Perceived Speech Quality Estimation Using DTW Algorithm

Ivan Kraljevski, Slavcho Chungurski, Zoran Gacovski, and Sime Arsenovski

Abstract – **In this paper a method for speech quality estimation is evaluated by simulating the transfer of speech over packet switched and mobile networks. The proposed system uses Dynamic Time Warping algorithm for test and received speech comparison. Several tests have been made on a test speech sample of a single speaker with simulated packet (frame) loss effects on the perceived speech. The achieved results have been compared with measured PESQ values on the used transmission channel and their correlation has been observed.**

Keywords – **Dynamic Time Warping, Mean Opinion Score, Perceptual Evaluation of Speech Quality**

I. Introduction

Speech quality measuring is a very important factor in the process of providing the quality of service for voice telecommunications networks. This applies particularly to wireless communication networks such as 3G mobile networks, IEEE 802.11 and Bluetooth. For voice information transport in these networks - packets or frames are used, and communication channel characteristics are very different compared to those found in wired communications networks (regarding transmission errors presented in the form of packet or frame error ratio). Understanding and estimation of these parameters are of great significance for the design or optimization of telecommunication services or network infrastructure [1].

Speech quality is defined by the way how the listeners value the perceived speech signals on the receiver side of the communication channel. Due to the evermore increasing complexity of communication networks and the number of parameters which characterize the communication channel, it is much more difficult to establish a straightforward relation between transport parameters and the perceived speech quality. Besides that, there is a problem of accurate extraction or estimation of the exact communication parameters, and not every time does the perceived speech quality correspond with the measured or estimated transport parameters of the received speech.

In this paper, research of the usability of DTW (Dynamic Time Warping) method for speech quality estimation is presented. This is a sequence matching algorithm between the test and received speech sequences performed after transmission over packet-switched or mobile communication channels. The DTW algorithm compares arrays of mel-cepstral coefficients which simulate the perception of human auditory system and it is usually used as a building block for simple speech recognizers [2].

Three speech codecs have been used in the experiments, G.711 [3], AMR 12.4 kb/s (compatible with GSM-EFR) [4] and G.729. The effects of packet receiving errors are modeled for a random and bursty packet loss. Low bit rate (high compression ratio) codecs are used to reduce the required bandwidth, but distort the original waveform significantly before it is even transmitted. The compressed speech produced by such codecs is also more sensitive to packet loss [5].

Different values for similarity metrics are observed after comparing the test and received speech sequences with varying the values of the possibility of packet loss errors and the possibility of introducing burstiness during packet errors (expressed as a percent of lost packets or frames). Achieved results have been compared with PESQ measured values (P.862 ITU-T) [6] on the transmission channel. They introduce high correlation values which justify the usability of this technique as a simple tool for perceived speech quality measurement in VoIP and GSM networks.

II. Speech Quality

Speech quality is most accurately measured by subjective opinion. The traditional speech quality measurement in telecommunications is the Mean Opinion Score (MOS). The MOS test is also called the Absolute Category Rating (ACR) test and it is described in detail in ITU Recommendation P.80.

Speech quality estimation could be performed by intrusive and nonintrusive methods. Nonintrusive methods monitor the received speech information, where particular characteristics are extracted for further processing for speech quality estimation. The drawback is the unavailability of the original speech sample for comparison with the received one, so it is possible to oversee some distortions effects of the speech signal that are impossible to be detected or measured, but have significant influence on perceived speech.

I. Kraljevski, Faculty for ICT, FON University, bul. Vojvodina bb, 1000, Skopje, Republic of Macedonia; (phone: +389 (2) 2445 593; fax: +389 (2) 2445 550; e-mail: ivan.kraljevski@fon.edu.mk).

S. Chungurski, Faculty for ICT, FON University, bul. Vojvodina bb, 1000, Skopje, Republic of Macedonia; (phone: +389 (2) 2445 555; fax: +389 (2) 2445 550; e-mail: chungurski@fon.edu.mk).

Z. Gacovski, Faculty for ICT, FON University, bul. Vojvodina bb, 1000, Skopje, Republic of Macedonia; (phone: +389 (2) 2445 592; fax: +389 (2) 2445 550; e-mail: zoran.gacovski@fon.edu.mk).

S. Arsenovski, Faculty for ICT, FON University, bul. Vojvodina bb, 1000, Skopje, Republic of Macedonia; (phone: +389 (2) 2445 590; fax: +389 (2) 2445 550; e-mail: sime.arsenovski@fon.edu.mk).

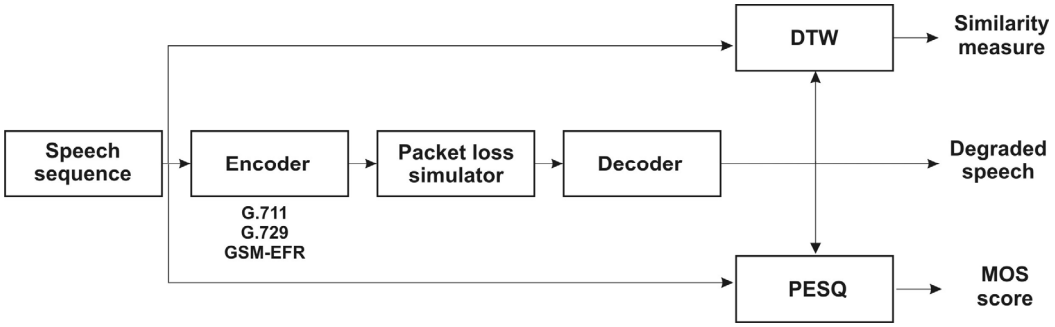

Fig. 1. Block diagram of the simulation system

Intrusive methods for quality estimation use test speech sequences that are transmitted over the communication channel. The received speech is compared with the test sequence in a similar way as the human speech perception and the quality is graded, as the listeners should do in traditional subjective tests (like MOS).

An example of one of the most popular and used algorithms for intrusive tests in packet switched and mobile networks is PESQ, defined in P.862 ITU-T. PESQ is capable of predicting the subjective quality expressed by MOS values with good correlation in a very wide range of conditions, which may include coding distortions, errors, noise, filtering, delay and variable delay. In PESQ - the original and degraded signals are represented with a perceptual model. The perceived speech quality is estimated by a cognitive model and the difference between the reference and transmitted signal [6].

This algorithm introduces some disadvantages regarding computing complexity and it cannot be used on codecs with a data rate below 4 kbps.

As already mentioned before, in this paper another intrusive procedure for speech quality measurement is described, where the DTW algorithm [7] has been used for comparison of a reference and a transmitted speech signal. The correlation between the measured values obtained by this method, and the well known and widely used PESQ model, has been observed.

III. Simulation System Description

A. Speech codecs

Three speech codecs have been used in the conducted experiments. A traditional a-Law PCM G.711 codec with a sampling frequency of 8 kHz and 8 bit per sample [3], G.729 - 8 Kbps (CS-ACELP) codec, are commonly used in VoIP applications and AMR (Adaptive Multi-Rate, ACELP) is used for speech codec standardized by ETSI for GSM applications, and it is chosen as mandatory for 3GPP networks [4]. AMR is a speech codec with 8 different narrowband modes of operation and data rates between 4.75 and 12.2 Kb/s. In simulations, the variant of 12.2 Kb/s compatible with GSM-EFR codec (3GPP TS 26.071) is used. This speech codec is mainly used for toll quality speech compression in the 2nd and 3rd generation mobile telephony applications.

The size of the packet/frame, for G.711 and AMR codecs is based on a 20 ms speech sample, for G.711 the packet size is 160 bytes, and for AMR-NB 160 samples are speech-coded to 31 bytes without VAD or PLC option used. G.729 coded one 10 ms speech sample within an 80 bit frame.

The reference speech sample has a duration of 13 seconds, and it is recorded by a male speaker in the Macedonian language.

A block diagram of the system that has been used for simulation is given in Fig. 1. The system is designed and coded in MATLAB, it allows simulation of reference sequence transmission over the communication link with packet or frame loss events. On the receiver side a comparison between the reference and received speech sequence is done. The system consists of a set of voice coders, a packet loss simulator, decoders, DTW and PESQ comparator of the degraded and the reference sequence.

B. Packet loss simulation

Packet or frame loss is a major source of speech impairment in VoIP and GSM applications. Such a loss could be caused by discarding packets in the IP networks due to congestion or by dropping packets at the gateway/terminal, due to a late arrival or faulty received and erased frames in mobile networks.

The impact of packet loss in perceived speech quality depends on several factors, including loss pattern, codec type, and packet loss size [5]. It may also depend on the location of loss within the speech, for example losing unvoiced frames/packets has a smaller impact in perceived speech quality, than losing voiced packets. Even more - as most real communication channels exhibit burstiness of packet loss, occurrence of burst of lost packets has a significant impact on speech perception. Some speech codecs like ITU G.711, G.729, and G.723.1 compression standards have implemented packet loss concealment (PLC) methods to deal with packet loss.

A 2-state Markov model [5], known as a Gilbert model, can model such channels. p is the probability that the next packet will be lost, when the previous one has arrived; q is the probability that the next packet will not be lost, given that the previous one has been lost. If $p+q=1$, the Gilbert model reduces to a Bernoulli model. The Gilbert model is a well known method for representing the packet loss

behavior 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 a jitter buffer).

The *mlp=p/(p+q)* parameter is the mean (unconditional) loss probability, and the *clp=1-q* is the conditional loss probability, conditioned on the event that the previous packet has been lost.

C. Mel Frequency Cepstral Coefficients

In order to represent the spectral features of a speech sequence that has been transmitted over a simulated communication channel, Fast Fourier Transformation (FFT) is computed for each 16 ms speech frame with advance of 1/2 frame duration. The standard Hamming window is applied to each frame.

The signal is filtered with a bank of triangular filters on Mel frequency scale (simulating the perception of the human hearing). Finally, Discrete Cosine Transformation is applied and 12 MFCC coefficients for each frame are produced.

D. DTW – Dynamic Time Warping

After signal conversion from a waveform to an array of MFCC vectors in time domain, it is necessary to perform pattern-matching algorithm to measure the distance between the transmitted and the reference speech sequence. Since the feature vectors have multiple elements, a means of calculating the local distance is required. The distance measure between the feature vector of reference signal and the feature vector of transmitted signal is given by the L2 or Euclidian distance metric.

Although the Euclidian metric is computationally more expensive than some other metrics, it does give more weight to large differences in a single feature. It can also be shown that this metric has several desirable properties when comparing cepstral features. A simple global distance score between two vector arrays is given by the sum of all local distances for the lowest distance path. The algorithm Dynamic Time Warping (DTW) always finds out the path with a minimal global distance – maximum likelihood between two utterances [7].

IV. Simulation Results

Several experiments have been conducted for each codec, G.711, G.729 and AMR, with different values for *mlp* and *clp* parameters. That produces a set of measured difference values for the speech sequence under different channel conditions, as well as MOS values for each of the experiments using the PESQ algorithm.

The speech sample has been distorted by introducing a packet loss through a simulated channel with the Gilbert model while varying values for *mlp* and *clp*. A higher value for *mlp* increases the possibility of a random packet loss, and higher values for *clp* increases the burstiness of the packet loss event. The size of the speech payload used in one packet or frame has been 160 samples, corresponding to 20 ms duration (except for G.729, where 10 ms frame is considered).

It can be seen that the observed values for similarity measure for voice codecs AMR-NB (Fig. 3) and G.729 (Fig. 4) are higher from the start of introducing a packet loss compared to G.711 (Fig. 2). The reason is that these are loosy codecs and degrade the signal even before the transmission process. In case of DTW measurements, in order to avoid the effects of signal attenuation due to used codec (which is more or less unrelated with perceived intelligibility) – the mean cepstral subtraction is applied to both - reference and the degraded signals.

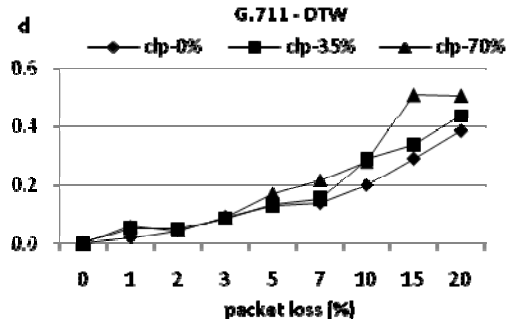

Fig. 2. Measured speech difference using G.711.

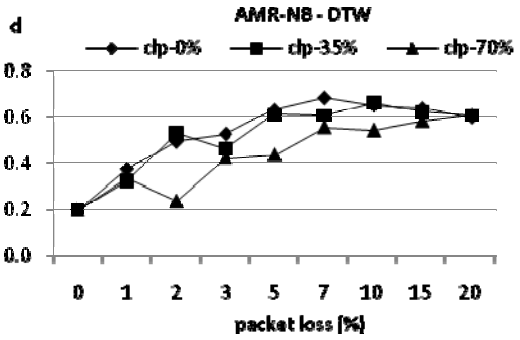

Fig. 3. Measured speech difference using AMR-NB.

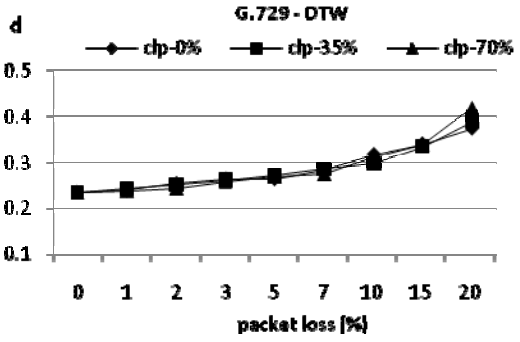

Fig. 4. Measured speech difference using G.729.

Increasing the packet loss, the difference measure (noted as „d“) for AMR-NB compared to G.711 increases in larger steps due to bigger influence of the packet (frame) loss events that induces a larger amount of degradation of the transmitted speech signal. Increasing the *clp* values causes bigger variations in the degradation growth, mainly due to the location of occurrence of bursty packet loss. On the other hand, the codec G.729 introduces smaller measured difference values with increasing the *clp*

parameter; the reason for that is the relatively small speech frame of 10 ms, sampled with 8 bits.

During simulations, besides difference estimation by the DTW algorithm, for each set of parameters - the MOS score is also predicted by the PESQ method. The regression analysis of measured MOS score (horizontal axis) and DTW difference values (vertical axis) is presented in Figs. 5, 6 and 7.

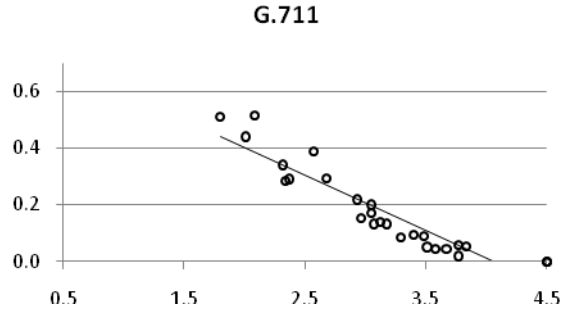

Fig. 5. G.711 – correlation coefficient = -0,93.

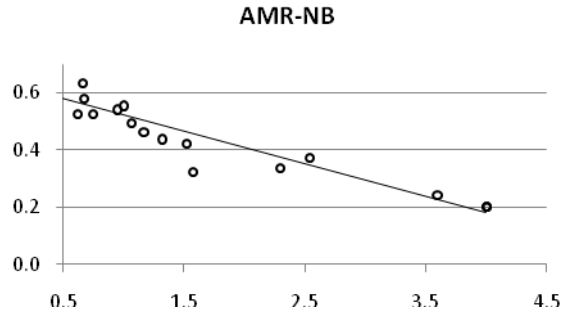

Fig. 6. AMR-NB – correlation coefficient = -0,96.

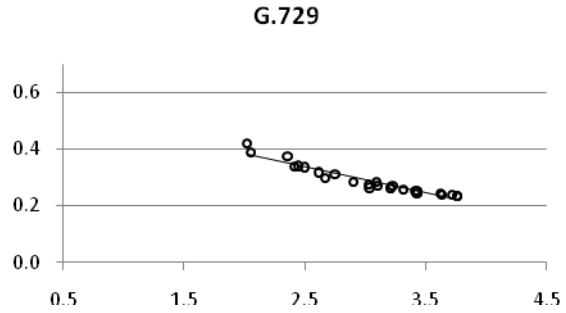

Fig. 7. G.729 – correlation coefficient = -0,96.

The results presented in Table 1 show that there is a high negative correlation between DTW measured difference values and measured MOS score for different values of *p* and *clp*. For all three codecs, achieved correlation coefficients are respectably higher in comparison to other well known objective speech quality measures: SNR, BSD, PAMS, MBSD, EMBSD and comparable with: PSQM, PSQM+, MNB2 and PESQ [8].

TABLE 1: CORRELATION COEFFICIENTS

Codcc	***G.711***	***AMR-NB***	***G.729***
Correlation coefficient	- 0,93080	- 0,96035	- 0,96039

That gives an opportunity to estimate the linear regression parameters for a particular codec and to use the DTW algorithm for perceived speech quality assessment instead of a PESQ based system.

V. CONCLUSIONS

In this paper a speech quality measurement method has been evaluated by simulating the transfer of speech over packet switched and mobile networks. The system is based on the Dynamic Time Warping algorithm for sequence matching, and several tests have been made on a reference speech sample from a single speaker, simulating packet loss effects on the perceived speech. Different values for similarity are obtained after comparing the test and received speech sequences with varying values of the possibility for packet loss errors and the possibility of introducing burstiness during packet errors.

Achieved results have been compared with values measured by the PESQ model and it has been shown that the DTW-based measurement system behaves just as a human listener in conditions with present packet or frame loss. The high degree of correlation between MOS score and measured difference values justifies the use of this technique as a simple tool for perceived speech quality measurement in VoIP and GSM networks instead of basic model of Perceptual Evaluation of Speech Quality (PESQ).

REFERENCES

[1] Ping Ji, Benyuan Liu, Don Towsley and Jim Kurose, "Modeling Frame-level Errors in GSM Wireless Channels", *IEEE Globecom, Internet Performance Symposium* 2002.

[2] I. Kraljevski, Z. Gacovski, S. Arsenovski, M. Mihajlov, "Performance of DTW Speech Recognizer on Packet Switched Network", *I6-2, VIIth ETAI Conference*, Ohrid, Macedonia, 2005.

[3] G.711 Recommendation, "Pulse Code Modulation (PCM) of Voice Frequencies", *ITU-T, Geneva*, 1988

[4] Digital Cellular Telecommunications System (Phase 2+), Universal Mobile Telecommunications System (UMTS), AMR Speech Codec, 3GPP TS 26.071 Version 6.0.0 R6.

[5] L. Sun and E. Ifeachor, "Perceived Speech Quality Prediction for Voice Over Ip-Based Networks", *Proceedings of IEEE International Conference on Communications (IEEE ICC'02),* New York, USA, April 2002, Pp.2573-2577.

[6] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs", International Telecommunication Union, Geneva, Switzerland (2001 Feb.)

[7] S. N. Wrigley. Speech Recognition by Dynamic Time Warping http://www.dcs.shef.ac.uk/~stu/com326/dtw.htm

[8] S. Mohamed, G. Rubino, M. Varela. "A method for quantitative evolution of audio quality over packet networks and its comparision with existing techniques", in Measurement of Speech and Audio Quality in Networks (MESAQIN). 2004