Abstract

A method to estimate speech intelligibility has been developed that mimics the multi-resolution nature of human hearing and can be used with any non-repeating input signal. Previous intelligibility metrics are defined and implemented assuming that background noise is measured separately or that the transfer function is measured with a specific signal, such as modulated noise, Maximum Length Sequence (MLS), or sine sweep. The new metric uses the coherence function, which is related to the complex-valued frequency response and can directly yield a signal-to-noise ratio (SNR) in the frequency domain, given a known input signal and a measured output signal. The coherence function is sensitive to any energy not present in the input, whether the extra energy comes from distortion, reverberation, or background noise. Significantly, coherence can be calculated in real time for any source or any noise level, even if the signal level changes with time. When used with different analysis lengths for different frequencies (multi-resolution), as is common practice in the audio industry, coherence can be used to replace the modulation transfer function in the Speech Transmission Index (STI) standard. The new metric is called Speech Coherence Index (SCI), and is compared to STI under simulated conditions with uncorrelated noise; with a single reflection; and with synthesized reverb, uncorrelated noise, and changes in direct level. The SCI value responds similarly to the STI, but is more consistent under certain conditions.

Keywords: Speech Intelligibility, Coherence, Frequency Response, Multi-Resolution, Signal to Noise Ratio (SNR)
An intrinsically source independent method to estimate intelligibility

1 Introduction

Most speech intelligibility metrics make use of the concept of Signal-to-Noise Ratio (SNR), whether explicitly or implicitly. When expressed in decibels, the SNR ranges from minus infinity with zero signal energy to positive infinity with zero noise energy, and is zero when the signal and noise energies are equal. Changes in SNR values do not map linearly to changes in perception [1]: a change from 0 to +10 dB SNR is much larger perceptually than a change from 100 to 110 dB SNR.

In general, speech intelligibility metrics are expressed as an index that ranges between 0 and 1, often expressed as a percentage from 0-100%. In the case of the Speech Transmission Index (STI), the result ranges from 0.0 with no signal to 1.0 with no noise, and is 0.5 when the signal and noise are equal. Changes in STI have been shown to match changes in intelligibility reasonably well [1]. For very simple cases with no reverb, STI and SNR are related by

\[ \text{STI} = \frac{\text{SNR}}{\text{SNR}+1} \quad \text{and} \quad \text{SNR} = \frac{\text{STI}}{1-\text{STI}}, \]

where SNR is the energy ratio (not expressed in decibels).

The coherence function expresses how much of the output energy of a system is correlated with the input energy, and how much is uncorrelated (i.e. noise). It has a minimum value of 0.0 when there is no correlated output, a value of 0.5 when the correlated and uncorrelated energies are equal, and a value of 1.0 when there is no uncorrelated output. The coherence function can be calculated in real time along with the frequency response; does not require a separate measurement, or even prior knowledge, of the background noise; and can be calculated with many kinds of signal. This paper describes a way to compute intelligibility metric similar to STI but using coherence instead of the modulation transfer function. This new method is called the Speech Coherence Index (SCI).

2 Background

2.1 The Coherence Function

The coherence function is a measure of the average correlation between input energy and output energy over time, and is given by

\[ \gamma^2 = \frac{G_{xy}^2}{G_{xx}G_{yy}}, \]

where \( \gamma^2 \) is the coherence, \( G_{xy} \) is the average input-output cross-spectrum, \( G_{xx} \) is the average input auto-spectrum, and \( G_{yy} \) is the average output auto-spectrum [3]. The cross-spectrum \( G_{xy} \) can be thought of
simply as the correlation between the output and input. In this paper we will refer to $\gamma^2$ as the coherence (as opposed to just $\gamma$) because it fulfills the relationship
$$\gamma^2 = \frac{SNR}{SNR+1}$$
and can be more directly integrated into the new metric.

### 2.2 Factors Influencing Coherence

Any signal which is present at the output of a system but not present at the input will lower the coherence. At a music or sports venue, coherence is reduced by all signals present at a measurement microphone but not at the output of the mixing console, such as HVAC noise, crowd sounds, or 60 Hz hum in the PA system. With a two channel frequency response analyzer, the reduction in coherence can be measured in real time for any nonrepeating input signal. As mentioned in the introduction, the coherence produces a value of 0.5 when the speech and noise energies are equal.

For non-repeating signals, reverberation will reduce coherence when the reverberant energy arrives outside of the coherence time window, which is equivalent to the energy being uncorrelated noise. This happens with signals that don’t repeat, such as speech or noise. Any early reflections that occur within the time window will not reduce coherence, since they have a constant phase offset from the direct signal.

For repetitive input signals, such as a MLS or swept sine, with a short repeat length may have reverberation that arrives at the microphone one time window later than the input signal. If the input signal also repeats one time window later, the energy is correlated and coherence will not decrease.

### 2.3 Time and Frequency Resolution

The length of the analysis window affects how much coherence is reduced by reverberation, and must be chosen carefully. In order to have narrow frequency resolution, one must use a long analysis window. In order to have narrow time resolution, one must use a short analysis window. For the goal of measuring speech intelligibility, this metric will adopt a time and frequency resolution inspired by human hearing. Many phenomena demonstrate that the human hearing has narrow frequency resolution and poor time resolution at low frequencies (equivalent to a long analysis window), and broad frequency resolution and good time resolution at high frequencies (equivalent to a short analysis window) [4]. This is suggested by measurements of pitch just-noticeable-difference, critical bandwith, time integration of loudness, delayed evoked otoacoustic emissions, and even direct auditory nerve impulse responses.
Frequency response analyzers that use progressively longer time windows at lower frequencies have existed for several decades, and are known as dual channel source independent spectrum analyzers [5]. Figure 1 shows that the analyzer used for this paper (SIM3, Meyer Sound Laboratories) has time windows that follow the trend of human hearing, with constant time/frequency resolution at low frequencies and logarithmically narrower time resolution and wider frequency resolution at higher frequencies.

Using different analysis times at different frequencies means that a reflection may reduce coherence at high frequencies when it falls outside the time window, but not at low frequencies where it falls within. When a reflection falls within the time window, the combing effect is visible in the magnitude response: this does not affect coherence, because the combing effect is linear and has a constant phase offset.

Figure 2 shows how the reflections that arrive at longer delays reduce coherence over progressively more of the spectrum, starting with high frequencies and short time windows.

Reflections and reverberation have a similar affect on the Speech Transmission Index [6]. The STI is calculated by measuring the Modulation Transfer Function (MTF) at modulation frequencies from 0.63 to 12.5 Hz (corresponding to periods from 1587 to 80 ms). A reflection which arrives between 80 and 1587 ms will significantly reduce the MTF at high modulation frequencies, but minimally reduce the MTF at low frequencies. So the STI value is also reduced by reflections arriving at longer delays.

### 3 Speech Coherence Index

#### 3.1 Speech Coherence Index Calculation

The SCI metric is calculated by replacing the Modulation Tranfer Function (MTF) in the STI metric with the coherence measurement. All other steps of the STI are kept, which include calculation of masking...
Figure 2 Coherence, magnitude, and phase response for a single echo at -6 dB magnitude and variable delay, as calculated by the multi-resolution analyzer described in Figure 1.

Threshold, band weighting, linearization between ±15 dB SNR, band importance weighting, etc. Specifically, the substitution happens in the equation that corrected the MTF values by auditory masking, thresholding, and background noises:

\[ m' = m \times \frac{l_k}{l_k + l_n + l_{\text{Mask}} + l_{\text{Thresh}}}, \]

where \( m' \) is the corrected m value, m is the derived modulation transfer ratio value, \( l_k \) is the signal energy, \( l_n \) is the noise level, \( l_{\text{Mask}} \) is the auditory masking factor, and \( l_{\text{Thresh}} \) is the reception threshold. For SCI, this equation becomes:

\[ m' = \frac{l_k}{l_k + N_{\text{Eff}} + l_{\text{Mask}} + l_{\text{Thresh}}}, \]

where the effective noise is now \( N_{\text{Eff}} = l_k \frac{1-y^2}{y^2} \) and \( y^2 \) is averaged per octave band. Note that this replacement also obviates the subsequent summation over modulation frequencies.
3.2 Noise from an uncorrelated source

For a noise source completely uncorrelated from the signal, and without any reflections or reverberation, STI and SCI have identical values, as seen in Figure 3.

3.3 Effect of a single reflection

For the case of a single reflection, SCI and STI both start at a high index close to 1 when the reflection delay is small, and slowly drop to a plateau level for longer delays, as seen in Figure 4. For very long latencies, the reflection is entirely outside the analysis window and reduces coherence at all frequencies. When the reflection is equal in energy to the direct sound (0 dB) and arrives after the longest analysis window it is equivalent to uncorrelated noise with an SNR of 0 dB, and results in an SCI of 0.5. Other reflection levels converge to the SCI value expected for the equivalent SNR at long latencies (Equation 1). At 0 dB magnitude and short latencies (around 1 ms), SCI is not equal to 1 because the complete combing produces energy dips in the output signal where the coherence goes to 0: when averaged over octave bands, these dips pull the average down from 1.

STI values do not converge to a single value at long echo delay times, but rather oscillate as the echo delay times correspond to the nodes and anti-nodes of the modulation frequencies. Note also that a 0 dB reflection does not reach an STI value of 0.5, as it did for uncorrelated noise.

Figure 3 STI and SCI versus SNR for an uncorrelated noise source and no reverberation.

Figure 4 SCI and STI as a function of latency for a single reflection at three different levels.
3.4 Effect of reverberation

Both metrics were evaluated with impulse responses synthesized to have different reverberation times by time stretching the original impulse response. Acoustically, this corresponds to a fixed surface area of absorption in rooms of different cubic volume. The total energy in the reverberation was -13 dB in arbitrary units, and the input signal was pink noise. Figure 5 shows the STI and SCI values for different values of noise levels, direct level, and reverberation time. Both STI and SCI decrease when reverberation time increases, noise level increases, or direct level decreases. SCI is more sensitive to direct level decreases, especially at low reverberation times. A preliminary analysis has shown that this is due to the reverberation of loud sounds masking subsequent quieter sounds, even with a quasi-stationary stimulus such as pink noise.

Figure 5 STI and SCI as a function of noise level, reverberation time, and direct level.
4 Conclusions

SCI has several advantages over STI, at least for the simple conditions studied here. Importantly, it can be estimated in real time and is independent of the source signal. Although both metrics share many trends, they do not produce identical results. For simple stimuli such as a single reflection, SCI does not have the artifacts seen with STI, which include oscillations for long single reflection latencies and lack of convergence to 0.5. For more realistic reverberation, both metrics are affected similarly by noise level, direct level, and reverberation time. SCI is affected more strongly by decreases in direct level, especially at low reverberation times.

5 Future work

Given the promising results shown here, SCI will be evaluated under more realistic conditions. This includes considering reverberation measured from a variety of rooms, especially for spaces where the difference between STI values does not correspond well to the subjectively perceived difference. Also, other input signals such as speech will be evaluated, since this includes the modulation frequencies relevant for intelligibility in a way that both pink noise and the STI’s modulation frequencies lack. In addition, the direct relationship between SCI and intelligibility scores will be measured.

References