APPLICATION OF SPEECH TO TEXT AS CRITERION OF AUDIO QUALITY ESTIMATION IN MULTIMEDIA COMMUNICATION SYSTEMS

S. Pleshkova, K. Peeva

Department of Telecommunications,
Technical University - Sofia, Kliment Ohridski 8, Sofia
snegpl@tu-sofia.bg ; kala_peeva@yahoo.com

Abstract

In the multimedia communication systems is very important to guaranty not only the video quality, but also the quality of the sounds or speech, usually accompanying the video information. Additionally when the new speech coding methods are developed is necessary to test and estimate the quality of the decoded with the new method speech. There exists two main possible ways to estimate the speech quality in these above mentioned cases. First there are subjective methods based on subjective tests averaging of the individual estimations of each of the participants of these tests. The second possibility is to apply objective methods for speech quality determination. Most of these objective methods are the base of existing standards for audio quality measurements and estimations. It is necessary to underline, that each of two mentioned subjective and objective methods have appropriates advantages, but also the important disadvantages. Therefore, it is the goal of this article to propose and describe the new method combining the advantages of subjective methods to estimate the speech of receiving quality, changing the estimator to be not a person, but a speech to text system.

Keywords: *Audio Quality, Speech to Text, Multimedia Communication System, Speech Quality Estimation*

1. INTRODUCTION

The audio signal quality estimation is of leading importance in multimedia communication systems, in which generally there are two information sources: video and audio. There are several different group of methods [1, 2, 3] for subjective and objective quality measurements and estimation in multimedia systems of received and decoded speech signals, which are became the base of appropriates speech quality standards, known as ITU-T Recommendations [4, 5, 6]. Each of these methods or standards are very popular, but are prepared for specific cases of speech signals coding and concrete characteristics of communication channel in multimedia systems. The goal of this article is to propose the application of text to speech method in transmission part and speech to text method in receiving part of a multimedia system, as means to replace human as speaker in transmission part and human as listener in receiving part of the multimedia system. The main advantages from this proposition are to eliminate the human subjective factor in speech quality estimation process and to approach the precision of objective speech quality methods to the higher precision of subjective methods.

2. APPLICATION OF SPEECH TO TEXT AS CRITERION OF AUDIO QUALITY ESTIMATION IN AUDIO COMMUNICATION SYSTEMS

Fig. 1 presents the developed block diagram of objective quality estimation of speech signals in multimedia communications systems. The main difference of the proposed method for objective evaluation consists in the application of converting speech signals into text file.

As an initial component is used an original text (marked as block “Original text”) from printed document or computer file, which is read into a microphone device connected to the computer system and is converted into a speech signal. The input speech signal is recorded as audio file (marked as block “Audio Record 1”) in the computer system and simultaneously is converted into a digital text file (referred as block “Speech to Text Conversion 1”). The speech signal is transmitted via multimedia communication channel (block “Multimedia Communication System”) and is received from the receiver part of the multimedia system and is reproduced by loudspeaker device (presented as “Speaker” in Fig 1). At the same time the received speech signal is recorded on the computer as audio file (marked as block “Audio Record 2”).

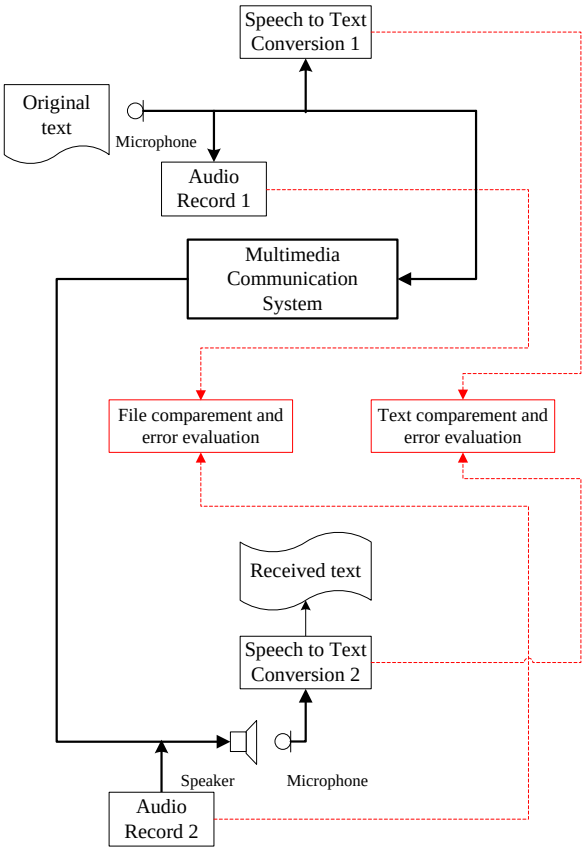

Fig.1. Main functional block scheme of simulation model for objective speech quality assessment by conversion of speech signals into text file

In front of and nearby the speaker device is placed another microphone, which transmits the speech signal for conversion into a new text file (referred as block "Speech to Text 2").

The goal of the simulation model is the comparison between the two digital text files (block "Text comparement and error evaluation"), generated in the transmission and the receiver part and the detection and determination of the number of incorrect words in the text. As a result of error evaluation is defined an exact objective quality assessment of the speech signal. In addition of the main method in the proposed simulation model is also applied a comparison between the two audio records "Record 1" and "Record 2" (marked as block "Audio comparison and error evaluation") like an extra function for more precise objective speech quality assessment.

3. SIMULATION MODEL OF OBJECTIVE AUDIO QUALITY ESTIMATION IN MULTIMEDIA COMMUNICATION SYSTEMS, IMPLEMENTED ON MATLAB PROGRAM SYSTEM

On Fig. 2 is presented a general model of simulation program using Matlab Simulink system. The schema involve the transmission and receiving parts of a multimedia communication system with the ability to choose the type of the communication channel (in this case it is shown on Fig. 2 an AWGN Communication Channel). There are presented on Fig. 2 two types of possibilities to choose the source of the speech signal: real speech signal direct from microphone (From Audio Device) or speech signal converted from a speech to text system (Data Type Conversion). In the transmission part the speech signal is saved as audio file (To Multimedia File) and in the same time is transmitted via communication channel of the Multimedia Communication System, in which is possible to define the level of noise and disturbances. In the receiving part are prepared similar operations like as in the transmission part. The received speech signal is saved back as audio file (To Multimedia File 1) and in the same time is reproduced with speaker (To audio device). With a microphone (From audio device 1), placed in front of the speaker is possible to made an inverse speech to text conversion (Data Type Conversion 2). This text is saved as a new received text document, which is used in the next step of the proposed method – the relative objective measures or estimations of speech quality in the multimedia system, described in next paragraph.

4. OBJECTIVE SPEECH QUALITY ESTIMATION BASED ON ORIGINAL AND RECEIVED TEXTS COMPARISON AND ERROR EVALUATION

The results from execution of the simulation program, presented in Fig. 2 are as following:

- received speech signal saved as speech file **rev.wav**;
- text document created after speech to text transformation in receiving part of the multimedia system and saved as text file **rev_stt.txt**;

Also in is known, that in the transmission part of the multimedia system (simulation model from Fig. 2) are available the corresponding speech information as saved files:

- original speech signal saved as speech file **orig.wav**;
- original text document **orig.txt**;
- text document created after speech to text transformation in transmission part of the multimedia system and saved as text file **stt.txt**.

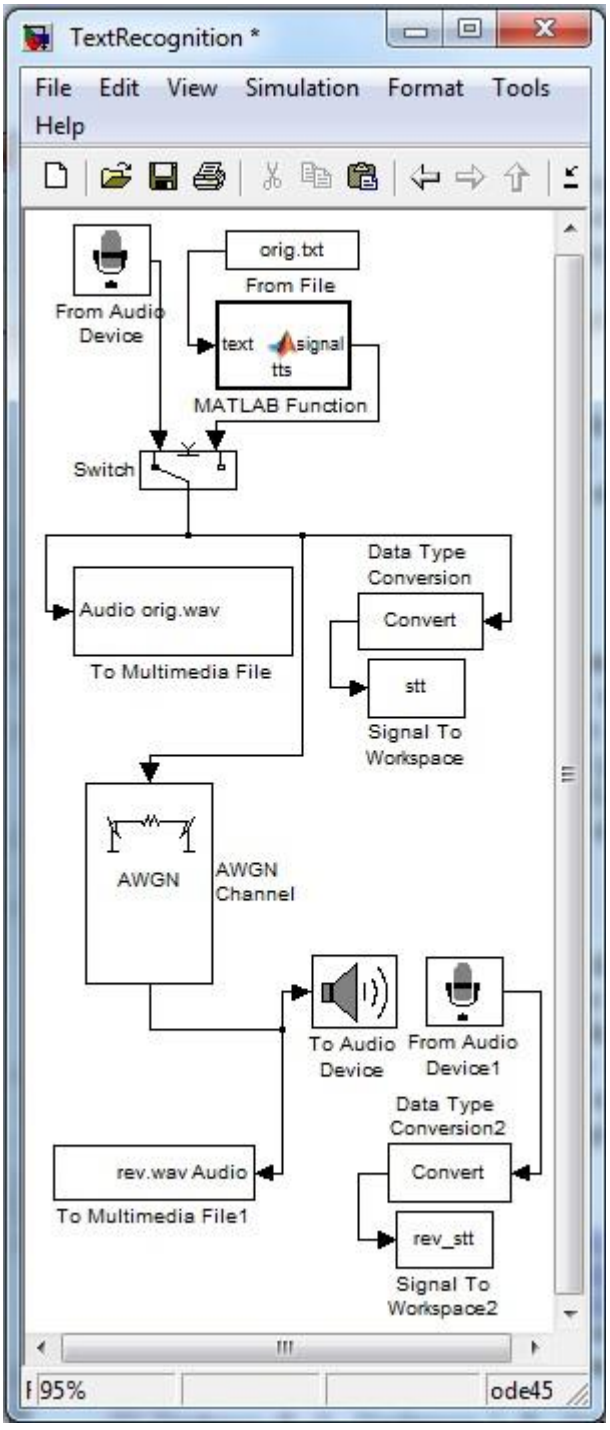

Fig. 2. Block scheme of the simulation program model for objective quality assessment of speech quality, by Speech to Text conversion

Usually each speech to text system gives some number of erroneous words when after the transformation spoken words are converted to corresponding text words. Therefore it can define the appropriate values of number of erroneous words in transmission ($NErW_{tr}$) and receiving ($NErW_{re}$) parts of the multimedia system. These definitions can be used to present the objective speech quality estimation (OSQE) as difference (DNErW) or as ratio (RNErW) between the number of erroneous words in receiving ($NErW_{re}$) part and the number of erroneous words in transmission ($NErW_{tr}$) part of the multimedia system:

$$OSQE_D = DNErW = NErW_{re} - NErW_{tr} \quad (1)$$

or

$$OSQE_R = \frac{NErW_{tr}}{NErW_{re}} \quad (2)$$

The proposed above equations (1 and 2) are two possible relative objective measures or estimations of speech quality in the multimedia system. They can be compared with some of existing subjective methods of speech quality estimation in with are used the well-arranged speech databases. If in the simulations of the proposed here method are applied the same test speech signals from the mentioned above speech databases, then the results from the existing subjective speech quality estimation methods can serve as criterion of precision of the proposed in this article method and also, which is more important to calibration of the scale of values calculated from the equations (1 and 2).

5. SIMULATIONS AND EXPERIMENTAL RESULTS OF THE PROPOSED OBJECTIVE SPEECH QUALITY ESTIMATION BASED ON ORIGINAL AND RECEIVED TEXTS COMPARISON

The schema bock of the simulation program model shown in Fig. 2 is used to carried out the experiments for objective quality assessment of speech quality, by speech to Text conversion. On Fig. 3 and Fig. 4 are shown as a simple example of one of the simulations: the original text after speech to text transformation **stt.txt** and text document created after speech to text transformation in receiving part **rev_stt.txt**

Get started ...

1. add the text2speech folder to your Matlab path
2. Test your new function:

Get started, if you use SAPI (before .NET)...

1. Make sure SAPI is installed on your computer
   a) get the Speech SDK 5.1 (86MB) for free from Microsoft:
   b) test your default computer voice
2. add the text2speech folder to your Matlab path
3. Test your new function: ('This is a test.')

I would like to thank "Desmond Lang" for his Text-To-Speech tutorial
and my wife for letting me play with the computer ;).

Fig. 3. Original text after speech to text transformation **stt.txt**

Get started ...

1. add the text2speech folder to your Matlab path
2. Test your new function:
Get started, if you use SAPI (before .NET)...
1. Make sure SAPI is installed on your computer
   a) get the Speech SDK 5.1 (86MB) for free from Microsoft:
   b) test your default computer voice
2. add the text2speech folder to your Matlab path
3. Test your new function: ('This is a test.')
I would like to thank "Desmond Lang" for his Text-To-Speech tutorial
and my wife for letting me play with the computer ;).

Fig. 4. Text document created after speech to text transformation in receiving part **rev_stt.txt**

It can be seen from Fig. 3 and Fig. 4, that there are differences of the number of erroneous words in the original text after speech to text transformation **stt.txt** and text document created after speech to text transformation in receiving part **rev_stt.txt**. This difference is used to calculate with the equations (1 and 2) the values the objective speech quality estimation (OSQE) as difference (DNErW) or as ratio (RNErW) between the number of erroneous words in receiving ($NErW_{re}$) part and the number of erroneous words in transmission ($NErW_{tr}$) part. For this example the concrete values of $NErW_{re}$ and $NErW_{tr}$ are:

$$NErW_{re} = 10 \;;\; NErW_{tr} = 5 \qquad (3)$$

Then from the equations (1,2 and 3) are calculated the values:

$$OSQE_D = DNErW = NErW_{re} - NErW_{tr} = 10 - 5 = 5 \qquad (4)$$

$$OSQE_R = \frac{NErW_{re}}{NErW_{tr}} = \frac{10}{5} = 2 \qquad (5)$$

The values calculated in equations (4 and 5) are only a demonstration of the methodology necessary to apply for the proposed method and in real simulation using the texts with larger number of words in the text the results are more realistic and precise. These results exists, but are not shown here a cause of limited size of this article.

6. CONCLUSION

In this article is proposed the application of text to speech method in transmission part and speech to text method in receiving part of a multimedia system, as means to replace human as speaker in transmission part and human as listener in receiving part of the multimedia system. The proposed method is developed as simulation model and a lot of simulations are prepared from which it is seen that the proposition of using text to speech and speech to text methods gives good results for objective speech quality estimation in multimedia system with the advantage of elimination the human subjective factor in speech quality estimation and of achievement of a near to in objective speech quality methods near to the precision of subjective methods.

Acknowledgment

This paper was supported by Technical University – Sofia inner program to support PhD research projects under Contract 132 PD0025-07: "Development of algorithms for quality estimation of audio-visual information in multimedia computer systems and networks".

References

[1] L. Thorpe and W. Yang, "Performance of current perceptual objective speech quality measures," in Proc. IEEE Speech Coding Workshop, 1999, pp. 144–146.

[2] A. Bayya and M. Vis, "Objective measures for speech quality assess- ment in wireless communications," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 1996, vol. 1, pp. 495–498.

[3] V. Turbin and N. Faucheur, "A perceptual objective measure for noise reduction systems," in Proc. Online Workshop Meas. Speech Audio Quality Netw., 2005, pp. 81–84.

[4] http://www.itu.int/rec/T-REC-J.148-200305-I/

[5] http://www.itu.int/rec/T-REC-J.143-200005-I/

[6] http://www.itu.int/rec/T-REC-J.144-200103-S/