COMPARISON OF SUBJECTIVE AND OBJECTIVE RATINGS OF INTELLIGIBILITY OF SPEECH RECORDINGS

Bradford N. Gover and John S. Bradley
Institute for Research in Construction, National Research Council, Ottawa, ON K1A 0R6, Canada
brad.gover@nrc-cnrc.gc.ca

1. INTRODUCTION

The intelligibility of speech recordings made in rooms and other spaces can be affected by a range of factors, including reverberation, noise, distance from talker to microphone, and properties of the microphone system itself. This paper presents some results regarding the evaluation of the intelligibility of speech recordings made under controlled conditions with a variety of microphone systems. Test speech sentences were recorded, and the intelligibility of those recordings was determined by a subjective listening test. Additionally, STIPA, a form of the Speech Transmission Index intended to predict intelligibility for electroacoustic systems [1] was determined for each recording condition.

2. METHOD

Recordings of test speech sentences [2] were made under controlled conditions in 4 test spaces having widely varying acoustical properties: see Table 1.

Test Space	*Volume (m^3)*	*Noise (dBA)*	*RT (s)*	*Description*
C	148	46.1	1.5	Reverb chamber with added absorption.
K	77	49.9	0.4	Domestic space.
R	892	67.3	–	Large, noisy.
T	–	63.8	0.05	Pickup truck.

Table 1: Descriptions of test spaces. The noise in each was generated via playback of recorded noise over loudspeakers.

In each space, a small loudspeaker played the test speech and the STIPA test stimulus at a fixed level, and recordings were made at 3 recording positions with each of 7 microphone systems, some in several different configurations. In all, 83 unique recordings were made, 44 of which were 2-channel. The 2-channel recordings were also analyzed as 1-channel, resulting in a grand total of 83 + 44 = 127 recordings. Each recording contained 5 unique test sentences, each of which was rated by 10 participants in a subjective intelligibility test. A total of 40 subjects participated, and the testing was approved by the NRC Research Ethics Board (O-REB Protocol 2006-47). One at a time, participants listened over headphones to the recorded sentences. After each, they repeated to the test operator the words they understood, as well as their judgments on a 7-point scale of *"How difficult did you find it to understand the speech?"* (1=Extremely Difficult, 7=Not Difficult At All) and *"How would you rate the quality of this speech recording?"* (1=Low Quality, 7=High Quality). The fraction of words correctly identified was determined for each sentence, and for a particular recording, the average score was determined from 50 responses (10 subjects × 5 sentences).

3. RESULTS

The intelligibility results for recordings made with four of the microphone systems at one location within each test space are shown in Fig. 1. These devices all used omnidirectional microphones, and differed in terms of size, preamplification, and processing. The black bars indicate the average intelligibility score for the 1-channel recording made with device D*i*, in configuration C*j*, in position M*k*, for the test space indicated (C, K, R, or T). The light grey bars are the scores for the 2-channel recording, where one existed. An asterisk at the left indicates that the difference between 1 and 2-channel scores was significant ($p < 0.05$).

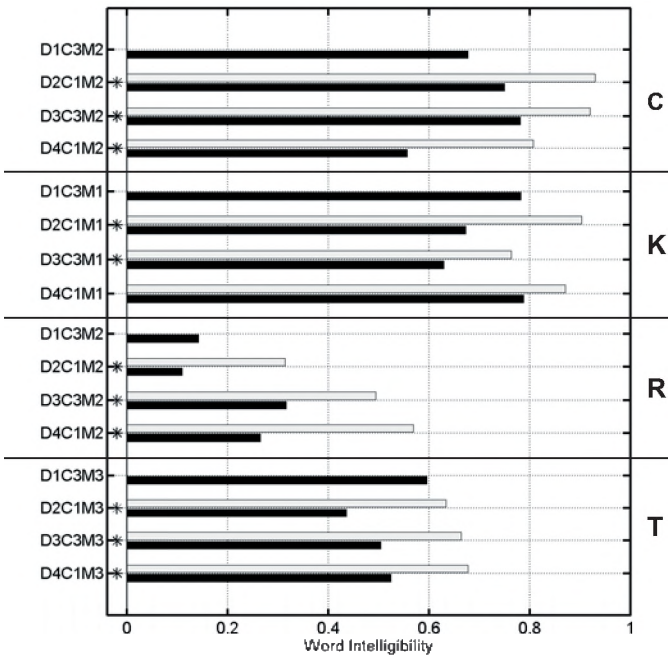

Figure 1: Selected word intelligibility scores: black bars for 1-channel, light grey bars for 2-channel. An asterisk at left implies a significant difference between the two ($p < 0.05$).

Within each test space, the variation in score indicates the variation with recording system at a fixed location: among 1-channel recordings, a substantial difference in intelligibility of almost 0.2 (20%) can result.

The scores for the 2-channel recordings were always higher than the corresponding 1-channel recordings. This is further demonstrated in Fig. 2, which plots the 2-channel score versus the 1-channel score for all 44 such recording conditions tested. The 2-channel intelligibility scores were on average 0.13 (13%) higher.

The judgments of difficulty and of quality are plotted versus intelligibility score in Fig. 3 for all 83 1-channel measurement cases. There is the expectation that as intelligibility increases; a sense of difficulty will decrease [3], and presumably, a sense of quality will increase. This is seen in the figure. Only for cases with intelligibility greater than about 0.8 (80%) were ratings greater than 4 (i.e., on the "Not Difficult", "High Quality" side of "Neutral").

Figure 4 shows the intelligibility scores plotted against STIPA for the 83 1-channel recordings. Overall, the correlation is not very high (R^2=0.35), but the breakdown by test space shows that it is lowest for test space R (R^2=0.21) and highest for test space K (R^2=0.71). The accuracy of STIPA as an objective predictor of speech intelligibility is not guaranteed. In some noise and reverberation conditions, it does not accurately predict intelligibility.

4. CONCLUSIONS

For a single recording location, the intelligibility of speech recordings varied by up to 0.2 (20% word score) depending on recording device type.

The intelligibility of 2-channel recordings was significantly higher than corresponding 1-channel recordings, by an average of 0.13 (13%).

Ratings of difficulty and quality improved (i.e., decreased, increased respectively) as intelligibility scores increased.

STIPA was not generally well correlated with subjective intelligibility, but for a subset of the data (from particular test spaces), the relationship is stronger.

REFERENCES

[1] IEC 60268-16 "Objective rating of speech intelligibility by speech transmission index," IEC Switzerland (2003).

[2] "IEEE recommended practice for speech quality measurements," IEEE Trans. Aud. Electroacoust., **17** (1969).

[3] Kobayashi *et al.*, "Optimum speech level to minimize listening difficulty in public spaces," J. Acoust. Soc. Am., **121** (2007).

ACKNOWLEDGEMENTS

Thanks to M. Stinson, G. Daigle, J. Quaroni, and R. Hartwig of NRC Institute for Microstructural Sciences for assistance and guidance in making the measurements.

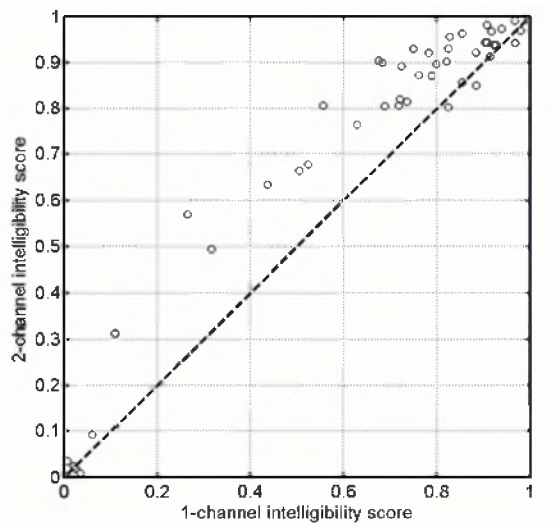

Fig 2: Intelligibility scores for 2-channel recordings versus corresponding 1-channel recordings for 44 cases. The dashed line indicates where the points would lie if equal.

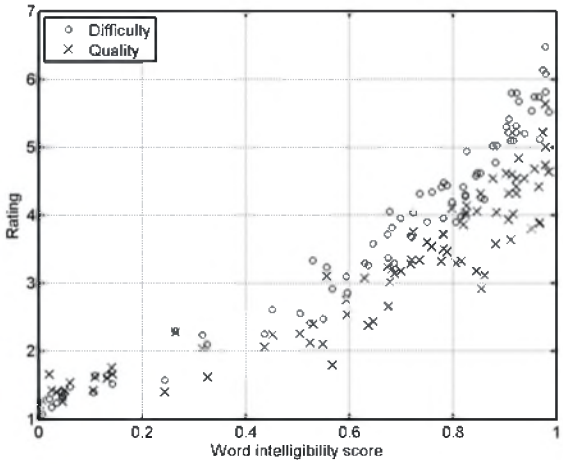

Fig 3: Ratings of Difficulty (1=Extremely Difficult, 7=Not Difficult At All) and Quality (1=Low Quality, 7=High Quality) versus intelligibility scores.

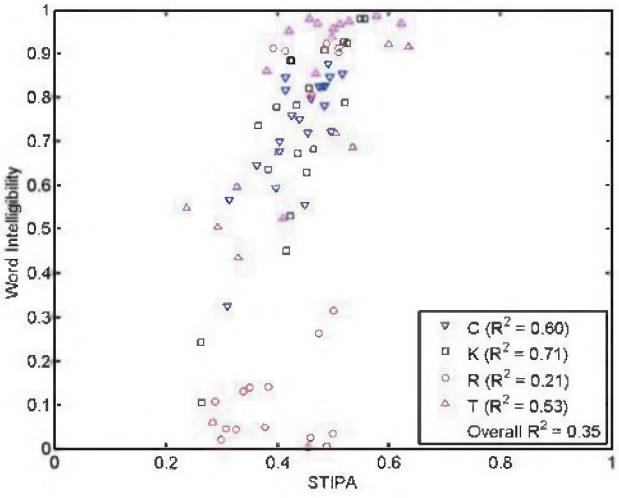

Fig 4: Intelligibility score versus STIPA for all 83 1-channel recordings, broken down by test space (C, K, R, and T).