Proc. of the Intl. Conf. on Advances In Information Processing And Communication Technology - IPCT 2014.

ISBN: 978-1-63248-021-7 doi: 10.15224/ 978-1-63248-021-7-79

Intelligent Public Address by Means of Adaptation to Speech Transmission Index or Ambient Noise Profile

Rationale and signal processing algorithms

Francis F. Li

***Abstract*—The seemingly mature technologies for public address systems are not often as straightforward as they are thought to be. The usability of such systems depends, to some extent, on their capability to adapt to ambient noise so that the required intelligibility can be achieved, while unnecessary loudness is avoided for the tranquility of environmental sound. The time varying and unpredictable nature of noise in occupied spaces necessities the deployment of intelligent adaptation. A new scheme to achieve a specified speech transmission index, an objective acoustic parameter for speech intelligibility, is developed in this paper based on a set of blind estimation algorithms using machine learning. The paper details the rationale of the method and associated algorithms. In addition, for systems designed to achieve a specific signal to noise ratio, a simplified version is derived.**

***Keywords*—public address system, speech intelligibility, speech transmission index, signal to noise ratio, blind estimation, machine learning, envelope spectrum.**

I. Introduction

Public address (PA) systems or Tannoys have found prevalent applications over a hundred years. In a simple form, a PA system may only have a microphone, an amplifier and a loudspeaker or a cluster of distributed loudspeakers. However PA systems in modern days still encounter some unsolved problems and challenges. In many large public venues such as airports, train stations, convention centres, and shopping malls, sophisticated PA systems are required to comply with relevant standards (in terms of speech intelligibility) for critical applications such as emergency broadcast, and on the other hand to achieve descent tranquillity for less important speech communication. Intelligibility and sound tranquillity depend not only on loudness of sound from speakers but also the ambient noise levels, which are varying. This necessitates the use of adaptive systems with ambient noise sensing.

In essence, to maintain required intelligibility, louder speech is broadcasted to compensate for the disturbance from increased noise level. What noise sensors acquire is a mixture of broadcasted speech and the ambient noise. Simple feedback control is not feasible: the regime of “the higher the noise level, the higher the amplification gain” would place the system in a positive feedback loop, leading to a saturated high level of output.

DSP algorithms are used to solve the problems. The current state of the art is to measure the noise level when there is no speech signal going through the system, i.e. in the interval between speech announcements. Obviously, this method does not continuously adapt to environment: during the course of running speech, the system gain is fixed, regardless of the varying ambient noise. A possible solution to the problem is to estimate the ambient noise from received mixture of speech and noise. Nonetheless, the complexity and the lack of accuracy hampered its tangible applications in the past.

A PA system is often designed to a set of specifications as laid down in the standards and building regulations. One of the most commonly quoted design criteria is the speech transmission index (STI), a combined physical measure of reverberance and signal to ambient noise ratio that gives a good correlation to perceived speech intelligibility. The STI is typically measured with artificial test signals. This makes in-situ measurement of STI difficult. A dual channel method (with prior knowledge of source) to accurately estimate the STI from running speech using machine learning has been developed [1, 2]. The STI can also be estimated from received arbitrary speech (single channel blind method) [3, 4]. These methods can be applied to estimate the STI from the signals received from the noise sensing units. The estimates can then be used to adjust the system gain to achieve the projected STI value, thus approaching the design specification. This paper mainly explores how these estimation tools can be extended to in-situ applications as a means to adapt PA systems continuously to environment to achieve design criteria.

As an acoustic parameter for speech intelligibility, the STI takes into account adverse implications of reverberation and noise on speech. Unlike crowded venues such theatres and arenas, reverberation in some large venues such as airports and railway stations, reverberation is not significantly changed over time, but noise does. When the systems only need to adapt themselves to ambient noise, the method and the algorithms can be simplified significantly.

II. Rationale and Methods

Speech Transmission index, as an objective parameter for the quality of speech transmission systems, was developed to quantify overall effects of adverse conditions of reverberation and noise on perceived speech intelligibility, and has be included in related international standards [5,6,7]. It combines the reverberation time (RT) and ambient noise level in a space and converts them into a single index related to the subject perception of speech intelligibility. The STI has been widely

Francis F. Li

School of Computing, Science and Engineering, University of Salford
UK

used as an objective parameter to assess speech intelligibility in spaces with or without PA systems, and is an international standard requirement many venues, where speech communication is important, need to meet.

The STI method uses low frequency sine waves (from 0.63 Hz to 12.5 Hz) modulated noise to emulate speech signals, where the noise carrier models the noise from voice chord, and the modulator simulates the modulation effects of voice production system when uttering words. Modulation Transfer Function (MTF) is then measured in 98 data points over 14 different modulation frequencies in 7 octave bands. The standard STI measurement method is well defined in literature and laid down in the standards[5,6,7]. The procedure is, in brief, as follows:

A speech spectrum shaped noise carrier *n(t)* is modulated by a very low frequency signal (0.63-12.5 Hz), which emulates the envelope of speech

$$m(t)=\sqrt{1+m\cos(2\pi Ft)} \qquad (1)$$

to generate a test stimulus

$$i(t)=n(t)\cdot\sqrt{1+m\cos(2\pi Ft)} \qquad (2)$$

where *F* is the modulation frequency and *m*, the modulation index. The intensity of the stimulus and response from the speech transmission system are

$$I(t)=\mathbf{I}\mathbf{i}[1+m\cos(2\pi Ft)] \qquad (3)$$

and

$$O(t)=\mathbf{I}_o[1+m_o\cos 2\pi F(t-\varphi)] \qquad (4)$$

where *mo* is the modulation index of the output intensity function and φ is time delay due to transmission. ***I, Ii*** and ***Io*** are amplitudes of corresponding sinusoidal function (mean intensities). The MFT of a channel is defined as the ratio of *mo* to *m* as a function of modulation frequencies.

$$MTF(F)=\frac{m_o}{m} \qquad (5)$$

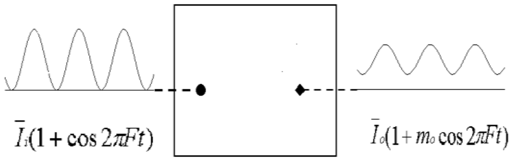

Figure 1. Speech transmission system as a modulation transfer function (for modulation index for the stimulus m=1)

TABLE I. 98 DATA POINT FOR STI CALCULATION

(Hz)	125	250	…..	4 k	8 k
0.63					
0.80					
……					
10.0					
12.5					

Row:7 Octave bands of speech interest
Column:14 Modulation Frequencies representing speech envelope

For good speech intelligibility, the envelope of speech signals should be preserved but noise interference minimized. The MTF takes the advantage of counting both envelope shaping effect and noise, since the input and output signal intensities are considered. The principle of apply the MTF and the standard specified 98 data points are illustrated in Figure 1 and Table 1.

The STI is a nonlinear combination of the MTF data which gives good correlation to perceived speech intelligibility obtained by a large number of subjective tests. It follows the calculation procedures of:

1. Converting *MTF(F)* into apparent *S/N* ratio

$$(S/N)_{app,F}=10\log\left(\frac{MFT(F)}{1-MFT(F)}\right) \qquad (6)$$

2. Limiting dynamic range to 30 dB

if (S/N)app > 15 dB >> (S/N)app = 15 dB
if (S/N)app < -15 dB >> (S/N)app= -15 dB (7)
else (S/N)app=S/Napp

3. Calculating mean apparent S/N ratio:

$$\overline{(S/N)}_{app}=\frac{1}{14}\sum_{F=0.63}^{12.5}(S/N)_{app},F \qquad (8)$$

4. Calculating overall mean apparent *S/N* by weighting the *(S/N)app,F* of 7 octave bands

$$\overline{(S/N)}_{app}=\sum w_k\overline{(S/N)}_{app,F} \qquad (9)$$

where *wk*=0.13, 0.14, 0.11, 0.12, 0.19, 0.17 and 0.14 respectively for the 7 octave bands .

5. Converting to an index ranging from 0 to 1

$$STI=\frac{\overline{(S/N)}_{app}+15}{30} \qquad (10)$$

The STI can also be obtained by a set of artificial neural networks and effective estimators from received running speech [2] by

$$MTF(dB)\approx Er(dB)-Es(dB) \qquad (11)$$

where *Es* and *Er* are estimated envelope spectra of source and received speech respectively. When statistical feature of source signal is further estimated, the STI can be estimated from received speech only, using a hybrid model developed by the author [3]. The key of in-situ application is an effective estimator for the envelope spectra of received speech to enable a fast estimation say within a few seconds.

Hilbert transform was used to estimate signal envelopes. The envelope *e(t)* of a sound signal *s(t)* is obtained via

$$e(t)=\sqrt{s^2(t)+s_h^{\,2}(t)} \qquad (12)$$

where $S_h(t)$ is the Hilbert transform of *s(t)*

$$s_h(t) = H[s(t)] \equiv \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s(t-t')}{t'}dt' \qquad (13)$$

Since the sound signal is digitized. The envelope signal is evaluated using discrete Hilbert transform calculated via FFT. Let s*[n]* denote the *nth* sample of the sound *s(t)*

$$s[n] = s(nT) \quad for \quad n \in [0, n_1, n_2, ..., N-1] \qquad (14)$$

where *T* is the sampling period and *n* denotes the sample number. The discrete Hilbert transform of *s[n]* is then calculated by

$$H\{s[n]\} = \sum_{k=0}^{N}\{A[k]\sin\frac{2\pi kn}{N} - jB[k]\cos\frac{2\pi kn}{N}\} \qquad (15)$$

where the coefficients A and B are determined by the Fourier transform

$$A[k] = Re\{\sum_{n=0}^{N} s[n]e^{-j2\pi kn/N}\} \qquad (16)$$

$$B[k] = Imag\{\sum_{n=0}^{N} s[n]e^{-j2\pi kn/N}\} \qquad (17)$$

The envelope signal is obtained by

$$e[n] = \left|s[n] + jH\{s\{[n]\}\right| \qquad (18)$$

A rectangular window is moved along energy signals of running speech envelope normalised to that of long-time energy signal over 10 seconds (as opposed to the longer period use in [2, 3]) of received speech signals to detect long-term signal intensity. Spectra are calculated with a commonly used Welch algorithm and normalised to long term signal intensity. A set of typical anechoic speech envelope spectra are shown in Figure 2. It can be seen envelope spectra are quite stable statistical feature of running speech. The peak occurs at 4-5 Hz representing the rate of syllables.

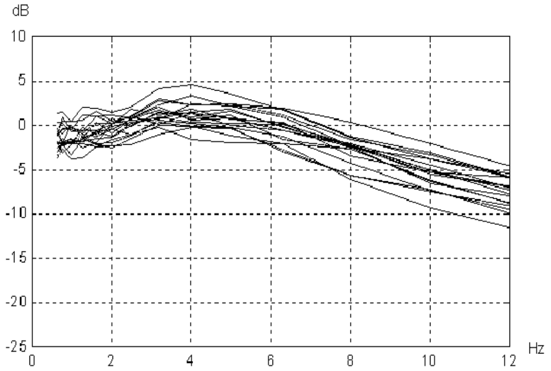

Figure 2. Typical envelope spectra from speech signals

As analysed, the envelope spectrum of a received speech signal is related to the original speech signal and the MTF of the transmission channel. The MTF comprises mixed information about both reverberation and ambient noises. The relation between RT and MTF under idealised situation supports this assertion. The MFT of a room with purely exponential impulse response is [8]:

$$MTF(F) = [1 + (2\pi F \cdot RT / 13.8)^2]^{-1/2} \qquad (19)$$

where *F* is the modulation frequency. The MTF is solely decided by RT and monotonically reduces when RT increases. When white noise is present in the baroground, the MTF becomes:

$$MTF(F) = [1 + (2\pi FRT / 13.8)^2]^{-1/2} \cdot \frac{1}{1 + 10^{(-S/N)/10}} \qquad (20)$$

where S/N is the signal to noise ratio at the listening position. Figure 20 shows how reverberation and noises affect MTFs. The reverberation contribution is modulation frequency related, while the noise effect is independent of modulation frequencies when white noise with equal power per frequency unit is considered. More specifically, the noise effect shifts the MTF pattern downwards, but longer reverberation time causes MTF to decrease more rapidly with respect to the modulation frequencies.

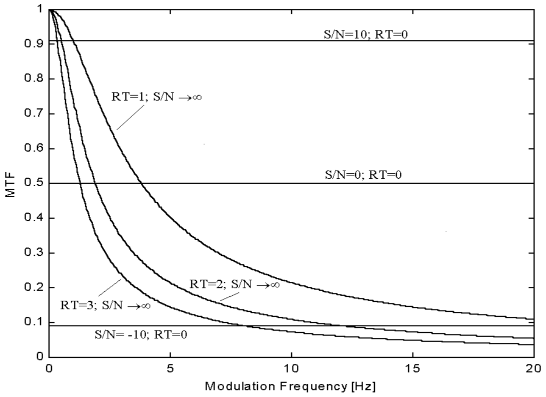

Figure 3. Contributions of RT and S/N to MTF

Given that fact that (1) anechoic speech shows fairly consistent envelop spectra, (2) the MTF can be estimated from envelope spectra of received and original speech signals. (3) the STI is defined on 98 data points from the MTF, a machine learning method using a hybrid artificial neural network and bespoke feature spaces can be used to estimate the STI from received speech signals with reverberation and ambient noise. Details about the algorithms can be found in [2,3]. A block diagram of the the blind estimation method is illustrated in Figure 4. It is worth mentioning that the STI is calculated in 7 octave bands, all the above motioned algorithms need to be performed in these sub-bands and the overall computing overhead is non-trivial, but generally manageable with a standard PC with moderate specification.

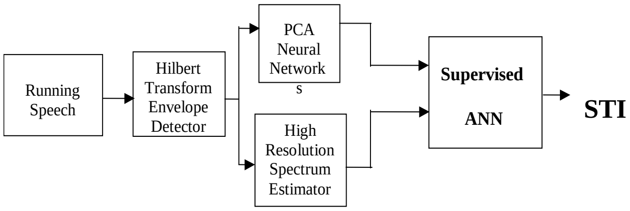
Running Speech
Hilbert Transform Envelope Detector
PCA Neural Network s
High Resolution Spectrum Estimator
Supervised ANN
STI

Figure 4. A block diagram illustration of the blind STI estimator

Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-1-63248-021-7 doi: 10.15224/ 978-1-63248-021-7-79

III. Implementations

A. *Adaptation with Blind STI estimation*

Many PA systems are design to meet certain STI criteria. For example to obtain excellent intelligibility, the STI should normally be greater than 0.75. In venues where reverberation and ambient noise levels both vary over time, full adaptation needs be considered. Examples of these venues include assembly halls, where the number of occupants can significantly change the sound absorption and the reverberation. The noise levels also change depending on activities, number of occupants, etc.. A full adaptation system deploying the blind STI estimator algorithms outline in the previous section is illustrated in Figure 5. Sensing units (measurement microphones) pick up noisy and reverberant speech signals in the venue. The envelope spectrum based blind STI estimator is applied. The detected STI value is used as a guide to adjust the gain of the amplifier. A set of adjustment rules can be adopted for diverse application, taking into account practical issues.

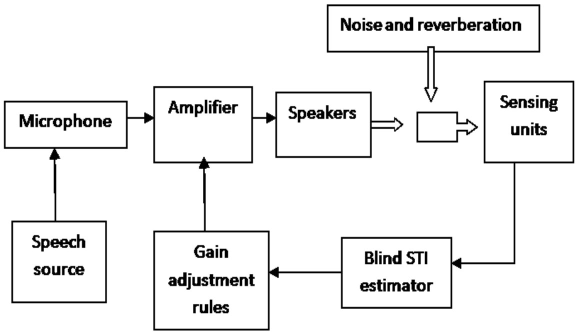

Figure 5. Adaptation to noise and reverberation

In [1,2,3], the purposes are to estimate STIs to very high accuracy, similar to those obtained from traditional measurements with deterministic testing probe stimuli, therefore the STIs are statistically estimated over 45 to 60 seconds. For continuous adaptation, shorter duration estimates, typically over 8 to 10 seconds, are taken to allow for timely updating and tracking of the ambient noise variation. Minor inaccuracy in STI values in this application is tolerable, since the standardised design criteria only call for STI banding.

The rules of amplification gain adjustment determine how the system behaves. They specify a required STI value, a tolerance band and adaptive intervals; in the meantime they should avoid unnecessarily high sound levels and maintain a natural sounding speech (i.e. avoid significantly perceivable volume fluctuation). The optimal design of these rules may need extensive psychoacoustic testing and this is beyond the scope of this paper. Some empirically identified simple rules are used to preliminarily test the proposed method. They are: targeted STI 0.8, tolerance band +/- 0.05, updating interval 1 s, gain adjustment step 1.5 dB. The system continuously monitors the sound over the past 10 seconds, but takes the STI estimation every second, increases the amplification gain by 1.5 dB, if the STI is 0.05 below the target; Decrease the amplification gain by 1 dB if the STI is 0.05 above the set target. A level cap may be applied so that the system cannot produce an unreasonably high level of output.

B. *Adaptation with STI estimates from envelope spectra subtraction*

In most PA system applications source signals of speech is readily available. An arguably more precise way to estimate the STI is to use the information from both the received speech signals and the speech source. Recall Equation (11), one can obtain the MTF by the difference between envelope spectra of transmitted speech and original speech. This is illustrated in as a block diagram in Figure 6. In this alternative method, envelope spectra are calculated by Equations 14-18, followed by Welch spectrum estimation. MTF estimation is calculated by Equation 11, and STI is obtained from Equations 6-11. The gain control rules adopted for this study is the same as the ones in the previous sub-section.

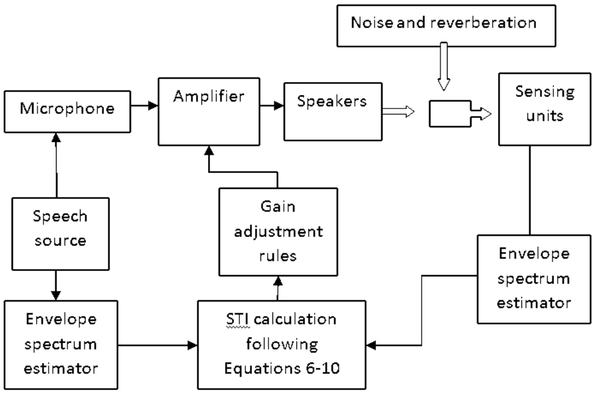

Figure 5. Adaptation via envelope spectra subtraction

C. *Adaptation to ambient noise only*

There are many venues where reverberation does not significantly change over time. In such cases, the adaptation can be made simpler. The problem becomes adaption of speech level to noise level. Figure 3 and Equation (20) give insight into this: in fact when modulation frequency *F=0,* the DC component of the MTF is determined by signal to noise ratio only. In other words, the DC component can be used to calculate the signal to noise ratio. Nonetheless, this is equivalent to the measurement of noise level during the pause period of speech.

Figure 2 shows that in the envelope spectra of speech there are quite significant energy in 0.63 Hz sub band. Figure 3 shows that in this very low modulation frequency sub band, the reverberation effect on the MFT is negligible (less than 10%). Therefore the MTF value is the 0.63 Hz band is used to approximate that of DC to estimate the signal to noise ratio. The estimated signal to noise ratio is compared with the pre-set value and to increase or decrees the amplification gain.

Proc. of the Intl. Conf. on Advances In Information Processing And Communication Technology - IPCT 2014.
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-1-63248-021-7 doi: 10.15224/ 978-1-63248-021-7-79

IV. Testing Results

The three methods were tested in a large room with a mid-frequency reverberation time RT=0.7 s. Background noise samples were taken from real world noise in various venues recorded in-situ. The background noise was played back during the tests using a separate system. The level was adjusted up and down automatically by a computer programme. The emulated ambient noise level varied from 40 to 70 dB(A). The STI estimator developed in [2] was used as an STI meter to monitor the STI. Pre-recorded male and female voices were used as speech sources. Methods A, B and C as described in sub Sections A, B and C respectively in the previous section were tested.

For Methods A and B, the target was to achieve a predefined STI value. In the experiments, the STI was set to be 0.8. During a 30 minute test with vary ambient noise. The amplification gains of both systems tracked the changes of noise level and maintained the STI reading within 0.71 to 0.87. Method A and B showed no significant difference in adaptation performance.

For Method C, the target was to maintain a signal to noise ratio at around 15 dB, +/- 3 dB. The signals applied to ambient noise emulator and PA speakers are monitored, through calibration, they were used to determine the signal to noise ratio. Change of amplification gain was with a step of 1.5 dB, updating every 10 seconds. The method did function to adapt, though the measured signal to noise discrepancy over a 30 minute test went up to 7.2 dB (Laeq over 60 seconds). This suggests that a better noise estimator needs to be developed.

V. Concluding remarks

This paper has presented three new methods to continuously adapt a PA system to ambient noise and reverberation or to ambient noise only. The former is achieved using a modified version of a blind STI estimation method developed previously or an envelope spectrum subtraction algorithm with added feedback control strategies, while the latter is based upon signal to noise ratio estimated from 0.63 Hz band in modulation transfer function. Tests show all three methods offer continuous adaptation of the volume of the PA systems to environment. While the STI based methods seem to outperform the signal to noise ratio based one, the noise level estimator in the latter can be improved.

The proposed methods only adjust the amplification gain of the PA systems, i.e. turning the volume up and down. As the intelligibility is affected by sub band masking, more advantageously, the spectrum of ambient noise might be analysed, and a strategic scheme based on psychoacoustics adopted to equalise the speech signal so that the best intelligibility is achieve at least increase of overall sound level. But this is beyond the scope of the current study, and will be left as future work.

References

[1] F. F. Li and T. J. Cox, “A neural network for blind identification of speech transmission index”, in the *Proceedings of IEEE ICASSP2003’*, v. II, pp. 757-760, 2003.

[2] F. F. Li and T. J. Cox, “Speech transmission index from running speech: A neural network approach,” *Journal of Acoust. Soc. Am.*, Vol. 113, Issue 4, pp.1999-2008, 2003.

[3] F. F. Li and T. J. Cox, “A Neural Network Model for Speech Intelligibility Quantification”, Applied Soft Computing, Vol. 7, Issue 1,pp. 145-155, January, 2007.

[4] F. F. Li, “Estimation of intelligibility from received arbitrary speech signals with support vector machine”, proceedings of the 2005 IEEE international conference on Machine learning and cybernetics, (ICMCL 2005), Vol. 6, pp. 3755-3760, 2005.

[5] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica, Vol. 28, 1973, p. 66-73.

[6] H. J. M. Steeneken and T. Houtgast, “A physical method for measuring speech transmission quality,” J. Acoust. Soc. Am. Vol. 67, No. 1, 1980, p. 318-326.

[7] IEC Standards, 60268-16, (also ISO, BS EN 60268-16), Sound system equipment, Part 16: Objective rating of speech intelligibility by speech transmission index, 4th Rev., 2011.

[8] M. R.Schroeder, “Modulation trasnfer function: Definition and measurement,” Acustica Vol 49, pp. 179-182, 1981.

About Author:

Francis F. Li was born in Shanghai, China. He received a B.Eng. from the East China University of Science and Technology, an MPhil from University of Brighton, and a PhD from the University of Salford, UK. Francis is currently with the School of Computing Science and Engineering at the University of Salford, where he teaches a variety of modules on BSc and MSc levels, supervises PhDs, and carries out research.

Prior to his current appointment, he was a senior lecturer in Computer Science at the Manchester Metropolitan University. His research interests include speech, music and multimedia signals processing; artificial intelligence and soft-computing; architectural acoustics; data and voice communications; bio-medical engineering; and instrumentation.

Dr. Li has published 100 research papers and a book. He is Associate Editor in Chief for SPIJ and Associate Technical Editor for J. AES.