overlapping between frames. The LAR parameters are obtained for each frame and then the target value T(n) is computed as given in (1). We append the frame energy to the feature vector X(n). (Refer to Fig. 4.) Now we have a list of input/output pairs to train our 2-layer time-delay network. Different network structures are tried. For training, we used two training algorithms. The results obtained with each algorithm are given below.

Algorithm 1: Resilient Backpropagation [8]. This is a variation of the standard steepest-descent backpropagation that avoids the "dead ends" of the sigmoid function.
Algorithm 2: Bayesian regularization [9]. This is a Modification of the Levenberg-Marquardt [10] training algorithm to produce networks that generalize well. It reduces the difficulty of determining the optimum network architecture.

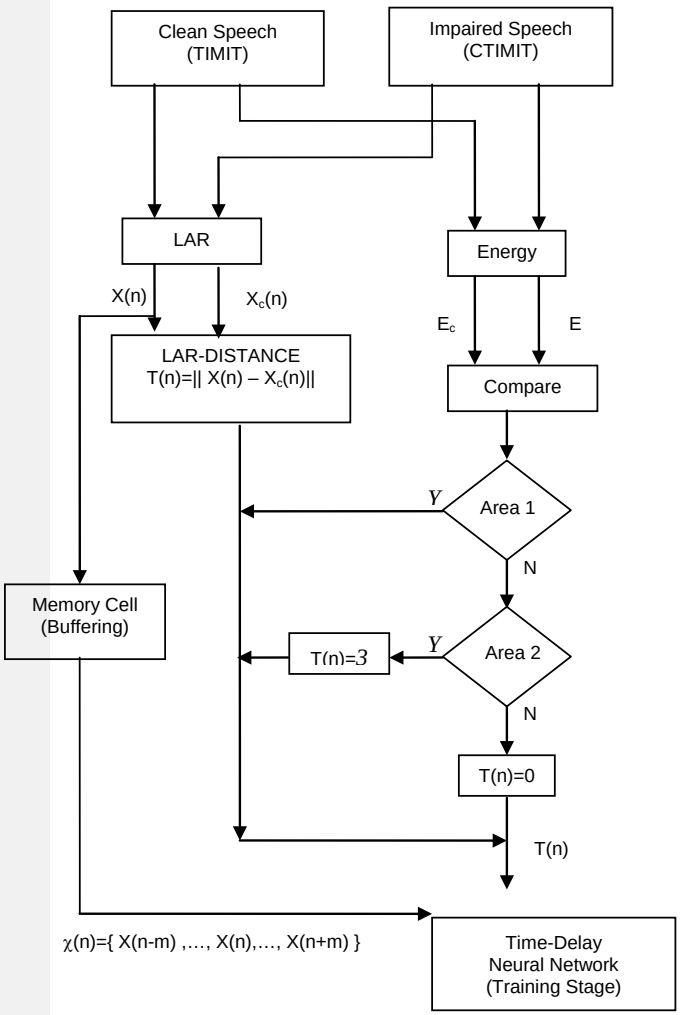

Fig. 4. Flow graph for preparing the training set to the time-delay neural network. Area 1 corresponds to loud frames and Area 2 corresponds to muted or noise frames as defined in Fig. 1.

In the training stage for both algorithms, we used 16 of the 5-second speech signals along with their associated per-frame targets T(n). Different numbers of hidden nodes were examined. We found that 7 to 10 hidden nodes generally result in good performance. The number of taps is 3. We also performed some experiments without time delay, however, the performance was not very impressive.

Figure 5 shows the output of a 2-layer (10 hidden nodes) neural network, trained with Algorithm 1, due to a noisy input signal. In the MOS scale, this speech signal is rated as "fair". Figure 6 shows the output when we train the network with Algorithm 2. Here the speech signal is rated as "poor". One can see from both figures that the output fluctuates around actual per-frame quality score. As noted before, for the noise-free regions the prediction is less accurate due to he overlap between the noise and speech. It is possible, even for a human, to confuse some speech signals with noise. Since we rely on the overall (or average) score, these fluctuations would be averaged out as shown in correlation analysis.

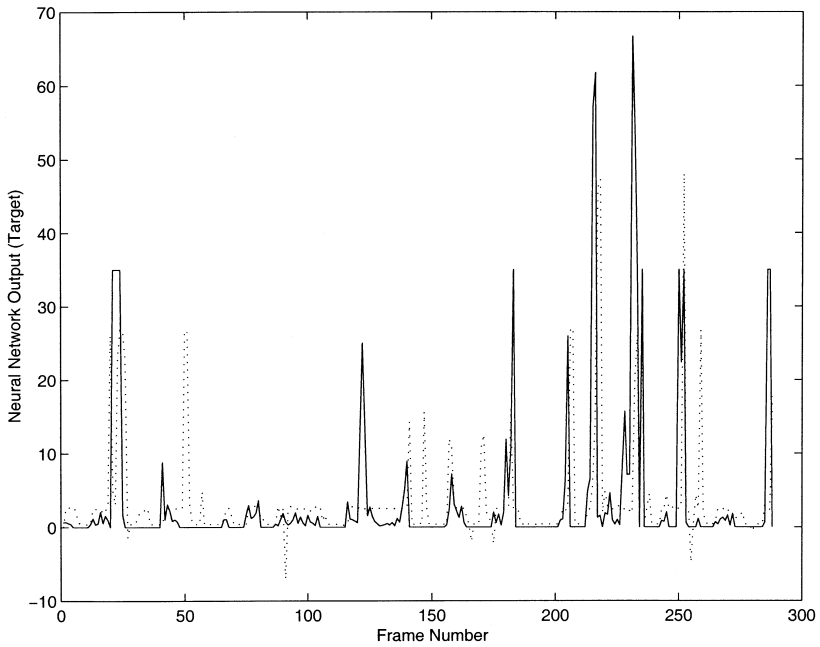

Fig. 5. Per-frame quality score (solid line) and its prediction by the neural network (dashed line) for speech through GSM channel. Algorithm 1 is used for training.

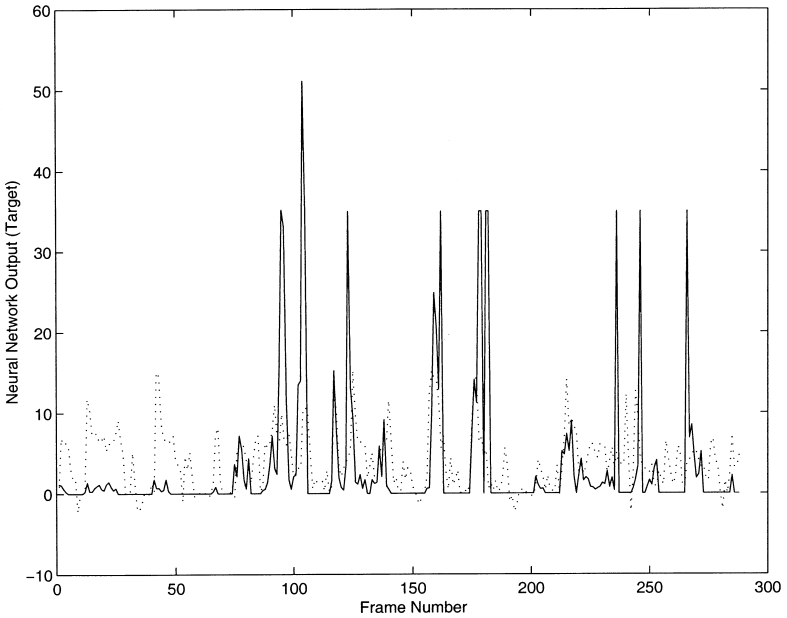

Fig. 6. Per-frame quality score (solid line) and its prediction by the neural network (dashed line) for speech through GSM channel. Algorithm 2 is used for training.

The more standard way in comparing the objective tests is to find their correlation coefficients with known test measures. Here we find the correlation between some distance measures with the average per-frame actual speech quality score, Q_a, defined as

$$Q_a = \frac{1}{L} \sum_{n=1}^{L} T(n) \tag{4}$$

The following three measures are considered.

Measure 1: The predicted quality, Q_p, which is the average of the neural network's output defined as

$$Q_p = \frac{1}{L} \sum_{n=1}^{L} Y(n)$$

(5)

Measure 2: (Histogram.) We apply the features of a clean signal and find its output $Y_c(n)$. This output is quantized using a step of size 16 to get $q_c(n)$. Then we find the associated histogram vector, h_c, of this quantized output. We repeat the same procedure to the noisy signal and obtain the quantized output q(n) and the histogram vector, h. The distance measure is defined as the Euclidean distance $||h_c-h||$. Note that the clean speech signal could be different from the noisy one.

Measure 3: (Transition Probability matrix.) We first obtain $q_c(n)$ and q(n) as described in Measure 2. We model $q_c(n)$ and q(n) as Markov processes. We calculate the first-order transition probabilities matrices, P_c and P for the clean and noisy speech, respectively. The distance measure is defined as the Euclidean distance between the elements of the two matrices, $||P_c-P||$.

Table 3 shows the correlation coefficients for the three distance measures with the average per-frame quality score. It is obvious from this table that Algorithm 2 has better generalization when the first distance measure is used. The training time and memory are, however, more exhaustive. We see also that for a simpler training algorithm, Algorithm 1, the histogram and transition matrix measure could increase the correlation coefficient. On the other hand, if a time-delay network is not used, the best correlation found does not exceed 0.88.

Table 3. Correlation Coefficients with different distance measures

	Measure 1	**Measure 2**	**Measure 3**
Algorithm 1	0.9294	0.9305	0.9503
Algorithm 2	0.9491	0.9316	0.9268

Cellular TIMIT (CTIMIT) database

In this part, the speech samples used for the training and testing stages are taken from CTIMIT database [6]. CTIMIT is the cellular version of the well-known TIMIT phonetic database. TIMIT was established in a joint effort among the Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI), and Texas Instruments, Inc. (TI). There are 630 speakers from 8 dialects US regions saying 10 sentences each, 2 sentences were common for all, 5 were chosen from a list of 450 phonetically balanced sentences selected by MIT, and 3 were randomly selected by TI. All speakers are adult with 30% female and 70% male.

CTIMIT was developed under the Lockheed-Martin Sanders Inc. sponsorship in 1996 by Kathy, Brown and Bryan [6]. They performed drive tests and recorded the transmitted TIMIT speech samples via a digital cellular system. Some of TIMIT speech files were missed from CTIMIT because of dropped-out call occurrences.

Comment [ρ.1]: Kathy L., Brown E. and Bryan George, "CTIMIT: A speech corpus for the cellular environment with applications to automatic speech recognition", Proc. ICASSP95, p-p 105-108, May 1995.

As a pre-processing step for the speech samples, we normalize the long-term mean and variance to zero and one, respectively, for both input and output speech signals and we removed any time shifts. Using a sampling rate of 8000 sample/sec, we form 64-ms time domain frames with 512 samples/frame with 50% overlapped. As discussed in the last subsection, the LAR coefficients and frame energy are then extracted from each speech frame to form the input feature vectors X(1), X(2), …, X(N), where N is the number of frames in the speech sample. The "customized" quality measure, T(n), of the n^{th} frame is calculated according the frame-energy plane partitioning given in the previous subsection (Please see Figure 4). Again we have the training se of the 2m+1-tap time-delay network, (χ(n), T(n)), where

$$\chi(n)=\{X(n-m), X(n-m+1), ..., X(n), ..., X(n+m)\}$$

We performed three types of experiments based on the used speech samples. In the first experiment, we use speaker-dependent speech samples. The speaker is fixed and the pronounced sentences are varying. The second experiment uses text-dependent speech files, where different speakers pronounce a single sentence. The third experiment is the more general one where we have a mixture of speakers pronouncing different sentences. Our simulation results are presented in the following.

A. Speaker-dependent speech samples. In this part, we have two speakers (two males and one female); nine different speech samples of each are selected. We train the neural network with the speech sample of one speaker. A leave-one training method is used, where eight of the speech samples are used for training and the last sample is used for testing. We repeat this procedure leaving a different test sample each time until all samples are covered. The trained neural network gives the predicted quality score of each frame in the test sample. We average these scores to get the overall predicted quality score Q_p (as defined in (4)). Therefore, for each speaker we have nine predicted quality scores corresponding to the nine sentences.

We obtain the correlation between the predicted quality score, Q_p, and the actual one, Q_a (as defined in (5)). These experiments are carried out for 3 and 5 taps and the LPC orders are 4,5 and 6. In Table 4 we show the correlation results averaged for the two male speakers and the one female speaker. Clearly this table indicates high correlation between the neural network quality prediction and the actual one. There is no major different between 3 and 5 taps. However, the LPC order should be maintained small.

Table 4. Correlation between the actual quality score and the neural network prediction for the speaker dependent case.

	LPC Order

	Taps	4	5	6
Male Speakers	3	0.88	0.86	0.65
	5	0.89	0.92	0.56
Female Speaker	3	0.91	0.92	0.81
	5	0.93	0.91	0.82

A detailed score result is also shown in Fig. 7. The actual quality score is shown as compared with the neural network prediction for the nine speech samples uttered by one of male speaker.

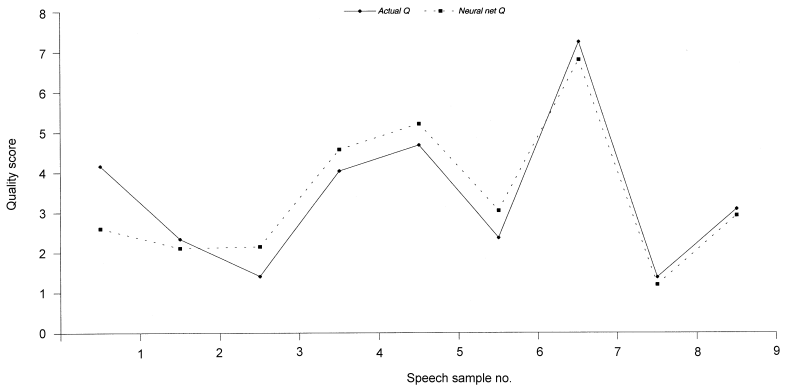

Fig. 7. Comparison between the actual quality score and the predicted one. The LPC order is 5.

B. Text-dependant case. Here the text of the speech is fixed, and the speakers are different. We have four different texts. Each text is uttered by nine speakers. The first two texts are uttered by a group of male speakers. The other two texts are uttered by females. Again a leave-one training/testing method is used with the time-delay neural network. Table 5 shows the testing results averaged separately for each speaker gender. We can see that the lower LPC orders perform better. This agrees with the conclusion reached by others as mentioned in the introduction.

Table 5. Correlation between the actual quality score and the neural network prediction for the text-dependent case.

		LPC Order		
	Taps	4	5	6
Male Single Speech	3	0.86	0.84	0.46
	5	0.90	0.78	0.67
Female Single Speech	3	0.95	0.92	0.85

5	0.93	0.95	0.90

It can be observed from the previous tables that the network predicts the speech quality with high accuracy for female rather than male samples. Single speech (text-dependent) case has the best prediction performance. There is a practical application for the single speech test where standard words (or phrases) can be transmitted (from any speaker) and a system (pre-trained to this phrase) can rate the received speech quality without referring to the original one.

C. Speaker-and-text-independent case. This case is the most general one where we collected speech samples of different sentences as uttered by different speakers. A training algorithm with better generalization (Resilient Back Propagation) is used to get the best fit to the training data set. We use 63 different speech samples, which include the previous samples in case A and B along with other samples. Similar to the previous cases we use a leave-one training/testing strategy. As shown in Table 6, the obtained results have less correlation than two preceding special cases, but still within accepted ranges.

Table 6. Correlation between the actual quality score and the neural network prediction for the speaker-and-text-independent case.

	LPC Order		
Taps	**4**	**5**	**6**
3	0.72	0.87	0.68
5	0.76	0.80	0.71

The 5th order LPC-based quality prediction gives the best correlation compared to the higher or lower orders. Figure 3 shows the difference between the actual Euclidean distance and the simulation output for the data set samples consisting of 9 speech files of mixed text and speakers.

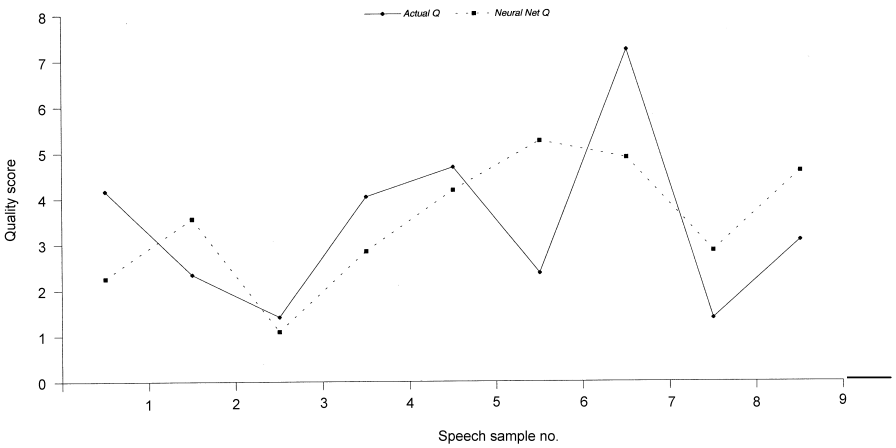

Fig. 8. Comparison between the actual quality score and the predicted one for the speaker-and-text-independent case. The speech samples are the same used in Fig. 7.

Conclusion

In this paper, we studied the problem of speech quality rating using only the output speech signal without the need to the reference signal. A time-delay multilayer neural network is trained to automatically perform the speech quality rating. A customized per-frame quality score is introduced to make the training possible. For the simple simulated channel, the task was simple and the performance is remarkable. However, for the actual GSM channel the problem is more difficult. Training should consider many possible forms of noise. Nevertheless, the obtained correlation reached 0.95 for the small speech sample size used in our experiment.

For the CTIMIT database, which contains much more speech samples, we consider three cases. The first case deals with a speaker-dependent speech signal. The second case is restricted to the text-dependent speech signal. The third case considers a mixture of different speakers and different texts. The first two cases gave very high correlation (above 90%) between the actual quality score and the predicted one. In the more practical case, the third one, the trained neural network achieved also good performance (above 80% correlation). However, as expected, the performance in this case is less than the other two special cases.

One can be sure that the output-based speech assessment methods cannot reach that of the input/output (reference signal) methods. The Output-based speech quality assessment is a difficult problem. Any algorithm will never be perfect since there is an overlap between noise and speech signals. For example, when a muted frame is received, it is difficult to decide if a silence occurs in the real speech or the GSM receiver made that mute to avoid annoying noise sound.

This paper could well lead to an automated speech assessment mechanism in mobile phones and possibly in the IP protocol over the Internet. To reach a more conclusive result, a larger speech database is needed. However, this will make the training task more difficult. To make the training of the neural network more tractable, one needs to use different networks for different environments. Then the output of the networks is combined using a well-designed statistical method. Another extension to this

research would be the utilization of the Markovian speech model for better distinguishing between the speech signal and noise.

Acknowledgment. I would like to acknowledge the support from the College of Engineering Research Center, King Saud University. I would also like to thank Engr: Abdullah Al-Mubadal and Engr: Mohammad Al-Shaie, both from Lucent Technology, for their help in obtaining and preprocessing of CTIMIT database. My thanks are also extended to Eng. Zaygham Nawaz for proofreading this manuscript.

References

[1] Deller, John, R., Hansen, John H. L. and Proakis, John G. *Discrete-Time Processing of Speech Signals.* IEEE Press, 1999.

[2] ITU-T COM 12-62: “Results of Processing ITU Speech Database Supplement 23 with the End-to-End Quality Assessment Algorithm”. 'PACE', 09/98.

[3] Voran, S. "Advances in Objective Estimation of Perceived Speech Quality". *Proceedings of the 1999 IEEE Workshop on Speech Coding for Telecommunications*, Porvoo, Finland, (1999), pp. 716-721.

[4] Jin, Chiyi and Kubichek, Robert. “Vector Quantization Techniques for Output-Based Objective Speech Quality”. *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*, (ICASSP-96), (1996), vol.1, p-p. 491-494.

[5] Haykin, S. *Neural Networks*, New Jersey: Prentice Hall, Inc., 1999.

[6] Kathy, L, Brown, E and Bryan, George. “CTIMIT: A Speech Corpus for the Cellular Environment with Applications to Automatic Speech Recognition”. *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*, (IACSSP-95), (1995), 105-108.

[7] Juric, P. "An Objective Speech Quality Measurement in the Qvoice". *Proceedings of the 1998 IEEE 5th International Workshop on Systems, Signals and Image Processing* (IWSSIP'98), (1998), 156-163.

[8] Riedmiller, M. and Braun, H. “A Direct Adaptive Method for Faster Backpropagation learning The RPROP algorithm”. *Proceedings of the IEEE International Conference on Neural Networks*, San Francisco, (1993), 586-591.

[9] Foresee, F. and Hagan, M. “Gauss-Newton Approximation to Bayesian Regularization”. *Proceedings of the 1997 International Joint Conference on Neural Networks*, (1997), 1930-1935.

[10] Hagan, M and Menhaj, M. “Training Feedforward Networks with the Marquardt Algorithm”. *IEEE Transactions on Neural Networks*, 5, No. 6 (1994), 989–993.

تقويم جودة الصوت في شبكات الهاتف النقال

خالد بن عبد العزيز المعشوق

قسم الهندسة الكهربائية، كلية الهندسة، جامعة الملك سعود
ص.ب. ٨٠٠ ، الرياض ١١٤٢١، المملكة العربية السعودية
البريد الإلكتروني: mashouq@ksu.edu.sa

(استلم في ٢٠٠١/٠٦/٠٥م،وقبل للنشر في ٢٠٠١/١٢/٢٩م)

ملخص البحث. يعتمد مستخدمو الهواتف النقالة على حدسهم الشخصي لتقويم جودة الخدمة المقدمة من مشغل الشبكة. في هذا البحث نقترح طريقة لتقويم الجودة في الهاتف النقال معتمدة فقط على الخرج. هذا التقويم عندما يؤخذ له المتوسط عبر الزمن سيكون مفيداً للمقارنة بين شبكات الهاتف النقال المختلفة. أضف إلى ذلك أن هذا التقويم يمكن أن يدخل في حساب تعرفة المكالمة الهاتفية.

هنا نستخدم شبكة تأخير الوقت العصبية ذات الطبقات المتعددة لعمل التقويم لجودة الصوت، وذلك بعد مرحلة مناسبة من التدريب. تتكون معطيات التدريب من خصائص الصوت (مثل معاملات التنبوء الخطي LPC وقدرة الإطار). وهدف التدريب هو نوع معدل من أنواع قياس الجودة. قمنا بعمل تمثيل حاسوبي موسع باستخدام الطريقة المقترحة. واستخدمنا عينات صوت من قناة محاكاة وقناة GSM حقيقية بالإضافة إلى قاعدة بيانات CTIMIT الموسعة، و قد أوضح هذا التمثيل الحاسوبي أن الطريقة المقترحة بإمكانها التنبؤ بجودة الصوت بدقة معقولة.