Applied Acoustics 113 (2016) 58–63

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Time-compressed speech intelligibility in different reverberant conditions

Jędrzej Kociński *, Dawid Niemiec

Institute of Acoustics, Faculty of Physics, Adam Mickiewicz University in Poznań, Umultowska 85, 61-614 Poznań, Poland

ARTICLE INFO

Article history:
Received 5 April 2016
Received in revised form 13 June 2016
Accepted 15 June 2016
Available online 21 June 2016

Keywords:
Speech intelligibility
Time Compression Threshold
Time-compressed speech

ABSTRACT

The aim of this study was to determine how different reverberant conditions affect subjective perception of fastened speech. The objective parameters used for a description of room acoustics quality were Reverberation Time (RT) and Speech Transmission Index (STI). In relation to these parameters the Time-Compression Threshold (TCT) had been measured for normal hearing listeners younger than 30. The measuring test was based on artificially generated room impulse responses, MATLAB implementation of a phase vocoder, the Polish Matrix Test (PMT) combined with the 1-up/1-down adaptive procedure. As expected, the value of TCT decreases with an increase of RT and is highly correlated with STI. Therefore, the TCT parameter can be treated as an another sufficient estimator of acoustical quality of an enclosure in the context of speech intelligibility.

Introduction

In general, speech intelligibility is a measure of the extent to which an entire system (i.e. electric circuit, amplification, enclosure, etc.) affects the amount of successfully received information provided in speech (phonemes, numbers, words, sentences, etc.). It is defined as the percentage of speech units that are understood correctly during a performance of speech intelligibility test [27].

The problem of speech intelligibility can be analyzed from two different approaches. The first one treats background noise as an additive distortion, described by the parameter called signal-to-noise ratio (SNR). The SNR value at which the psychometric function that describes speech intelligibility is equal to 0.5 (50% of intelligibility) is called SRT – the Speech Reception Threshold [33]. It is worth noticing that the SRT values depend not only on a system but also on a subject and material used in the test. Moreover, intelligibility of speech in the presence of background noise also depends on spectral structure of both signals, presentation direction [35] and the knowledge of the context [37].

The second general type of interferences with significant influence on speech intelligibility are convolutive distortions. In each room the output signal consists of a direct wave merged with a number of reflections from various walls and obstacles. These time-shifted and somehow filtered reflections more or less affect both time and spectral structure of recorded/perceived signal. As a normal speech has a certain dynamics of about 25 dB, the presentation of utterances in highly reflective enclosures can result in masking quiet phonemes by the sustained previous loud ones and therefore in significant decrease of successful information understanding [5,22]. It is worth to note that masking occurrence depends not only on the intensity level, but also on a spectral structure, a type of presentation and a distribution of stimulus in time [27].

The problem of speech intelligibility in the context of room acoustics has been analyzed for many decades. Generally, there are two main approaches to speech intelligibility measurements in a room, namely objective and subjective. The former is based on the analysis of distortions provided by the enclosure and model-based prediction of intelligibility. The latter one is simply obtained by conducting special speech intelligibility tests which are presented to the listeners.

Many different objective predictors of speech intelligibility have been introduced. The idea that speech intelligibility can be calculated on the basis of the influence of different frequency bands has been proposed in the late twenties of the last century by Fletcher [7], and afterward modeled by French and Steinberg [8]. This idea was simplified by Kryter [19,20] who introduced Articulation Index (AI). To obtain AI value the spectrum of the speech signal has to be divided into 20 bands in which the consecutive SNRs are analyzed. Knowing the value of SNRs and weighting factor for each band, defined by the American National Standards

* Corresponding author.
E-mail address: jedrzej.kocinski@amu.edu.pl (J. Kociński).

http://dx.doi.org/10.1016/j.apacoust.2016.06.009
0003-682X/

Institute [2], the AI can be calculated according to the equation given by authors of this concept.

AI's successor in a field of a speech intelligibility prediction was the Speech Intelligibility Index (SII). The SII weighting values slightly differ from those proposed for AI. Moreover, a correction factor dependent on a vocal effort and its influence on a spectral structure of the utterance was introduced.

In 1970s and 80s Houtgast and Steeneken [10–14] analyzed the problem of speech intelligibility estimation and finally introduced the parameter called Speech Transmission Index (STI), which significantly differs from the AI in the method of SNR estimation. The STI calculation is based on the assumption that intelligibility of the output signal is strongly dependent on a depth of low frequency amplitude modulations (AM), namely AM rates occurring in a speech signal [48]. Generally, STI method is based on analysis of a specially designed AM test signal with long-term spectrum identical to speech in predetermined number of frequency bands after passing through the target system (or after convolution with its impulse response). In those bands a so-called modulation transfer function (MTF) is determined [14]. Then, similarly to AI, a system of weights of bands reflecting their importance in the context of overall speech intelligibility is required. Afterward, a single number between 0 and 1 is obtained that defines speech intelligibility (0–0.3 bad, 0.3–0.45 poor, 0.45–0.6 fair, 0.6–0.75 good and 0.75–1 excellent intelligibility). A modification of this procedure was introduced, called RaSTI (Rapid STI or Room Acoustics STI). The general idea was the same but the number of considered bands was limited to 2 (500 Hz and 2000 Hz). Since RaSTI was declared obsolete, it was replaced by STIPA (Speech Transmission Index for - Public Address Systems). In STIPA, each octave band is modulated simultaneously with two modulation frequencies. The modulation frequencies are spread among the octave band. It gives a reliable STI values based on a sparsely sampled Modulation Transfer Function matrix.

A large number of studies, e.g. Houtgast and Steeneken [15], Anderson and Kalb [1], Barnett [3], van Wijngaarden and Drullman [40], Houtgast and Steeneken [11], van Wijngaarden and Steeneken [41], Mapp [24], validated results obtained using the STI calculation to experimental values of speech intelligibility in psychoacoustic research. The comparison was made between following groups of material: CVC words (consonant–vowel–conso nant), phonetically balanced word list, logatomes, and complete sentences. It must be emphasized that all of those studies compared STI to subjective speech intelligibility at a normal speech rate presentation. Changes in the speed rate of playback of speech signal obviously must cause also changes in AM rates. This leads to the problem of accelerated speech intelligibility.

Time-compressed speech finds its vast amounts of applications in modern technology. For example, in automated call centers with systems based on speech recognition, voice commands are often time-normalized to standard values of wpm/spm (words/syllables per minute) in order to facilitate the comparison of the input signal with base patterns [23]. Saving time required to transfer information is also remarkable advantage in early warning systems, when during emergency situations such as an evacuation of a building or a sudden burst of bad weather, every second is valuable and can help minimizing damage to human or material resources. Speech acceleration is also used in teaching. Sticht [39] showed that twice accelerated piece of information presented twice was much better remembered than the same material presented once at a normal speed rate. It is worth noting that this effect occurs only during the presentation of the material in the listener's native language. If the listener receives information in a foreign language, slower speed rates improve its understanding [6]. It is also well known fact that some languages are "faster" than other, e.g. Spanish native speakers talk with a higher speech rates than English ones. According to National Center for Voice and Speech for English, this value reaches about 150 wpm (words per minute). The upper limit of the acceleration threshold of natural speech derives more from physiological than neurological nature [4]. The aforementioned physiological barrier is not an issue for artificial speech. Modern speech synthesizers can reach speed of 550 wpm, which is achieved mainly by reducing the duration of phonemes and pauses between words. In practice it can be found that audiobooks are recommended to be 150–160 words per minute (wpm) [46], while slide presentations tend to be closer to 100–125 wpm for a comfortable pace [47].

In the area of artificial speech acceleration many different algorithms can be used. First attempts to an automatic time-compression were carried out using an intuitive method that is faster playback of the recorded material [18]. The next step in the search for the optimal time-compression technique was shortening of between words gaps. In this approach, the sentence retains its natural pitch, but due to a lack of breaks in breathing, after some time, listeners reported discomfort and fatigue [28]. There are several techniques to eliminate a silence from speech – one of them was developed by Maxemchuk [25]. He used a 62.5 ms signal windowing and then compared the energy of obtained segments. At the point at which the energy of several consecutive segments felt below the threshold, the algorithm marked this region as a silence, which was cut out from the signal. The main problem was the dynamics of the speech signal, which is ca. 25 dB. In non-reverberant conditions with high SNRs, the algorithm works satisfactory. However, in other conditions or for incorrect initial threshold value, it cuts out also the most quiet phonemes.

Another family of algorithms used for speech acceleration is a group of techniques based on sampling. Two methods are worth mentioning here. The first one is called Fairbanks sampling. In this technique, the signal is sampled at regular intervals and depending on the rate of acceleration, the appropriate number of samples is omitted. Portnoff [36] states that the sampling time should be at least equal to one period of a fundamental frequency of voice, but also shorter than a length of a single phoneme. Fairbanks sampling method is very simple and therefore not computationally expensive. Unfortunately, it effects in clearly noticeable artifacts, noise and generally much worse output signal quality compared to the input. Quite similar method is based on the dichotic presentation of the sampled signal. Dichotic means that two different signals are presented to each ear. In this case, there are two signals sampled using Fairbanks method, delayed from each other by half the sampling time.

The next one of the most common and simple algorithm is based on the overlap-add method (OLA). The principle of this method is based on splitting the signal into overlapping segments, then in proportion to a desired acceleration/deceleration, the multiplication or reduction of windowed fragments occurs. This is a very simple method, however, reflecting negatively on the quality of the output signal. Due to the absence of any mechanism for tracking of behavior of the input signal while adding adjacent portions constituting the output signal, there are many artifacts resulting from border mismatch. This problem is somewhat solved in the SOLA method (synchronous overlap-add algorithm), by introducing the cross-correlation function, which is to make the best match of adjacent segments, unfortunately resulting in a much greater computational complexity.

For the synchronization of adjacent windows peak values, a different approach can be implemented. For a speech signal, the natural marker of the periodicity can be used, namely the fundamental frequency. This fact is used in the PSOLA algorithm (the pitch-synchronous overlap-add), however, only with an assumption of a constant fundamental frequency. If this frequency varies in time, it is necessary to use an additional tracking system,