Using Neural Networks to Evaluate the Intelligibility of the Speech in Voice over Internet Protocol

Angel Garabitov
Fakultät für deutsche Ingenieur- und Betriebswirtschaftsausbildung
TU – Sofia
Sofia, Bulgaria
e-mail: angel.garabitov@fdiba.tu-sofia.bg

Aleksandar Tsenov
Fakultät für deutsche Ingenieur- und Betriebswirtschaftsausbildung
TU – Sofia
Sofia, Bulgaria
e-mail: aleksandar.tsenov@fdiba.tu-sofia.bg

***Abstract* — An important and unresolved problem is the automatic quantification of voice quality. Many parameters affect the voice quality, but only subjective assessments are crucial. The article makes a proposal to create a solution for the prediction of subjective voice quality assessment. A neural network is used to evaluate the quality of Voice over IP (VoIP) solutions. Computational models of similarity, Musical Instrument Digital Interface (MIDI) analysis and Speech Processing Tools are used to assess intelligibility. Statistics techniques and algorithms are used for building neural networks. The end result is an analysis and an automated hypothesis about sound quality in the Mean Opinion Score (MOS) scale.**

Keywords-speach; VoIP; intelligibility; neural network.

I. INTRODUCTION

The intelligibility of speech is the defining characteristic of the speech transmission channel, since if the transmission channel does not provide complete comprehensibility of it, then its other advantages are of no importance - it is not suitable for operation. There is only one method for direct determination of the intelligibility characteristic: a statistical approach with a large number of participants (listeners and speakers). There are many attempts to create modern predictors of intelligibility. Two important issues exist with current intelligibility predictors. Many of these methods cannot reliably predict the effect of more advanced nonlinear signal processing algorithms on speech intelligibility. Typically, these measures are based on very complex auditory models or use average statistics of minutes of running speech, which makes it difficult on how to design new speech processing solutions in an optimal manner given such a measure.

To this end, we propose several new measures, which show good prediction results with the intelligibility of nonlinear processed speech. The newly proposed measures are of a low computational complexity and mathematically tractable, which make them suitable for optimization of new signal processing solutions, which aim for improving speech intelligibility.

An important and unresolved problem is the automatic quantification of voice quality. Many parameters influence the voice quality, but only subjective assessments are crucial. The article makes a proposal to create cost-effective and efficient solution for the prediction of subjective language quality assessment. A Multi Layered Perceptron (MLP) neural network is used as a "customer" to evaluate and predict the quality of VoIP solutions [1]- [3]. In Section 2, we will review the existing solutions. Then, we will propose a solution. After this, the dataset will be defined. In Sections 5 and 6, we will present the results and the conclusion.

II. RELATED WORKS

A number of measures have been proposed to predict speech intelligibility in the presence of background noise. Among these measures, the Articulation Index AI and Speech-Transmission Index STI are by far the most commonly used today for predicting speech intelligibility in noisy conditions. The AI measure was further refined to produce the speech intelligibility index STI. The STI measure is based on the idea that the intelligibility of speech depends on the proportion of spectral information that is audible to the listener and is computed by dividing the spectrum into 20 bands contributing equally to intelligibility and estimating the weighted average of the Signal-to-Noise Ratios SNRs in each band [1].

DIRAC PC software is a system used for measuring a wide range of room acoustical parameters. Based on the measurement and analysis of impulse responses, DIRAC supports a variety of measurement configurations. For accurate measurements according to the ISO 3382 standard, you can use the internally generated maximum-length sequence (MLS) or sweep signals through a loudspeaker sound source. Survey measurements are easily carried out using a small impulsive sound source, such as a blank pistol or even a balloon. Speech measurements can be carried out in compliance with the IEC 60268-16 standard [2], for male and female voices, through an artificial mouth-directional loudspeaker sound source or through direct injection into a sound system, taking into account the impact of background noise.

A lot of attempts have been made to find the direct connection between intelligibility of speech, on the one hand, characteristics of speech transmission routes and the

 ISBN: 978-1-61208-634-7

conditions for its reception and transmission, on the other, but no acceptable results were obtained. Only through the formant theory this connection could be established [3].

III. PROPOSED SOLUTION

We want to make an analogy between the melodic structure of tonal music and the understandability of speech. Using some approaches to determine the melodic structure to determine the possible speech intelligibility.

A melody is a linear succession of musical tones that the listener perceives as a single entity. In its most literal sense, a melody is a combination of pitch and rhythm. Melodies often consist of one or more musical phrases or motifs, and are usually repeated throughout a composition in various forms. Melodies may also be described by their melodic motion or the pitches or the intervals between pitches, pitch range, tension and release, continuity and coherence, cadence, and shape. All this is fully valid for the intelligibility of speech.

Before we do the analysis is necessary to invert the synthesis of speech process using formant synthesis and MIDI format. MIDI is a technical standard that describes a communications protocol and allows a wide variety of electronic musical instruments, computers and other audio devices to connect and communicate with one another [3]. MIDI carries event messages that specify notation, pitch and velocity (loudness or softness), control signals for parameters such as volume, vibrato, audio panning from left to right, cues in theatre, and clock signals that set and synchronize tempo between multiple devices.

MIDI file itself is simply an ordered series of 0s and 1s and this gave a way to analyze it using standard information theory. We will analyze MIDI sequences and perform a comparative analysis of audio signals (acoustic recordings) and symbolic representations, i.e., MIDI. By symbolic representation we mean a representation of speech as a set of notes, which is equivalent to music notation. This information is stored in MIDI files. We considered note numbers sequenced in the order of their appearance and regarding their exact onset times, durations and dynamics (so called MIDI velocities).

Formant synthesis is a special but important case of subtractive synthesis. Part of what makes the timbre of a voice or instrument consistent over a wide range of frequencies is the presence of fixed frequency peaks, called formants. These peaks stay in the same frequency range, independent of the actual pitch being produced by the voice or instrument. While there are many other factors that go into synthesizing a realistic timbre, the use of formants is one way to get reasonably accurate results.

Formants of speech sounds fill the entire frequency range from 150 to 7000 Hz. The average probability of occurrence of formants in this or that part of the range for each language is quite definite. The entire frequency range is divided into bands, so that in each of them the probability of occurrence of the formant was the same. They are defined for a number of languages. It turned out that with a sufficiently large amount of transmitted material, the probability of appearance of the formant obeys the additivity rule [4].

Our idea is to convert the speech recordings in MIDI format. To create a MIDI sequence for a speech recorded in audio format we must determine pitch, velocity and duration of each note being played and record these parameters into a sequence of MIDI events. We need to determine the order of sounds. It contains two parameters – height and the duration of the note. The structure of the sound can be analyzed mathematically by finding functions, which connect the parameters mentioned above. These functions are important for understanding the intelligibility of speech.

There are many such functions designed to analyze the melody of music [5]- [8]. We will use their ideas, but we will redesign the approach in order to analyze the sound of speech.

In brief, we want to turn the speech synthesis process to analyze speech intelligibility. There are two stages involved: sound to phonemes and phonemes to words.

The processing of this idea is too complicated; this is why we try to realize it with a neural network, because of its adaptive, deep learning capabilities and the ability to accurately separate highly correlated classes.

IV. DEFINING DATASET

A. *Mellaccent*

It calculates melodic accent salience according to Thomassen's model [9]. A musical accent can be defined as an increased prominence or noticeability associated with some note or chord.

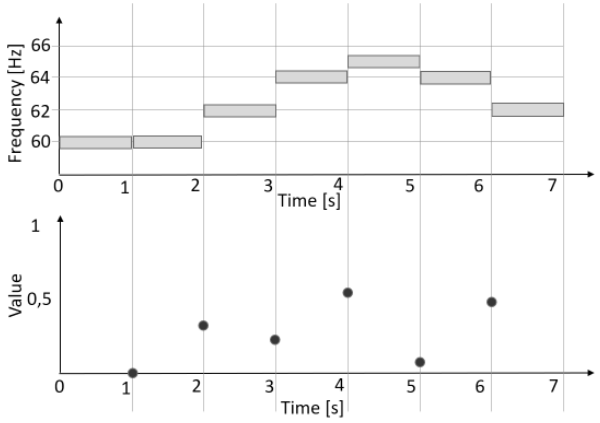

Figure 1. Melodic accent salience according to Thomassen's model

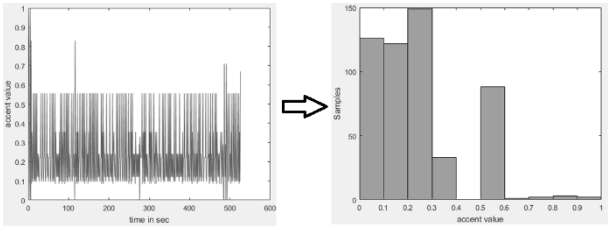

Figure 2. Mellaccent = [126, 122, 149, 33, 0, 88, 1, 2, 3, 2]

By "melodic accent" in Figure 1, music theorists mean accent arising from pitch-related changes such as changes of pitch height, pitch interval, or pitch contour. This model assigns melodic accents according to the possible melodic

contours arising in 3-pitch windows. Accent values vary between 0 (no salience) and 1 (maximum salience) in figure

The resulting value is a vector. In order to be usable for us, we represent the output vector in the form of histogram. Each value of the bins is separate input parameter of the neuronal network.

B. Narmour proximity (pr) and consonance(co)

The model draws on the Gestalt based principles [6] of proximity, similarity, and good continuation and has been found to predict listeners' melodic expectancies. The model operates by looking at implicative intervals and realized intervals. The former creates implications for the melody's continuation and the next interval carries out its implications. The model contains five principles (Registral Direction, Intervallic Difference, Registral Return, Proximity, and Closure) that are each characterized by a specific rule describing the registral direction and the distance in pitch between successive tones. The principle of Registral Return, for example, refers to cases in which the second tone of the realized interval is within two semitones of the first tone of the implicative interval in Figures 3 and 4.

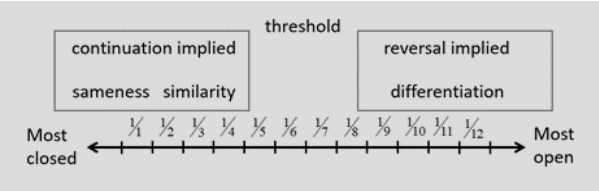

Figure 3. The Parametric Scale for Melodic Intervals

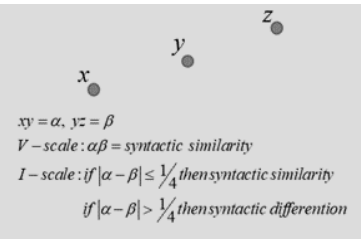

Figure 4. Cases of Intervallic Similarity and Differentiation

From all five principles, we used only Proximity and Consonance [6]. The output is also a vector and it is processed as above.

V. EXPERIMENTAL DATASET

We are using two criteria as identifier of the speech probe. They are in the form of vectors. Each value of the vector will be an input neuron of the neuronal network.

To investigate the impact of network defects is done with simulation of the loss of packets, the delay and jitter. Changing the parameters directly affects the quality and intelligibility of speech. The testing tool can be seen in Figure 5. In the performed experiment, we have collected 708 samples. Here we will perform some statistical processing of all these samples. The total count of input values is 23, and after removing a column containing only zeros the count is reduced to 22.

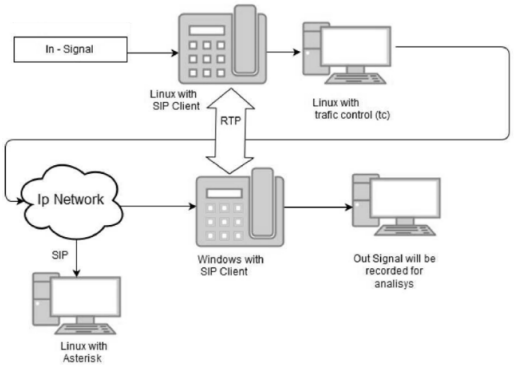

Figure 5. The used detwork simulation

A. Task description

The data set contains the parameters for creating the predictive model. It comprises a data matrix in which columns represent variables and rows represent samples.

The independent variable is 8 but in form of vectors. The total count of all input parameters will be 22 independent variables. The targets will be 8 dependent variables, representing the voice quality in MOS scale. Figure 6 shows the mean value of each input parameters depending of the target. Figure 7 shows the correlation between the parameters.

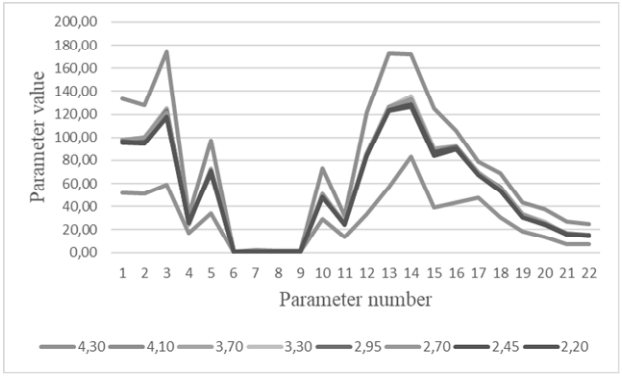

Figure 6. Data mean values

The number of layers in the neural network is two. The following table depicts the size of each layer and its corresponding activation function. The architecture of this neural network can be written as 22:22:8. Activation function in all layers is Logistic. The mathematical expression represented by the neural network can be written. It takes the inputs Mellaccent (1 – 9), NamorCo (1 – 5) and NamorPr (1 – 8), to produce the outputs in MOS (2,20 – 4,40). The NN errors are displayed in the next table.

TABLE I. ERRORS TABLE

	Training	Selection	Testing
Sum squared error	25.5304	12.253	10.9318
Mean squared error	0.0599305	0.0869006	0.0775302
Root mean squared error	0.244807	0.294789	0.278442
Normalized squared error	0.796492	0.788809	0.778646
Cross-entropy error	2.37538	2.3531	2.31614
Minkowski error	38.4993	17.0072	15.5343

 ISBN: 978-1-61208-634-7

	Melaccent	Melaccent	Melaccent	Melaccent	Melaccent	Melaccent	Melaccent	Melaccent	Melaccent	NamorCo	NamorCo	NamorCo	NamorCo	NamorCo	NamorPr	NamorPr	NamorPr	NamorPr	NamorPr	NamorPr	NamorPr	NamorPr
Melaccen	1,000																					
Melaccen	0,992	1,000																				
Melaccen	0,998	0,992	1,000																			
Melaccen	0,981	0,957	0,972	1,000																		
Melaccen	0,997	0,998	0,995	0,966	1,000																	
Melaccen	0,570	0,498	0,543	0,689	0,514	1,000																
Melaccen	0,655	0,577	0,667	0,675	0,600	0,651	1,000															
Melaccen	0,538	0,470	0,565	0,539	0,487	0,562	0,914	1,000														
Melaccen	0,719	0,651	0,734	0,726	0,670	0,653	0,988	0,933	1,000													
NamorCo	0,984	0,965	0,989	0,963	0,972	0,594	0,768	0,668	0,825	1,000												
NamorCo	0,992	0,995	0,987	0,968	0,996	0,550	0,567	0,465	0,644	0,960	1,000											
NamorCo	0,997	0,993	0,995	0,981	0,996	0,549	0,617	0,494	0,682	0,973	0,992	1,000										
NamorCo	0,998	0,997	0,995	0,976	0,999	0,545	0,608	0,490	0,676	0,973	0,996	0,999	1,000									
NamorCo	0,994	0,994	0,997	0,954	0,995	0,510	0,649	0,558	0,721	0,985	0,987	0,988	0,992	1,000								
NamorPr	0,997	0,992	0,998	0,971	0,996	0,524	0,651	0,542	0,715	0,984	0,986	0,996	0,996	0,995	1,000							
NamorPr	0,943	0,967	0,937	0,908	0,963	0,383	0,371	0,260	0,452	0,875	0,967	0,958	0,961	0,938	0,946	1,000						
NamorPr	0,973	0,990	0,971	0,939	0,987	0,439	0,479	0,357	0,554	0,927	0,985	0,984	0,985	0,971	0,975	0,989	1,000					
NamorPr	0,984	0,994	0,980	0,952	0,993	0,497	0,519	0,415	0,601	0,944	0,997	0,988	0,991	0,982	0,980	0,979	0,994	1,000				
NamorPr	0,990	0,991	0,989	0,961	0,989	0,578	0,645	0,543	0,716	0,979	0,989	0,983	0,989	0,992	0,984	0,929	0,966	0,982	1,000			
NamorPr	0,990	0,979	0,993	0,960	0,983	0,572	0,720	0,634	0,786	0,996	0,975	0,979	0,982	0,994	0,989	0,901	0,943	0,962	0,989	1,000		
NamorPr	0,980	0,952	0,979	0,977	0,961	0,657	0,788	0,659	0,835	0,994	0,951	0,969	0,966	0,969	0,974	0,859	0,912	0,931	0,969	0,985	1,000	
NamorPr	0,980	0,958	0,981	0,967	0,964	0,643	0,780	0,671	0,831	0,996	0,954	0,967	0,968	0,976	0,976	0,862	0,916	0,935	0,978	0,991	0,996	1,000

Figure 7. Correlation between the parameters

For classification problems, the information is propagated in a feed-forward fashion through the scaling layer, the perceptron layers and the probabilistic layer.

The mathematical expression represented by the neural network can be written. It takes the inputs Mellaccent (1 – 9), NamorCo (1 – 5) and NamorPr (1 – 8), to produce the outputs in MOS. For classification problems, the information is propagated in a feed-forward fashion through the scaling layer, the perceptron layers and the probabilistic layer.

Table II contains the elements of the confusion matrix. 567 – Train samples; 141 (25%) – Test Samples. The correctly classified samples are 113, and the misclassified instances are 28. The obtained classification accuracy is calculated as 80%. Table II shows the classification results of the neuronal network according to the MOS scale:

TABLE II. PREDICTION MATRIX

		Predicted							
		4,3	***4,1***	***3,7***	***3,3***	***2,95***	***2,7***	***2,45***	***2,2***
Actual	***4,3***	18	0	0	0	0	0	0	0
	4,1	0	16	0	0	0	0	0	0
	3,7	0	0	17	0	1	0	0	1
	3,3	0	0	1	16	0	1	0	0
	2,95	0	0	0	1	11	0	1	3
	2,7	0	0	0	1	1	17	0	1
	2,45	1	2	1	4	2	1	11	0
	2,2	0	0	0	1	1	0	3	7

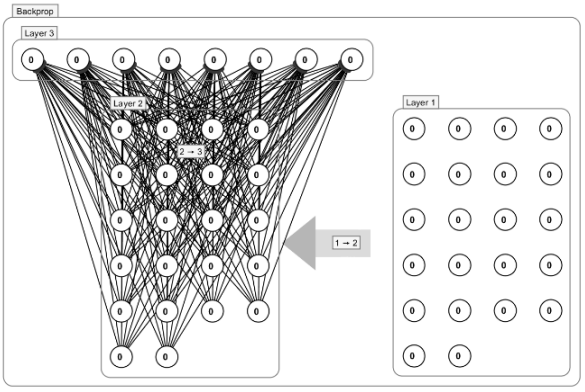

Figure 8. Experimental MLP topology

VI. CONCLUSION

MLP classification neural network with given criteria can be used to estimate voice quality according to MOS scale. The network is used as a "customer" to evaluate and predict the quality of VoIP solutions with about 80% successful predictions. An advantage of the method is the possibility of easy adjustment when changing the input parameters - only by means of re-training the MLP.

ACKNOWLEDGMENT

Research, the results of which are presented in this publication are funded by internal competition, TU-Sofia 2017, CONTRACT № 172PD0010-07 for a research project to help PhD

REFERENCES

[1] J. Ma, Y. Hu, and P.C. Loizou "Objective measures for predicting speech intelligibility in noisy," The Journal of the Acoustical Society of America, vol. 125, no. 5, pp. 3387–3405, 2009

[2] Xscala Sound & Vibration http://xscala.com/list/products/dirac-room-analysis-software/, retrieved: April, 2018

[3] M. A. Sapojnikov, Electoakustic. Moskva, Sviaz, 1978

[4] Ph. Burk, L. Polansky, D. Repetto, M. Roberts, and D. Rockmore, "Music and Computers", http://sites.music.columbia.edu/cmc/MusicAndComputers/chapter4/04_04.php, retrieved: April, 2018

[5] A. Swift, "A brief Introduction to MIDI," Imperial College of Science Technology and Medicine http://www.doc.ic.ac.uk/~nd/surprise_97/journal/vol1/aps2/, retrieved: April, 2018

[6] P. Toiviainen and T. Eerola, "miditoolbox (Midi toolbox) GitHub", https://github.com/miditoolbox/, retrieved: April, 2018

[7] M. Brookes, "VOICEBOX: Speech Processing Toolbox for MATLAB", http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, retrieved: April, 2018

[8] E. Pampalk, "Computational Models of Music Similarity and their Applications in Music Information Retrieval," TU Wien, Wien, 2016

[9] J. Thomassen, "Melodic accent: Experiments and a tentative model," Journal of the Acoustical Society of America, vol. 71, no. 6, pp. 1598-1605, 1982

 ISBN: 978-1-61208-634-7