Evaluating Psychoacoustic Parameters and Quality of Transmitted Speech over Wireless Networks

A. Olatubosun, Patrick O. Olabisi

Abstract—Psychoacoustic parameter of sound known as loudness is a major quality factor for assessing the perceptual quality of service of speech signals transmitted through telecommunication networks. The Zwicker and Fastl loudness model is a preferred loudness model and in this work has been programmed to obtain both loudness and loudness level of speeches transmitted over wireless. Here, the best maximum instantaneous loudness of the transmitted speeches is 42.55% of that of the original speech. While the best maximum instantaneous loudness level of the transmitted speeches is 87.06% of that of the original speech. These showed an intuitive and innovative representation of the degradation suffered by the transmitted speeches with respect to the original speech.

Index Terms—Loudness Model; Psychoacoustic Parameter; Quality of Service; Transmitted Speech.

I. INTRODUCTION

Psychoacoustic as the study of the perception of sound by the human auditory system is a field in which acoustic stimuli of sounds generally reaching the human ear is statistically related with the hearing sensations produced in the auditory system. It relates the physical quantities of sound with the perception of sound. Perception of acoustic stimuli and the quality of perception by the human auditory system are usually assessed by evaluating the psychoacoustic properties or parameters of speech signals, which include loudness, timbre, pitch, fluctuation of strength, sharpness, and roughness [1].

End-to-end transmission of speech over telecommunication networks happens between the mouth of a speaker (the caller) and the ear of the corresponding listener (the called). This occurs notwithstanding the type and dimensioning of the telecommunication network facility installed in-between the two parties. In measuring the quality of such transmission in what is known as end-to-end (E2E) speech quality of service (SQoS) and all the associated issues, consideration of what the psychoacoustic parameters are, becomes very important.

Speech production at one end and speech perception on the other end, between them sits the network with its facilities, and conditions, which include several atmospheric, transmission and systemic distortion and attenuation conditions. Efforts at assessing the quality of transmitted speech may as well derive meaning and measures from evaluating the psychoacoustic parameters of speech, which is what this work is all about.

Of all psychoacoustic parameters of sound, [2] claimed that loudness is the primary perceptual correlate of the level of quality of a sound. This is because its value and variation depend upon such physical factors of sound as the frequency, bandwidth, duration, spectral complexity of sound, the presence of other sound, and so on.

II. PSYCHOACOUSTIC PARAMETER OF LOUDNESS

Loudness with unit as Sone is referred to by [3] as the subjective perception of how strong or powerful a given sound is. It is a non-linear response of the human ear to sounds of different levels of intensity [4]. Loudness was defined by [2] as the perceptual strength of a sound that ranges from soft (very quiet) to very loud, but indicated that Scharf (1978) defined it as the attribute of sound that changes when the subjective intensity of sound is varied. The loudness of a sound is rather seen as a psychological attribute of sound than being a physical attribute [5].

In estimating speech quality, the integral quality is obtained from the combination of the perceptual transform and a simulation of the cognitive processes in the human auditory cortex. [6] indicated that this perceptual transformation follows the psychoacoustic model for loudness calculation developed as loudness models, for example that by Zwicker and Fastl and of course by a few others.

The quantity of acoustic sensation of speech in the auditory system is described by both loudness and loudness level. While loudness describes the absolute sensations of sound strength perceived by a person, loudness level describes the relative sensations, taking cognizance of the hearing condition of the subject. They are both statistically characterized by their measures of central tendency and dispersion [3].

Loudness level of a sound, measured in phon, defines the loudness sensation of a sound and it is equivalent to the sound pressure level (SPL) in dB of a 1 kHz tone as loud as the sound. It can be deduced from the equal-loudness contour graph for different tones [7].

Several computation models have over time been devised for estimating loudness of diverse categories of sound, namely: stationary sounds, non-stationary or time-varying sounds and impulsive sounds. Although, efforts at determining loudness of sounds dated back to mid-nineteenth century with the works of men like Fechner in what is known as the Fechner’s logarithmic law built upon both philosophical and mathematical intuitions derived from

Published on February 3, 2019.
A. Olatubosun is with the Electrical & Electronic Engineering Department of University of Ibadan, Ibadan, Nigeria. (e-mail: a.olatubosun@mail.ui.edu.ng).
P. O. Olabisi is with the Electrical, Electronic and Computer Engineering Department of Bells University of Technology, Ota, Nigeria (e-mail: poolabisi@bellsuniversity.edu.ng).

DOI: http://dx.doi.org/10.24018/ejers.2019.4.2.1088
Weber’s law of intensity discrimination [2].

The Steven’s power law was one of the earlier loudness models, and it gave a relationship between the magnitude of sound stimulus and its perceived intensity [8]. More recent loudness models include the Zwicker and Fastl model (1999) and the Moore and Glasberg model (2002) for both stationary and non-stationary types of sound, and the Bouillet model (2006) for impulsive sounds.

The loudness models were developed around the human auditory system and are based on spectra extraction of speech signal [9]. They provided computational procedures for estimating loudness and loudness level of sound perceived by individuals with normal hearing condition. This is done by analysing physical characteristics of sound under specified listening conditions.

The models were internationally standardized under the ISO 532 series. The ISO 532:1975 was the internationally standardized version of the Steven’s power law loudness model. The ISO 532-A is for calculating loudness and loudness level of stationary sounds while ISO 532-B is for calculating loudness and loudness level of non-stationary sounds [11], [12]. The German version of the ISO standards are the DIN 45631:1999 and the DIN 45631/A1:2010 respectively.

The Moore and Glasberg model originated in 1996 as Moore, Glasberg and Baer model based on excitation loudness pattern and was developed to cover time-varying sounds in 2002 by only Moore and Glasberg [13], [14]. It was formally standardized and defined in ANSI S3.4. [12].

In comparing the two major loudness models – the Zwicker and Fastl model and the Moore and Glasberg model, [9] noted that results have shown that Zwicker’s parameters are more suited to reflect physiological processes and perceptual results.

III. QUANTITATIVE MEASURE OF THE LOUDNESS OF SPEECH SIGNALS

Efforts were made by [15] at providing quantitative estimate of the loudness of speech based on the characteristics and estimation of the source of excitation, that is, glottal excitation of speech signal. This is determined by the size and shape of the vocal tract and position of the articulators, leading to variations in speech loudness.

The loudness of various environmental noise (as an example of non-stationary sound) was estimated by [16] by approximating to it the arithmetic average of the sound pressure level obtained in octave bands, and that the arithmetic average of sound pressure level in octave bands for 63 Hz to 4 kHz, \( L_{an/1} \) (63 – 4 kHz), correlates well with loudness level \( LL(Z) \) defined in ISO 532B and given by:

\[
LL(Z) = 40 + 10 \log_2 N \quad \text{(in phon)}
\]

where \( N \) is the total loudness (in sone).

Steps involved in Zwicker and Fastl model given by [7], [9] are: Microphone pick-up of sound, amplification of the signal, a two-third octave filter bank, one-way rectification, and 2 ms time-constant low-pass filtering.

Shown in Fig. 1 is a summary of stages involved in calculating the loudness of time-varying sounds using the two major loudness models, the Zwicker and Fastl model (2007) and the Moore and Glasberg model (2002). Linear time-invariant filters are used to model acoustic performance between outer and middle ear before the introduction of critical band filterbank to break the signal power into its spectral components [10].

Statistical indicators used to obtain the global loudness for the sound under test include percentile loudness (\( N_x \)) and 50% (\( N_{50} \)), which indicate loudness value exceeded during x percent of time, for Zwicker and Fastl model. For Moore and Glasberg model, we have the maximum of short-term loudness (STLmax) calculated in order to estimate overall loudness of time-varying sounds perceived at a particular time instance, and the maximum of long-term loudness (LTLmax) calculated in order to estimate the overall loudness of steady time sounds or sounds that vary in time very slowly presuming that overall loudness is saved in memory for time spanning several seconds before a new stimulus is felt [11].

Calculating the loudness particularly for Zwicker and Fastl model therefore, entails that the following are carried out:

1. To obtain the filter transfer representing outer to middle ear functions;
2. To make use of auditory filter banks in representing inner ear functions from which excitation, \( E \), are calculated. Critical bandwidth as well as the frequency selectivity of the auditory system play critical roles in loudness.
3. Transform the excitation patterns into specific loudness, \( N' \), by a power law relationship. Specific loudness is defined by [7], [8] as:

\[
N' = \alpha \left( \frac{E_{THQ}}{E_0} \right)^\beta \left[ 0.5 + 0.5 \frac{E}{E_{THQ}} \right]^{\frac{\beta}{5}} - 1 \quad \text{sone} \text{ Bark}^{-1}
\]

where: \( \alpha \) and \( \beta \) are constant factors given as \( \alpha = 0.08 \), \( \beta = 0.23 \), \( E_{THQ} \) is the excitation at threshold in quiet, \( E_0 \) is excitation that corresponds to the reference intensity, \( E \) is the excitation at the specific frequency, while \( I_0 = 10^{-12} \text{ W/m}^2 \). The index G appended to sone indicates that this specific loudness is obtained as a function of the critical-band levels or rates.

Total loudness is obtained as the integration of the specific loudness over frequencies in ERB or Bark scale. For the Zwicker and Fastl model adopted for this work, the total loudness is integrated over the Bark scale. It is obtained as summation of all neural activities caused by the sound/speech across the Basilar Membrane in the inner ear. It is therefore defined by:

\[
N = \int_0^{248 \text{Barks}} N' dz
\]

where \( N' \) is the specific loudness given above.

DOI: http://dx.doi.org/10.24018/ejers.2019.4.2.1088
IV. QUALITY OF TRANSMITTED SPEECH ON WIRELESS NETWORK

In attempt to evaluate end-to-end quality of service transmitted through wireless mobile networks, the key psychoacoustic parameter of speech mostly affected by distortions and attenuations caused by network equipment and transmission conditions is loudness.

Shown in Fig. 2 is the plot of waveforms (in dB) of an original speech sample, and speeches transmitted through three wireless networks (two intra-networks - EE network and GG network, and one inter-network - EM network). The speech plot was carried out using Audacity 2.1.2 software for recording and editing of sound [17].

As could be seen in Fig. 2, the waveform of the original speech has more robust amplitude (in dB) than the remaining three waveforms which are degraded speeches obtained from the three transmissions respectively. As a result of degradation of the speech during transmission, the amplitudes from the networks have attenuated very appreciably as shown.

Plot of the frequency analysis of these speeches was also carried out as shown in Fig. 3 to 6. Fig. 3 is the plot of the frequency analysis of the original speech, while Fig. 4, 5, and 6 are the plots for the transmitted speeches obtained from the three mobile wireless networks.

V. PROGRAMMING OF LOUDNESS ESTIMATION

Due to the complexity of the computations involved in the stages of obtaining the loudness of time-varying sounds, the developed loudness models have almost in every case come with computer software programmes to help in reducing computation stress, time and resources. [8] implemented the algorithms for the Zwicker and Fastl model and the Moore and Glasberg model using Matlab for a couple of pure tones, bandpass noise and industrial noise stimuli. There are many professional sound quality software which provide programme for the calculation of loudness values of different types of sound – speech, music, noise, and so on. This work made use of the programme developed by Genesis [18] written in Matlab.

The results of the computation of the loudness of sample speeches used are stated in Table I.

Fig. 1: Stages of Loudness calculation for time-varying sounds (Source: Ref. 11).

Fig. 2: Waveform (in dB) of the original speech and the transmitted speeches obtained from Networks: EE, GG and EM respectively.

Fig. 3: Frequency analysis of the original speech.

Fig. 4: Frequency analysis of EE network speech.

Fig. 5: Frequency analysis of GG network speech.

Fig. 6: Frequency analysis of EM network speech.
Fig. 3. Frequency analysis plot of the original speech (OrgM1Sp1.wav) before transmission.

Fig. 4. Frequency analysis plot of speech transmitted through EE wireless network (EEM1Sp1.wav).

Fig. 5. Frequency analysis plot for speech transmitted through GG wireless network (GGM1Sp1.wav).
VI. RESULTS ANALYSIS

Visual display of the plot of the waveforms of the original (reference) speech sample (OrgM1Sp1.wav) and the network degraded speeches (EEM1Sp1.wav, GGM1Sp1.wav and EMM1Sp1.wav) in Fig. 2 showed that the transmitted speeches suffered serious attenuation compared to the original speech. In Fig. 3, the frequency analysis plot of the original speech shows how robust the original sample speech is compared to those of the transmitted speeches shown in Fig. 4, 5 and 6.

As could be seen on Table I, all the instantaneous loudness and loudness level values of the transmitted speeches are much less than those of the original speech. In Table II particularly, the maximum instantaneous loudness of the transmitted speeches are 42.55, 37.08 and 35.64% respectively of the maximum instantaneous loudness of the original sample speech. Also, the maximum instantaneous loudness level of the transmitted speeches are 87.06, 84.98 and 84.37% respectively of the maximum instantaneous loudness level of the original sample speech.

VII. CONCLUSION

The results analyses are resounding objective proofs of the degrading effects that transmission network conditions have on the quality of the original speech. This was proven by analysis of loudness psychoacoustic feature of speech signal. This work is in line with efforts at extending the frontiers of discovery in the evaluation of causes and consequences of quality effects of speech degradation when transmitted over various telecommunication networks.

REFERENCES


