Perceived Speech Quality Prediction for Voice over IP-based Networks

Lingfen Sun and Emmanuel C. Ifeachor
Department of Communication and Electronic Engineering,
University of Plymouth, Plymouth PL4 8AA, U.K.

Abstract – **Perceived speech quality is the key metric for QoS for VoIP applications. The primary aims of the study reported in the paper are to carry out a fundamental investigation of the impact of packet loss and talkers on perceived speech quality using an objective method to provide the basis for developing an artificial neural network (ANN) model to predict speech quality for VoIP. The impact of packet loss (e.g. loss burstiness, loss patterns and packet size) and different talkers on speech quality was investigated for three modern codecs (G.729, G.723.1 and AMR) using the new ITU PESQ algorithm. Results show that packet loss burstiness, loss locations/patterns and the gender of talkers have an impact on perceived speech quality. Packet size has, in general, no obvious influence on perceived speech quality for the same network conditions, but the deviation in speech quality depends on packet size and codec. Based on the investigation, we used talkspurt-based conditional and unconditional packet loss rates (instead of network packet loss rates because they are perceptually more relevant), codec type and the gender of the talker (extracted from decoder) as inputs to an ANN model to predict speech quality directly from network parameters. Results show that high prediction accuracy was obtained from the ANN model (correlation coefficients for the test and validation datasets were 0.952 and 0.946 respectively). This work should help to develop efficient, non-intrusive QoS monitoring and control strategies for VoIP applications.**

Keywords -- **Voice over IP, Speech Quality, Artificial Neural Network, Packet Loss, Codecs, Talker Dependency**

I. INTRODUCTION

In real-time voice communication, perceived speech quality, expressed as a Mean Opinion Score (MOS), is the key metric for Quality of Service (QoS) as it provides a direct link to quality as perceived by the end user. MOS values may be obtained by subjective listening tests [1] or by objective perceptual measurement methods, such as the new ITU algorithm, the Perceptual Evaluation of Speech Quality (PESQ) [2].

In voice over IP applications, statistical and artificial intelligence methods are being developed to predict speech quality directly from IP network parameters for QoS monitoring and control purposes [3][4][5][6]. The E-model as well as artificial neural networks (ANN) models have recently been used to predict speech quality from network parameters [4][5][6][7]. Unlike the E-model which is static, artificial neural networks models can adapt to the dynamic environment of IP networks, such as the Internet, because of its ability to learn. However, the success of ANN approach in voice over IP depends on the ability of the models to fully learn the non-linear relationships between IP networks impairments (e.g. packet loss and jitter) and the perceived speech quality.

At present, both the E-model and ANN based methods rely on databases obtained by subjective tests. Unfortunately, subjective listening tests are costly and time-consuming and as a result the databases are limited and do not cover all the possible scenarios and network conditions. The impact of a variety of network parameters (e.g. loss rate, burstiness, loss pattern and packet size) on perceived speech quality remains unclear. Further, little attention has been paid to talker dependency and the development of current ANN models are based on a limited number of codecs. The assumptions about the behaviour of network losses do not reflect reality. For example, only the numbers of consecutively lost packets (e.g. 1 to 5) were used to represent different bursty losses.

The aims of the study reported in this paper are three fold: (1). to undertake a fundamental investigation of the impact of packet loss (e.g. loss rate and loss pattern) on perceived speech quality using an *objective* measurement algorithm (the new ITU PESQ algorithm), (2) to investigate the impact of different talkers on perceived speech quality, and (3) to develop a robust ANN model that exploits perceptually relevant information for speech quality prediction.

The remainder of the paper is organised as follows. In Section II, the experimental system used in the study is introduced. In Section III, a fundamental study of the impact of packet loss and different talkers on speech quality is presented. The study provides a basis for the development of the ANN model for speech quality prediction which is presented in Section IV. Section V concludes the paper.

II. SIMULATION SYSTEM

A block diagram of the system that was used in the study is depicted in Figure 1. It is a PC-based software system that allows the simulation of key processes in voice over IP. It enables the simulation of a variety of network conditions and objective measurement of the effects on perceived speech quality. The system includes a speech database, an encoder/decoder, a packet loss simulator, a speech quality measurement module, a parameter extraction and an ANN model. The speech database is taken from the TIMIT data set [15] and ITU dataset [2]. Speech files from different male and female talkers are chosen for talker dependency analysis and to generate a data base for ANN model development

Three modern codecs were chosen for the study. These are G.729 CS-ACELP (8 Kbps) [9], G.723.1 MP-MLQ/ACELP (5.3/6.3 Kbps) [10] and Adaptive Multi-Rate (AMR) codecs with eight modes (4.75 to 12.2 Kbps) [11].

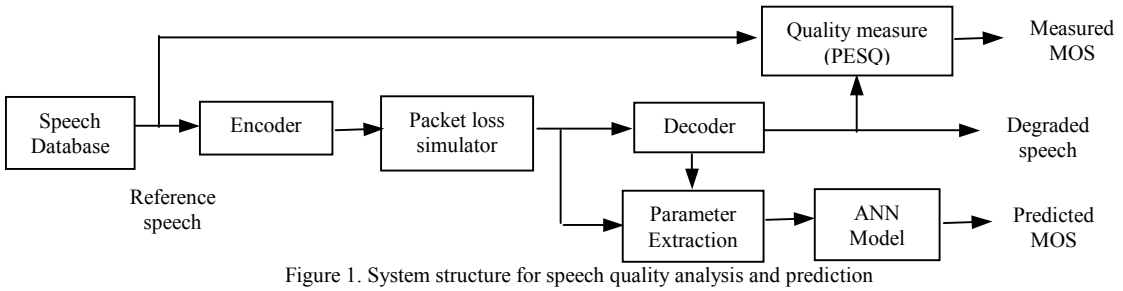


Figure 1. System structure for speech quality analysis and prediction

A 2-state Gilbert model was used to simulate packet loss (see Figure 2). The Gilbert model is well known to represent the packet loss behaviour of a real network, even after the late arrival loss due to jitter is taken into account (if a packet arrives too late, it will be discarded by jitter buffer) [8]. In the figure, State 0 is for a packet received (no loss) and State 1 is for a packet dropped (loss). p is the probability that a packet will be dropped given that the previous packet was received. q is the probability that a packet will be dropped given that the previous packet was dropped. q is also referred to as the conditional loss probability (*clp*). The probability of being in State 1 is referred to as unconditional loss probability (*ulp*). The *ulp* provides a measure of the average packet loss rate. It is given by:

$$ulp = p/(p+1-q)$$

The conditional loss probability (*clp*) and unconditional loss probability (*ulp*) are used in the paper to characterise the loss behaviour of the network.

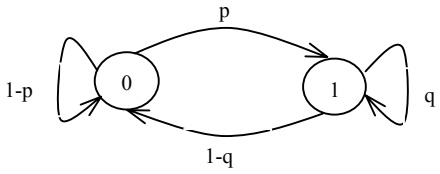


Figure 2. Gilbert model

In our system, the new ITU perceptual measurement algorithm, the Perceptual Evaluation of Speech Quality (PESQ), is used to measure the perceived speech quality under different network conditions and for different talkers. The PESQ compares the degraded speech with the reference speech and computes an objective MOS value in a 5-point scale. In the study, the MOS score obtained from the PESQ is referred to as the 'measured MOS' to differentiate it from the 'predicted MOS' obtained from the ANN model. The Parameter Extraction module is used to extract salient information from the IP network and the decoder (including the codec type and network packet loss). In real VoIP applications, codec type and packet loss would be parsed from the RTP header. After processing, the information is fed to the ANN model to predict speech quality.

To provide a basis for the development of a robust ANN model, a fundamental study of the impact of packet loss and gender on perceived speech quality was undertaken. This enabled us to determine the relevant parameters to be used as input to the neural networks model to predict speech quality. The study is based on three modern codecs described above.

III. Perceived Speech Quality Analysis

A. *The impact of packet loss on perceived speech quality*

We first investigated how packet loss burstiness affects perceived speech quality. A fixed packet size was set for different codec. Different network *ulp* and *clp* were chosen and the corresponding MOS score was calculated. To account for a wide range of possible type of packet loss patterns and locations, 300 different initial seeds for random number generation were chosen for each pair of *ulp* and *clp*. The average MOS score and 90% Confidence Internal (CI) were calculated. The results for G.729 and G.723.1 (6.3 Kb/s mode) are shown in Figures 3 and 4. The length of the test speech sentence was about 12 seconds. The packet size for G.729 and G.723.1 was 2 and 1 frames/packet, respectively. No VAD was activated.

From Figures 3 and 4, it can be seen that the *clp* has an obvious impact on the perceived speech quality even for the same average loss rate (*ulp)*. When burst loss increases (*clp* increasing), the MOS score decreases and the variation of the MOS score (shown in CI) also increases. This is because losses may occur more concentrated with high burst losses and this results in large variation in the MOS scores due to the locations of the losses, whereas it may occur evenly in low burst loss cases which results in small deviations. There is only a small difference between the results for G.729 and G.723.1, when *ulp* is 10%, and *clp* is from 40% to 70%.

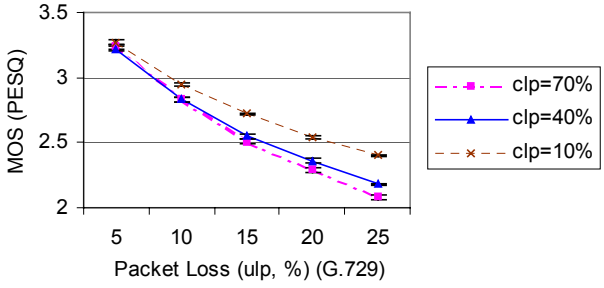


Figure 3. MOS vs packet loss for G.729

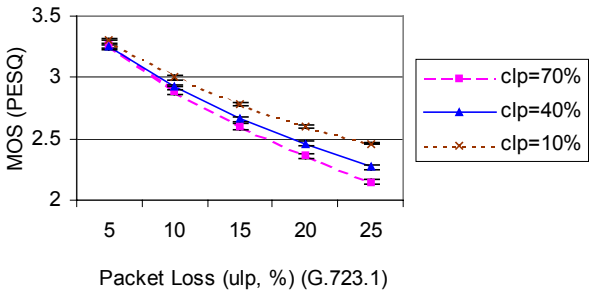


Figure 4. MOS vs packet loss for G.723.1

We then investigated how packet size affects perceived speech quality. A fixed *clp* (40%) was set and *ulp* was changed from 0% to 40% in 5% increment. The packet size was changed from 1 to 6 frames/packet. As before, 300 different initial seeds were generated. The average and the standard deviation of MOS scores for G.729 are shown in Figure 5 (a) and (b). The standard deviations for the MOS scores for AMR (12.2 Kb/s mode) are shown in Figure 6.

From Figure 5 (a), it can be seen that the packet size has in general no obvious influence on perceived speech quality for a given packet loss rate. Similar results were obtained for G.723.1 and AMR. However, the variation in speech quality for the same network loss rate depends on packet size and codec, as shown in Figures 5 (b) and 6.

When packet loss rate is lower and packet size is larger, the higher values of the standard deviation of MOS score means larger variation in speech quality for the same network conditions. The variation in quality is the main obstacle in the prediction of speech quality directly from network parameters. When packet loss (e.g. *ulp* and *clp*) was calculated from the Gilbert model, the loss is perceptual irrelevant as some losses may occur during a silent period which is imperceptible [13]. As a solution, we proposed to calculate losses only during talkspurts.

A network packet may include a speech talkspurt frame or a silence frame. The number of silence frames depends on whether VAD (Voice Activity Detection) is activated at the encoder side. If VAD is activated, silence frame only represents SID (Silence Insertion Description) frame. Here we combined the information from decoder's VAD indicator and network packet loss, and calculated the *ulp* and *clp* according to Gilbert model only during the speech talkspurt. In this case, State 1 in Figure 2 represents loss during a talkspurt, and State 0 represents no loss or loss during a silence period. We use *ulp(VAD)/clp(VAD)* to differentiate them from network *ulp/clp*. As the calculation of *ulp(VAD)/clp(VAD)* was based on speech frame, the loss pattern and the factor of packet size have both been taken into account. The codec type, *ulp(VAD) and clp(VAD)* were identified as inputs for neural network analysis.

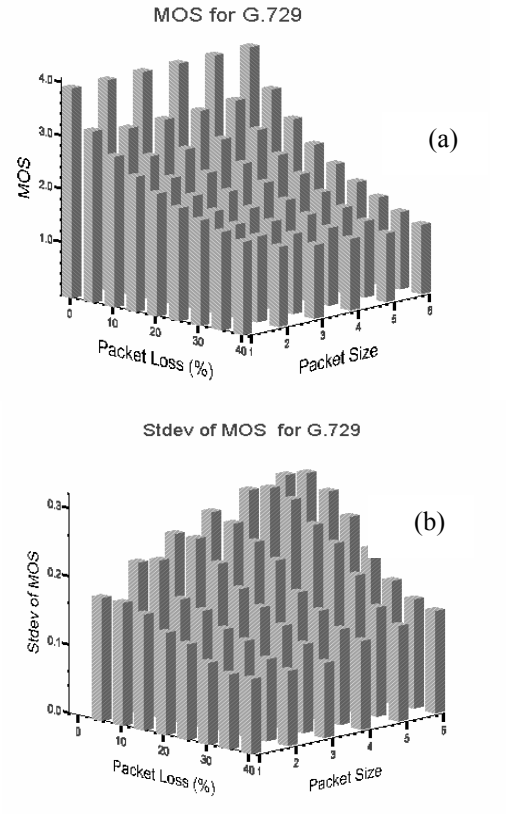


Figure 5. (a) Average MOS and (b) Standard Deviation of MOS for G.729

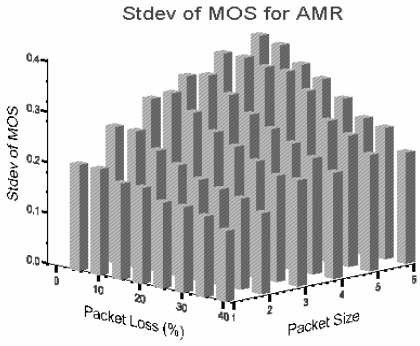


Figure 6. Standard Deviation of MOS for AMR

B. *The impact of talker on perceived speech quality*

This experiment was to investigate whether difference in talker (male or female) has an effect on perceived speech quality for the same network conditions. We first chose 6 English speakers (3 male and 3 female) from the TIMIT [15] Data Set (dialects 1 and 2). Speech files from the same talker were grouped to form a longer file (about 10s). The activity factor [16] was about 0.82 for all files.

We altered *ulp* from 0 to 30 % in 5% increment, set *clp* to 10% and packet size to 2 for G.729 (no VAD). As before, 300 different initial seeds were chosen. The average MOS

scores for the six talkers are shown in Figure 7. The speech file name starts with letter "f" for female and "m" for male.

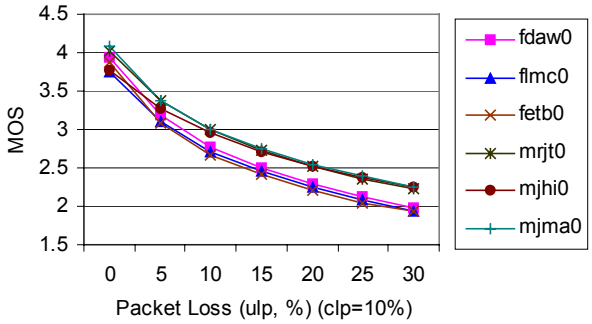


Figure 7. MOS vs Loss Rate for English speakers

We further tested another four speech files (2 male and 2 female of Dutch speakers) from an ITU data base [2]. Each speech file was about 8s, with about 45% to 49% activity. The results for the four speech files are illustrated in Figure 8.

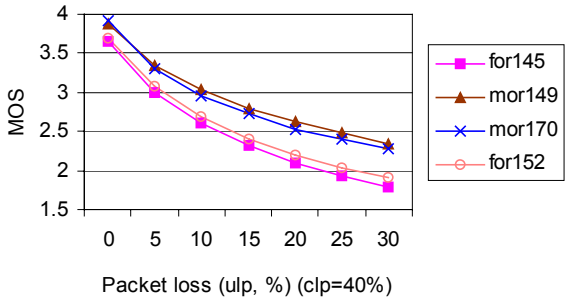


Figure 8. MOS vs Loss Rate for Dutch speakers

From inspection Figures 7 and 8, it can be seen that the impact of different talkers on perceived speech quality appears to depend mainly on the gender of the talker (male or female), irrespective of language and accent. The quality for the female talker tends to be worse than that of the male talker for the same network impairments. This effect is more obvious when loss increases.

The reason for talker dependency is due to the codec algorithm. As the G.729, G.723.1 and AMR are all CELP-based codecs, the use of linear predictive model of speech production can lead to variations in codec performance with different talkers and languages [14]. In this paper, we focused on gender issue, and identified gender as one of the input parameters for neural network analysis. The gender can be decided according to pitch delay derived from the decoder.

IV. Prediction of Perceived Speech Quality using Artificial Neural Network (ANN)

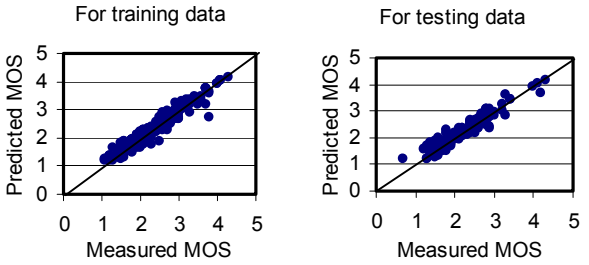
In order to model the relationships between network impairments and perceived speech quality, a neural network model was developed to learn the non-linear mapping from network parameters to MOS score.

Four variables were identified as inputs to the neural network model, namely: codec type, gender, *ulp(VAD)* and *clp(VAD)*. The predicted MOS score was the only output (see Figure 1). Stuttgart Neural Network Simulator (SNNS) package [12] was used for neural network training and testing. A three-layer feed-forward neural net architecture and the Standard Backpropagation learning algorithm were selected for simplicity.

In order to train and test the neural network, a database was generated from 2 talkers (1 male and 1 female) and three codecs, G.729, G.723.1 (6.3Kb/s) and AMR (12.2 Kb/s). For dual–mode G.723.1 and eight-mode AMR, only one mode was chosen for simplicity. The network loss *ulp* was set to 0, 10, 20, 30 and 40% and *clp* was set to 10, 50 and 90%. The packet size was set to 1, 2, 3, 4 and 5 frames/packet. For each case, an initial seed was chosen randomly to cater for a range of possible loss patterns. The state transitions were counted according to the Gilbert model (see Figure 2). In order to compare the results to real network loss and talkspurt-based network loss, the real loss rate at the end of the test sentence, *ulp(Real)/clp(Real)* and loss rate during talkspurt, *ulp(VAD)/clp(VAD)* were calculated at the same time. The difference between *ulp(Real)/clp(Real)* and *ulp/clp* is due to pseudo-random number generation and initial seeds selection. For each loss condition, the perceived speech quality between the reference and degraded speech files was calculated using PESQ. A total of 362 samples (patterns) were generated. 70% of the samples were chosen randomly as the training set and the remaining 30% as the testing set.

Different network structures (e.g. the number of neurons in the hidden layer and the parameters of learning algorithm) were investigated to determine a suitable architecture for ANN model. Comparing the predicted MOS score from the ANN model and the measured MOS, we obtained a maximum Correlation Coefficient (ρ) of 0.967 and an average error of 0.12 for the training set. For the testing set, ρ was 0.952 and the average error was 0.15. The learning rate (η) was 0.4 and the maximum difference (d_{max}) was 0.01 for a 4-5-1 net. The scatter diagrams of the predicted versus the measured MOS scores for the training and test data sets are illustrated in Figure 9 (a) and (b). Increasing the number of neurons in the hidden layer did not improve the prediction accuracy. However, when *ulp(Real)/clp(Real)* was used instead of *ulp(VAD)/clp(VAD),* the Correlation Coefficients for the training and testing datasets both dropped by 2-3 percent. This suggested that *ulp(VAD)/clp(VAD)* are better for speech quality prediction than *ulp(Real)/clp(Real)*. We also investigated the effect of including packet size as an input to the neural net (i.e. 5 inputs) and obtained similar results. This suggested that packet size may not be necessary as an input to the neural network.

As the training and test data sets were from the same talkers, we further generated a validation data set from another male and female talkers and set the different network loss conditions (*ulp:* 5, 15, 25, 35%, and *clp:* 30, 70%). A total of 210 new patterns were generated and used to validate the trained ANN model. We obtained ρ of 0.946 and an average error of 0.19. This suggested that the neural network model works well for speech quality prediction in general.



(a) (b)

Figure 9. Predicted MOS vs. Measured MOS for (a) training data and (b) test data

The correlation coefficients obtained from training, test and validation datasets are between 0.946 to 0.967. It seems difficult to improve the performance further from neural network side. We think this is mainly due to the following two reasons. (1). *ulp(VAD)/clp(VAD)* is still not accurate enough to express perceptually relevant loss information for some loss patterns/locations; (2). Objective MOS scores from PESQ may not be as accurate as subjective MOS scores for some loss conditions. Our subjective test results have also confirmed that PESQ shows higher sensitivity than human subjects in high bursty conditions, especially in the case of missing words, whereas, it shows lower sensitivity than human subjects in lower bursty cases for G.729.

V. CONCLUSIONS

We have investigated the impact of packet loss, codec and talker on perceived speech quality based on the new ITU PESQ measurement algorithm and developed an ANN model for speech quality prediction. Results show that the loss pattern, loss burstiness and the gender of the talker have an impact on perceived speech quality. Packet size has in general no obvious influence on perceived speech quality for a given packet loss rate, but the deviation in speech quality depends on packet size and codec. The quality for the female talker tends to be worse than that of the male talker for the same network impairments. Based on the investigation, we used talkspurt-based conditional and unconditional packet loss rates (instead of the network packet loss rates because they are perceptually more relevant), codec type and the gender of the talker (extracted from decoder) as inputs to an ANN model to predict speech quality directly from the network parameters. Results show that high prediction accuracy was obtained from the ANN model (correlation coefficients of the test and validation datasets are 0.952 and 0.946 respectively). This work should help to develop efficient, non-intrusive QoS monitoring and control strategies for VoIP applications.

Future work will focus on further analysis of the loss pattern in order to incorporate more information from speech content (e.g. signal energy, voiced/unvoiced information) and to obtain more accurate perceptually relevant loss information. The neural networks based model will be optimised using real Internet VoIP trace data. More speech data will be investigated for the analysis of talker dependency.

ACKNOWLEDGEMENT

We are grateful to Acterna for sponsorship.

REFERENCES

[1] ITU-T Recommendatiaon P.800, Methods for subjective determinatio of transmission quality

[2] 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.

[3] L.A.R. Yamamoto, J.G.Beerends, Impact of network performance parameters on the end-to-end perceived speech quality, Expert ATM Traffic Symposium, Mykonos, Greece, Sep. 1997

[4] S. Mohamed, F. Cervantes-Perez and H. Afifi, Integrating Networks Measurements and Speech Quality Subjective Scores for Control Purposes, IEEE Infocom 2001

[5] S. Mohamed, F. Cervantes-Perez and H. Afifi, Audio Quality Assessment in Packet Switched Networks: an "Inter-Subjective" Neural Network Model, Proc. International Conference on Information Networks, Japan, Jan. 2001.

[6] A. D. Clark, Modeling the Effects of Burst Packet Loss and Recency on Subjective Voice Quality, IPTEL'2001, pp. 123-127, April, New York, 2001

[7] ITU-T Recommendation G.107, The E-model, a computational model for use in transmission planning

[8] W. Jiang and H. Schulzrinne, QoS Measurement of Internet Real-Time Multimedia Services, Technical Report, CUCS-015-99, Columbia University, Dec. 1999, http://www.cs.columbia.edu/~wenyu

[9] ITU-T Recommendation G.729, Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP), March 1996

[10] ITU-T Recommendation G.723.1, Dual Rate Speech Coder for Multimedia Communication Transmitting at 5.3 and 6.3 kbit/s, March 1996

[11] ETSI EN 301 704 V7.2.1, Digital cellular telecommunications system; Adaptive Multi-Rate (AMR) speech transcoding

[12] http://www-ra.informatik.uni-tuebingen.de/SNNS/

[13] L. F. Sun, G. Wade, B. Lines and E. Ifeachor. Impact of Packet Loss Location on Perceived Speech Quality, IPTEL'2001, April, New York, 2001

[14] P. A. Barrent, R. M. Voelcker and A. V. Lewis, Speech transmission over digital mobile radio channels, BT Technol J Vol 14 No. 1 January 1996, pp.45-56

[15] J. S. Garofolo, L.F. Lamel, W. M. Fisher, et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993

[16] ITU-T Recommendation P.56 - Objective measurement of active speech level