Auris Nasus Larynx 41 (2014) 343–349

Contents lists available at ScienceDirect

Auris Nasus Larynx

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

Evaluation of speech intelligibility in short-reverberant sound fields

Ryota Shimokura [a,*], Toshie Matsui [c], Yuya Takaki [b], Tadashi Nishimura [a], Toshiaki Yamanaka [a], Hiroshi Hosoi [a]

[a] Department of Otorhinolaryngology—Head and Neck Surgery, Nara Medical University, Japan
[b] Department of Classroom Management, Faculty of Education of Future Generations, International Pacific University, Japan
[c] Life Science Centre of Tsukuba Advanced Research Alliance, University of Tsukuba, Japan

ARTICLE INFO

Article history:
Received 25 September 2013
Accepted 28 January 2014
Available online 21 February 2014

Keywords:
Speech intelligibility
Short-reverberant sound field
Deterioration procedure
Low-pass filtering

ABSTRACT

Objective: The purpose of this study was to explore the differences in speech intelligibility in short-reverberant sound fields using deteriorated monosyllables. Generated using digital signal processing, deteriorated monosyllables can lack the redundancy of words, and thus may emphasize differences in sound fields in terms of speech clarity.
Methods: Ten participants without any hearing disorders identified 100 monosyllables convolved with eight impulse responses measured in different short-reverberant sound fields (speech transmission index >0.6 and reverberation time <1 s), and we compared speech recognition scores between normal and deteriorated monosyllables. Deterioration was produced using low-pass filtering (cut off frequency = 1600 Hz).
Results: Speech recognition scores associated with the deteriorated monosyllables were lower than those for the normal monosyllables. In addition, scores were more varied among the different sound fields, although this result was not significant according to an analysis of variance. In contrast, the variation among sound fields was significant for the normal monosyllables. When comparing the intelligibility scores to the acoustic parameters calculated from eight impulse responses, the speech recognition scores were the highest when the reverberant/direct sound energy ratio (R/D) was balanced.
Conclusions: Although our deterioration procedure obscured differences in intelligibility score among the different sound fields, we have established that the R/D is a useful parameter for evaluating speech intelligibility in short-reverberant sound fields.

1. Introduction

Speech intelligibility can be affected by reverberation and noise in an environment. The acoustic features of a sound field can be quantified by examining the time-series of responses to a pulse signal in the environment (impulse response) [1,2]. From the measured impulse response, various parameters can be used to assess speech intelligibility, such as the speech transmission index (STI) [3,4]. The sound of the human voice is modulated by the vocal cords, and so environments that facilitate communication tend to preserve this modulation as the speech information travels from its source to a listener. The STI was developed to evaluate the preservation of modulation depth during the transmission of an acoustic signal. This can be done using a modulation transfer function, which is calculated by applying a Fourier transform to the squared impulse response [5]. An additional parameter for assessing speech intelligibility is clarity (C_{50}) [6–8]. The C_{50} parameter represents the early/late sound energy ratio (in dB) of an impulse response according to an early sound limit, i.e., 50 ms after the direct sound arrives at a destination. The early and late arriving sounds can improve or reduce speech intelligibility, respectively.

Although the parameters for assessing the degree of speech intelligibility in an environment are well defined with a generally standardized relationship [9], the relationship between variables does not hold for sounds traveling through a short-reverberant sound field. For example, in an environment with an STI greater than 0.6, the percentage of correct speech recognition for phonetically balanced words reaches a ceiling at around 90%, and minute changes in the intelligibility of the signal cannot be

* Corresponding author at: Department of Otorhinolaryngology—Head and Neck Surgery, Nara Medical University, 840 Shijo-cho, Kashihara, Nara 634-8522, Japan. Tel.: +81 744 29 8887; fax: +81 744 24 6844.
E-mail address: rshimo@naramed-u.ac.jp (R. Shimokura).

0385-8146/$ – see front matter © 2014 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.anl.2013.12.013

explained by the STI. Although they are generally less reverberant than other types of structures, residential homes may have an STI greater than 0.6. Therefore, the above-mentioned parameters may not be appropriate for assessing speech intelligibility in this environment. This is problematic because the residential structures are frequently spaces for conversation, and even a small amount of reverberation may present communication challenges for elderly people and people with hearing loss [10,11]. Considering the rapidly aging society of Japan and the associated communication challenges, there is a need for an effective evaluation of speech intelligibility in environments with a short-reverberant sound field. New data may lead to better designs for acoustic environments within residential structures.

In this study, we choose to manipulate speech intelligibility using deteriorated monosyllables [12,13]. By intentionally removing the redundancy of words by deteriorating the monosyllables and reducing the articulation, we hoped to characterize the differences in speech intelligibility in different short-reverberant sound fields. One deterioration method involves modulated noise reference unit (NMRU) processing [14]. By adding white noise with envelope modulation that is similar to that of the original speech signal, the instantaneous S/N ratio of the output is kept constant. Another method for deterioration involves low-pass filtering. By removing the high-frequency components of speech, the redundancy can be reduced, similar to how a voice sounds through a telephone [15]. As elderly people often have difficulty detecting high-frequency sounds, this method of deterioration may be particularly useful as a partial simulation of peripheral hearing loss. Deterioration of a sound signal can be used to distinguish speech intelligibility in sound fields with short reverberation periods. However, if a procedure changes the rank order of sound fields in terms of speech intelligibility, then it is not appropriate for evaluating speech intelligibility. We examined monosyllable articulations in eight different sound fields prior to and following a deterioration procedure. We choose to deteriorate the signals using the low-pass filtering method (<1600 Hz) because a previous study had reported that this method produced monosyllables that were sensitive to differences in sound pressure level [16].

There were two main purposes of this study. First, we wanted to examine whether the low-pass filtering method could be used to emphasize the differences in articulation produced by deterioration of the signal. Speech audiometry conducted in Japanese medical institutions usually uses 20 or 50 standardized monosyllables (67-S or 57-S monosyllable table edited by Audiology Japan) that frequently occur in conversation and text. Since listening in a short-reverberant sound environment may make speech intelligibility difficult to discriminate, we used 100 Japanese monosyllables in this study. Our second purpose was to identify the acoustic parameters that affect speech intelligibility of monosyllables in short-reverberant sound fields. In this study, we used a convolution technique to simulate different sound fields with measured impulse responses. As mentioned above, the impulse response indicates the transfer function of reflections in a room, so a dry source convolved with the impulse response represents the sound heard in the sound field. A convolution is an integral that expresses the amount of one function $f(t)$ (i.e., dry source) as it is shifted over another function $g(t)$ (i.e., impulse response), which can be defined as

$$[f*g](t) = \int_0^t f(\tau)g(t-\tau)d\tau, \quad (1)$$

where τ is a delay time for shifting. With this method, we can present various room acoustics using a headphone or in a non-reverberant room.

2. Methods

2.1. Participants

Ten native Japanese speakers (six males and four females ranging from 23 to 36 years of age) participated in this experiment. They did not report having any hearing disorders, and in an audiometry test, the averaged detection threshold for 500, 1000 and 2000 Hz tones did not fall below 20 dB.

2.2. Sound fields

The eight short-reverberant impulse responses used in this experiment were chosen from the environment/architectural acoustics "SMILE 2004" database, which includes sounds measured with a common procedure in actual environments (sampling rate: 44,100 Hz, sampling resolution: 16 bits) [17]. There were seven impulse responses in seven sound fields (drama theater, meeting room, lecture room, wooden house, movie theater, church, and multipurpose hall) and one artificially controlled impulse response (T_{30} = 0.5 s). Examples of the impulse responses are shown in Fig. 1. The reverberation times (T_{30}) calculated from the impulse responses were shorter than 1 s, and the STIs of the sound fields were greater than 0.6. Although these parameters appear to have the largest influence, other relevant parameters may be related to the rank order of speech intelligibility. From each impulse response, we calculated the relative sound pressure level (L_{re}), initial time delay gap (ITDG), reverberation times (T_{30} and Early Decay Time: EDT), early/late sound energy ratios (C_{50}, C_{80} and R/D), center of time (T_s), and STI, and then verified the parameters affecting speech intelligibility. These parameters are often used in the acoustical evaluation in concert halls and opera houses [18].

The L_{re} is the logarithmic ratio of sound pressure to minimum sound pressure, so the L_{re} in the lecture room is 0. The L_{re} represents the total sound energy of the direct and reflection sounds. The reverberation time T_{30} (or EDT) is the time required for

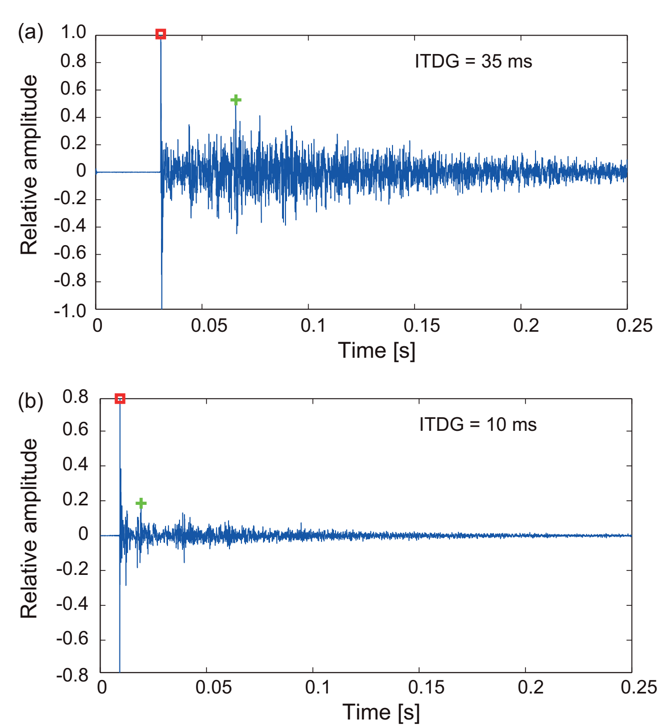

Fig. 1. Measured impulse responses in the drama theater (a) and meeting room (b). The symbols (□ and +) indicate the direct sound and the first reflection, respectively.