Speech security outside meeting rooms

Carl Hopkins and Matthew Robinson
Acoustics Research Unit, School of Architecture, University of Liverpool, Liverpool L69 7ZN, United Kingdom.

Ken Worrall and Tim Jackson
Her Majesty's Government Communications Centre (HMGCC), United Kingdom.

Summary

A new approach to provide speech security outside meeting rooms is described where a covert listener might attempt to extract confidential information. Experiments are used to establish a relationship between an objective measurement of the Speech Transmission Index (STI) and a subjective assessment relating to the threshold of information leakage. This threshold is defined for a specific percentage of English words that are identifiable with a maximum safe vocal effort (e.g., "normal" speech) used by the meeting participants. Experimental results show that it is possible to quantify an offset that links STI with a specific threshold of information leakage which describes the percentage of words identified. The offsets for male talkers are shown to be approximately 10 dB larger than for female talkers. Therefore for speech security it is possible to determine offsets for the threshold of information leakage using male talkers as the "worst case scenario." To define a suitable threshold of information leakage, the results show that a robust definition can be based upon 1%, 2%, or 5% of words identified. For these percentages, results are presented for offset values corresponding to different STI values in a range from 0.1 to 0.3.

PACS no. 43.55.Hy,43.71.Gv,43.72.Dv

1. Introduction

Before confidential discussions take place inside dedicated meeting rooms it is beneficial to be able to quantify the speech security. Such a measure can be used to ensure adequate protection from a casual overhearing or deliberate interception by a covert listener (aided or unaided by the use of electronic and/or electroacoustic equipment).

Research by Gover and Bradley [1,2,3] formed the basis for ASTM E2638-10 'Standard test method for objective measurement of the speech privacy provided by a closed room' [4]. This was intended for eavesdroppers that were unaided by the use of electronic and/or electroacoustic equipment. The test method combines the single number rating of sound insulation with background noise levels to obtain a Speech Privacy Class (SPC). The Standard states that the method does not set criteria for adequate speech privacy and provides guidance on interpreting different values of SPC. Potential issues concerning the application of this Standard to speech security are evident in its statement that "People speak at different levels and vary their voice level in reaction to room noise and other acoustical factors. Consequently it is not possible to say definitely whether a room is protected against eavesdropping. One can only assign a probability of being overheard.". This results in a non-mandatory appendix outlining an approach to setting criteria.

The approach in this paper develops an alternative approach for more rigorous accreditation of speech security which considers the talker's vocal effort inside the meeting room. The aim is to prescribe a simple figure of merit for a meeting room that accounts for speech security in terms of the talker's vocal effort, the transmission path which involves the sound insulating structure and the background noise in the vicinity of the listener. In practice it is the loudest utterances that are most likely to be heard or intercepted; hence a room which can securely contain quiet speech may not be secure for loud speech. For this reason, the performance of a meeting room can be quantified in terms of how loudly the talkers can safely converse. For technical users of a meeting room this quantity can be given in terms of a sound pressure level for a particular vocal effort. For non-technical users it can be translated into a vocal effort label which is purely descriptive and easily understood. Existing catalogues of vocal effort levels [e.g. 5] allow this translation. This allows a Vocal Effort Level (VEL) to be defined as the sound pressure level measured at 1m on-axis in anechoic conditions for a specified description of vocal effort relating to the labels: ‘normal’, ‘raised’, ‘loud’, or shouted’ speech. The Maximum Safe Vocal Effort Level (MSVEL) indicates how loudly a talker can safely speak for a specified degree of speech security. Different degrees of speech security are realized by defining a threshold of information leakage. The threshold of information leakage is the MSVEL for which a covert listener may be aware of the talker’s voice and its cadence, but is only able to identify a certain percentage of words. For example, in a high-security situation, a meeting room could be specified such that a maximum of only 1% of words are identifiable when the talkers in the room use a ‘raised’ vocal effort level. For most speech security applications, it is expected that the acceptable percentage of identifiable words will typically be much less than 20%. The threshold of information leakage is a quantitative measure that must be derived from subjective experiments and these are the subject of this paper.

An objective test procedure for the approach outlined above can be based around measurement of the Speech Transmission Index (STI). All STI values less than 0.3 are classified as indicating ‘bad’ intelligibility; hence, by itself, STI is not appropriate to quantify thresholds of information leakage. However, by measuring the signal level from an artificial mouth or loudspeaker which results in an STI less than or equal to 0.3, there is the potential to make a link between STI and a specific threshold of information leakage that describes the percentage of words identified by using an offset in decibels. This paper describes the listening tests used to quantify this offset.

The authors have recently published a journal paper [6] describing their approach to speech security and the reader is referred to this for more details than is possible to present in this conference paper.

2. Subjective experiments

In the experiments, subjects listen to the Harvard sentences at different VELs that have been processed to simulate the transmission of speech from inside to outside a meeting room. They identify the words that they can hear in each sentence to give the percentage of words identified as a function of the tested VEL, *W*(VEL). The aim is to identify the percentage of words identified at any VEL, denoted as *W*, by using interpolation to give *W* as a smooth function of VEL. It is then possible to determine the VEL for specific values of *W* which are used to define a threshold of information leakage with an offset relating to 0.3 STI. To define the required threshold of information leakage, *W* values of particular interest are likely to be 0%, 1%, 2%, 5%, 10%, and 20%.

Each subject is presented with speech stimuli that simulate transmission through a wall/door of a meeting room in the presence of masking noise outside the room. By considering a wall and a door each with seven VELs and repeating the combinations five times using different word lists, the experiment uses approx. 400 words to obtain W_{av}(VEL) for the wall and the door. This requires seventy word lists from the Harvard sentences which are presented in random order.

The approach primarily considers a covert listener using headphones to monitor a microphone placed outside a room. Hence stimuli were presented to subjects over headphones.

Figure 1 illustrates data from one listener showing the seven discrete VEL values that result in a fitted curve from which the offset is determined for 0.3 STI at any value of *W*. The offset for each listener and construction is calculated using the following steps: (1) determine the *W*(VEL) for each word list, (2) define W_{av}(VEL) as the average of the *W*(VEL) values from the five word lists, (3) calculate the line of best fit for W_{av}(VEL) using cubic spline interpolation, and (4) calculate the offset for the chosen value of *W*, listener and construction type by calculating the difference between VEL for a given W_{av}(VEL) from the line of best fit and the VEL which produces 0.3 STI.

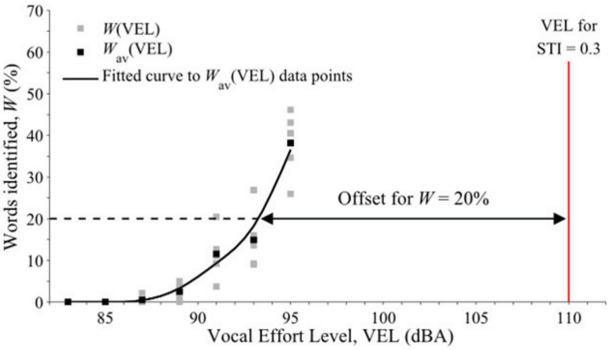

Figure 1. Example data from one participant indicating the percentage of words that were identified at different values of VEL.

1.1. Subjects

The threshold of hearing was determined for each potential participant. Only those with a hearing loss less than 20dB HL took part in the experiment. Forty untrained listeners between the ages of 18 and 58 were recruited as subjects (20 male and 20 female). All listeners used English as a first language. Five male and five female subjects listen to speech stimuli from one of the four talkers. All experiments received approval from the University of Liverpool ethics committee.

1.2. Listening tests

Listeners carried out the tests inside an audiometric booth. The preferred listening level was assumed to be 60dBA but each subject could choose their own listening level at –5, 0 or +5dB relative to this preferred level. The tests used a Matlab GUI into which the subject typed the words they heard within a time limit of 18s. Subjects were asked not to guess any words, the word order was ignored and incorrect spelling was identified with the following rules: (a) allow mis-spellings using ‘a’ instead of ‘e’, (b) ignore punctuation such as apostrophes, (c) allow homonyms, (d) allow either American or British English spelling.

Anechoic recordings of speech were processed to include the reverberation inside a hypothetical meeting room, sound transmission through the wall or door, and masking noise. The signal processing chain used to simulate transmission of the speech signal from inside to outside the room for the experiments is summarized on Figure 2.

The primary variable is the VEL of the speech because the masking noise level is fixed hence it is the reverberation inside the room and sound transmission through the wall/door that alter the signal level for the listener. The VEL used in the experiments is assigned depending upon the gender of the talker and the efficacy of the different sound insulating elements. The VEL needs to be adjusted to account for the sound power of the speaker. This sound power level needs to be based upon a generic directivity for a talker. It is determined by measuring the directivity of a B&K Type 4128 HATS with a loudspeaker in the oral cavity. To measure this directivity, the HATS is installed in an anechoic chamber with 17 microphones arranged in a hemisphere at 1m from the oral cavity. Measured sound pressure levels are converted to a sound power level specific to the HATS.

Reverberation is then added to the speech signal using an image-source method to generate a room impulse response. Room dimensions are 3m x 2m x 2.5m representing a small meeting room with a reverberation time of 0.5s at all frequencies. The impulse response is then combined with the speech signal using convolution via the frequency domain. The next step is to determine the incident sound upon the sound insulating element. As the sound reduction index, *R*, uses a ratio of transmitted to incident sound power, only the fraction of sound incident upon the wall/door has to be included in the calculation; hence the reverberant sound pressure level in the room is reduced by 6 dB based on the assumption of a diffuse field inside the room [7]. Transmission via the wall/door is discussed in Section 1.2.2. The masking noise is synthesised as described in Section 1.2.3 and then added to the processed speech signal to give the stimuli for presentation to the subjects.

This approach allows the speech to be auralised at a receiver position outside the room.

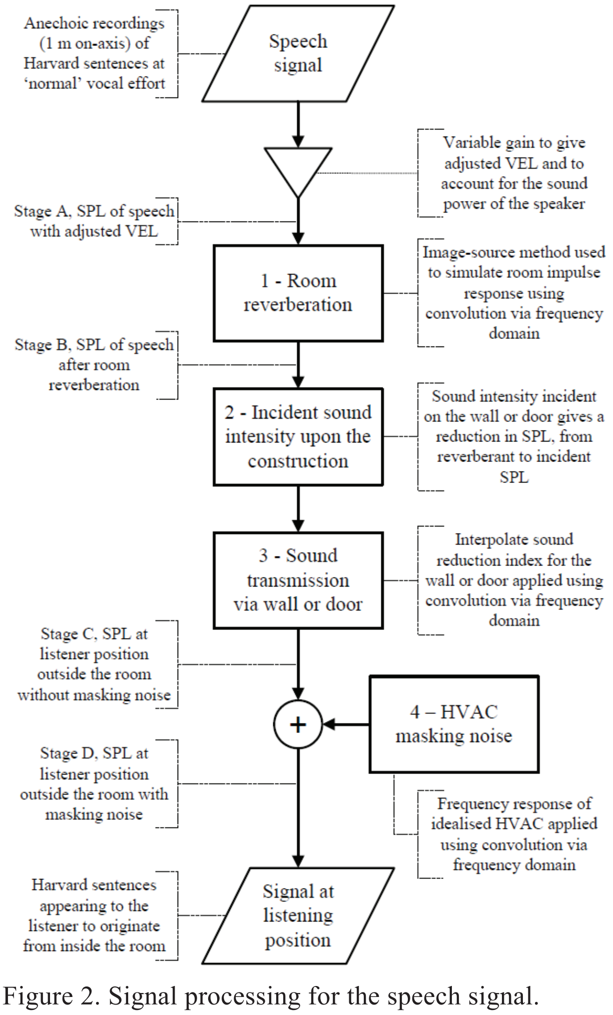
Anechoic recordings (1 m on-axis) of Harvard sentences at ‘normal’ vocal effort
Speech signal
Variable gain to give adjusted VEL and to account for the sound power of the speaker
Stage A, SPL of speech with adjusted VEL
1 - Room reverberation
Image-source method used to simulate room impulse response using convolution via frequency domain
Stage B, SPL of speech after room reverberation
2 - Incident sound intensity upon the construction
Sound intensity incident on the wall or door gives a reduction in SPL, from reverberant to incident SPL
3 - Sound transmission via wall or door
Interpolate sound reduction index for the wall or door applied using convolution via frequency domain
Stage C, SPL at listener position outside the room without masking noise
+
4 – HVAC masking noise
Stage D, SPL at listener position outside the room with masking noise
Frequency response of idealised HVAC applied using convolution via frequency domain
Harvard sentences appearing to the listener to originate from inside the room
Signal at listening position

Figure 2. Signal processing for the speech signal.

1.2.1. Talkers

Four talkers recorded word lists at normal vocal effort in an anechoic chamber with a 0.5” B&K Type 4155 microphone that was positioned 1m in front of the mouth, i.e. on-axis. These talkers were native speakers of British English with speech that was perceived as being close to Received Pronunciation without a strong regional accent. Male talkers A and B were aged 26 and 42 years old respectively and female talkers A and B were aged 63 and 30 years old respectively. The 720 Harvard sentences were used as source material for the recordings. These form 72 word lists where each list comprises 10 sentences with each sentence typically having approximately 8 words. On average each sentence comprises 87.5% one-syllable words and 12.5% two-syllable words. There are only three three-syllable words in the entire set of 720 sentences.

1.2.2. Sound insulation

It is cost effective to use of masking noise outside the room with more modest levels of sound insulation; hence two different sound insulating elements were considered. The wall construction comprised two sheets of plasterboard on both sides of 50mm light steel studs giving 43dB R_w. The door was a solid timber doorset with seals giving 35dB R_w. The sound transmission loss is simulated using the approach described by Vorländer [8].

1.2.3. Masking noise

Mechanical ventilation from air conditioning is chosen as the masking noise for the speech as this is common in many buildings. To ensure it is well-defined and repeatable, it is synthesised by shaping broadband noise between 20Hz and 20kHz to an idealised frequency response that is typical for mechanical ventilation systems in buildings. The frequency response for the idealised masking noise is approximately flat up to 200Hz then decreases at 6dB per octave. The experiments use a fixed masking noise level of 50dB $L_{A,eq}$.

1.3. Experimental procedure

The experimental procedure for each subject is as follows: (1) Randomise the order of the stimulus variables (construction type, and number of VEL values) using multi-dimensional random number generation. (2) Randomise the order of the word lists. (3) Randomise the order of the ten sentences in the selected word list. (4). Present the chosen sentence to the listener at the chosen VEL. (5) Ask the subject to identify all the words that they hear by typing them into a text box in the Matlab GUI within a time limit of 18s. The subject was asked not to guess any words. Note that the order of the words typed by the subject is ignored as there are no repeated words in the sentences (other than ‘the’). Subjects are allowed to check their spelling before submitting these words. Incorrect spelling is identified and assessed after the experiment using the following rules: (a) allow mis-spellings using ‘a’ instead of ‘e’, (b) ignore punctuation such as apostrophes, (c) allow homonyms and (d) allow either American or British English spelling. [Steps (4) and (5) are repeated for the ten sentences in the selected word list]. (6) Calculate *W*(VEL) for the selected word list at the chosen VEL. [Steps (3) to (6) are repeated for the 70 selected word lists.]

The STI of the simulated acoustic environment is measured according to EN 60268-16. Using the specified male and female speech filters, the STI test signal (Maximum Length Sequence) is set to the same A-weighted level as the largest VEL for male or female speech given in Table I. An STI test of the signal processing system is then performed using a test signal at this A-weighted level. This test is then repeated after altering the level of the test signal until 0.3 STI is measured for the system. The impulse response that is measured using speech shaped MLS is processed using the commercial software, DIRAC, to determine the STI. The A-weighted level of the test signal that is required to achieve 0.3 STI gives the VEL which is used to calculate the offset. The VEL required to achieve 0.3 STI for the wall is 110dBA (male speech), 111dBA (female speech), and for the door is 99dBA (male speech), 100dBA (female speech).

3. Results

A mean offset is calculated by taking the mean of the offsets for each listener, each type of construction and each value of *W* used to define the threshold of information leakage. In setting this threshold there may be circumstances where *W*=0% is desirable This might be possible with female talkers but with male talkers there were 4 of the 20 subjects that were able to identify at least one word in at least one word list at the lowest VEL. For this reason it is not possible to consider W=0%. Using 0.3 STI for the wall, Table II shows the offset data for male and female talkers from all listeners.

Table I. Range of VEL values used for the wall and door with male and female talkers.

	Element	Vocal Effort Level (dBA) (Lowest on the left to the highest on the right)						
Male talker	Wall	83	85	87	89	91	93	95
	Door	72	74	76	78	80	82	84
Female talker	Wall	92	94	96	98	100	102	104
	Door	79	81	83	85	87	89	91

Table II. Offsets for the wall.

		Words identified (%)				
		1	2	5	10	20
Male talkers	VEL for STI = 0.3 (dBA)	110	110	110	110	110
	Mean VEL for word ID (dBA)	87.7	88.7	90.2	91.7	93.2
	Mean offset (dB)	22.3	21.3	19.8	18.3	16.8
	Median offset (dB)	22.2	21.3	20.0	18.5	16.8
	Upper 95% CI of offset (dB)	25.1	23.7	22.1	20.6	18.6
	Maximum offset (dB)	24.8	23.1	21.4	19.8	18.2
Female talkers	VEL for STI = 0.3 (dBA)	111	111	111	111	111
	Mean VEL for word ID (dBA)	97.8	99.1	101.1	102.5	103.2
	Mean offset (dB)	13.2	11.9	9.9	8.5	7.8
	Median offset (dB)	13.2	11.9	10.0	8.6	7.6
	Upper 95% CI of offset (dB)	15.9	15.2	12.8	10.6	8.5
	Maximum offset (dB)	17.8	15.3	13.3	10.8	8.2

The words ‘a’ and ‘the’ are considered to have no information content and are therefore removed from the analysis as the two most common words in the Harvard sentences. The full data for the door and wall is published in [6] and indicates that the offset is different for male and female talkers; the offset for male talkers is approximately 10dB larger than for female talkers. Hence it is possible to define the threshold of information leakage based upon the ‘worst case scenario’ for the wall and the door; this will be for male talkers. For the wall and the door the offset is similar (within ≈2dB); however to make a decision on a suitable offset the larger value will be chosen (i.e. wall).

The offsets only apply to the specific sound insulation curves for the wall and door with the chosen masking noise spectrum. However the offsets are similar and therefore it is reasonable to assume that they will apply to any construction with sound reduction indices in individual one-third octave bands that fall between the values for this wall and door. Figure 3 shows the offset for male talkers. On the assumption that the threshold of information leakage is less likely to correspond to percentages of words identified of 10% and above, the analysis here focuses on 1%, 2% and 5%. The next step is to identify the parameter most suitable to quantify the offset and therefore indicate the MSVEL. The risk in choosing an offset using the mean or median is that covert listeners can be expected to be better than the average listener. Hence there are two viable options to set the offset, either use the upper 95% confidence interval or the maximum offset from the 40 listeners. Fortuitously, male talkers have a larger offset than female talkers; hence the decision to use the 95% confidence interval or the maximum offset determined by an individual’s responses is unnecessary because they are nominally identical. Therefore from Table II the offsets (integer values) for 1%, 2% and 5% of words identified are 25dB, 24dB and 22dB respectively.

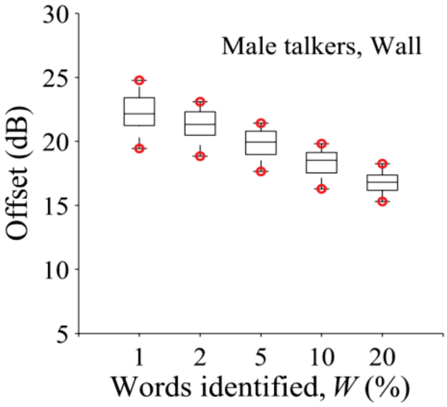

Figure 2. Offsets for male talkers for different *W* values for the wall. Median value lies within the box where the box represents lower and upper quartiles respectively. Whiskers indicate 95% confidence intervals. Minimum and maximum offsets are shown with open circles.

Table III. Offsets with male talkers.

STI (-)	Words identified		
	1%	2%	5%
	Offset (dB)		
0.1	14	13	11
0.2	20	19	17
0.3	25	24	22

Results from individual listeners indicate a spread in *W*(VEL). Hence there is a potential problem in defining threshold of information leakage when there is a large spread in the percentage of words identified for each word list. For example, if 5% of words identified is chosen to represent the threshold of information leakage then if the average value at a specific VEL had individual word lists with 30% of words identified then this would compromise speech security. As the offset is being defined by male talkers, an assessment can be made of the robustness by choosing 1%, 2%, 5%, 10% and 20% of words identified to define a threshold of information leakage. There is potentially a problem when the average number of words identified at each VEL is 10% or 20% because the maximum percentage of words identified in an individual word list are as high as 22% and 37% respectively. Note that for the door they were only slightly lower (20% and 34% respectively). This justifies choosing 1%, 2% or 5% of words identified to robustly define a threshold of information leakage. It could be beneficial for the measurer to have other offsets available when 0.3 STI cannot be measured due to low power artificial mouths or loudspeakers or when meeting rooms have high sound insulation. Hence, Table III shows the offset for STI values of 0.1, 0.2 and 0.3 and *W* values of 1%, 2% and 5%.

4. Conclusions

An objective approach to assess speech security outside meeting rooms is proposed using measurements based on STI. This differs from the ASTM approach [4] in that it requires explicit consideration of the talker's vocal effort inside the meeting room. This leads to the definition of a maximum safe vocal effort level to indicate how loudly a talker can safely speak inside a room for a specified degree of speech security. Different degrees are realized by defining a threshold of information leakage as the maximum safe vocal effort level for which a covert listener may be aware of the talker's voice and its cadence, but is only able to identify a certain percentage of words.

Acknowledgement

CH and MR gratefully acknowledge the research funding from HMGCC.

References

[1] B.N. Gover and J.S. Bradley. Measures for assessing architectural speech security (privacy) of closed offices and meeting rooms. J. Acoust. Soc. Am. 118 (6) 3480–3490 (2004).

[2] J.S. Bradley, M. Apfel and B.N. Gover. Some spatial and temporal effects on the speech privacy of meeting rooms. J. Acoust. Soc. Am. 125(5) 3038-3051 (2009).

[3] J.S. Bradley and B.N. Gover. Speech levels in meeting rooms and the probability of speech privacy problems. J. Acoust. Soc. Am. 127 (2) 815–822 (2010).

[4] ASTM E2638-10 Standard test method for objective measurement of the speech privacy provided by a closed room, American Society for Testing and Materials.

[5] I.R. Cushing, F.F. Li, T.J. Cox, K. Worrall, and T. Jackson. Vocal effort levels in anechoic conditions. Appl. Acoust., 72 695–701 (2011).

[6] M. Robinson, C. Hopkins, K. Worrall, T. Jackson. Thresholds of information leakage for speech security outside meeting rooms. J. Acoust. Soc. Am. 136(3) 1149-1159 (2014).

[7] C. Hopkins, Sound Insulation (First ed., Butterworth-Heinemann, Oxford, 2007, ISBN: 978-0-7506-6526-1).

[8] M. Vorländer, Auralization: Fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality (First ed., Springer, Berlin, 2008, ISBN: 978-3540-4882-93)