A MULTIPLE BANDWIDTH OBJECTIVE SPEECH INTELLIGIBILITY ESTIMATOR BASED ON ARTICULATION INDEX BAND CORRELATIONS AND ATTENTION

Stephen D. Voran

Institute for Telecommunication Sciences
325 Broadway, Boulder, Colorado, 80305, USA, svoran@its.bldrdoc.gov

ABSTRACT

We present ABC-MRT16—a new algorithm for objective estimation of speech intelligibility following the Modified Rhyme Test (MRT) paradigm. ABC-MRT16 is simple, effective and robust. When compared to subjective MRT data from 367 diverse conditions that include coding, noise, frame erasures, and much more, ABC-MRT16 (containing just *one* optimized parameter) yields a very high Pearson correlation (above 0.95) and a remarkably low RMS estimation error (below 7% of full scale.) We attribute these successes to concise modeling of core human processes in audition and forced-choice word selection. On each trial, ABC-MRT16 gathers word selection evidence in the form of articulation index band correlations and then uses a simple attention model to perform word selection using the best available evidence. Attending to best evidence allows ABC-MRT16 to work well for narrowband, wideband, superwideband, and fullband speech and noise without any bandwidth detection algorithm or side information.

Index Terms— ABC-MRT, ABC-MRT16, articulation index, modified rhyme test, MRT, objective estimator, speech intelligibility

1. INTRODUCTION

Many subjective testing techniques have been developed to quantify speech intelligibility and they can provide meaningful and consistent results for specific application areas. Tutorials can be found many places, including [1]-[3].

Common attributes of these tests include carefully prepared speech material presented to screened listeners in highly controlled environments. Listeners then answer content-based questions or repeat what was heard, and analysis of their responses drives speech intelligibility results. One approach is rhyme testing, and a specific form called the Modified Rhyme Test (MRT) [4] is standardized in [5]. The MRT is based on 50 lists. Each list consists of six English language keywords of the form consonant-vowel-consonant and the six differ only in the initial or final consonant (e.g., "not," "tot", "got," "pot," "hot," "lot".) The listener's task is to select which of the six keywords was presented and the rate of correct word identification produces a measure of speech intelligibility. The US National Fire Protection Association has specified the MRT for critical communications testing and thus we have performed substantial amounts of MRT work [6]-[9].

Subjective speech intelligibility test can be time-consuming and costly and there is good motivation to use signal processing algorithms to estimate speech intelligibility instead. Many such efforts have been published over the years and examples can be found in [1]-[3] and [10]-[12]. Work more specifically aligned with the MRT includes [13] and [14].

A fundamental understanding of how different frequency bands contribute to speech intelligibility dates to the work of Harvey Fletcher in the 1920's. Fletcher's studies produced the idea of articulation bands and the intelligibility estimator called articulation index (AI) [15].

We built upon this fundamental and venerable work when we developed the Articulation Band Correlation Modified Rhyme Test (ABC-MRT) algorithm [16]. ABC-MRT is a narrow band (NB) speech intelligibility estimator that emulates the Modified Rhyme Test (MRT). It directly compares temporal correlations across the 17 articulation index bands [1], [15] that cover the NB (roughly 250 to 3850 Hz in this case) spectrum. The maximum correlation drives the selection of one of six possible words in each of the 17 bands, and the rate of successful word identification becomes the measure of speech intelligibility, just as as in the MRT. This very simple approach sufficiently emulated human MRT behavior for NB test material and we reported good agreement between ABC-MRT output and MRT results across 139 NB test conditions in [16]. Several organizations have leveraged ABC-MRT for NB measurement work and we used it heavily to great advantage in [9]. In addition, ABC-MRT provided the basis for a Japanese language intelligibility estimator [17].

We also cited extension to wideband (WB) signals (approximately 50-7000 Hz) as a future priority, pending access to WB MRT data [16]. WB speech communication systems are becoming more common and superwideband (SWB) and fullband (FB) speech systems (with nominal bandwidths 50-16,000 Hz and 20-20,000 Hz, respectively) are emerging.

We have now completed WB, SWB, and FB MRTs and have used the data (properly partitioned) to develop and eval-

This paper was presented at and will appear in the Proceedings of the 42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP2017), New Orleans, LA, March 5-9, 2017.

uate a new speech intelligibility estimation algorithm called ABC-MRT16. This algorithm is much more than a simple bandwidth extension of ABC-MRT. ABC-MRT16 employs a very effective yet concise model of core human processes in audition and forced-choice word selection. On each trial, ABC-MRT16 gathers word selection evidence in the form of articulation index band correlations and then uses a simple attention model to perform word selection using the best available evidence. The word selection success rate provides the basis for speech intelligibility estimation. Attending to best evidence allows ABC-MRT16 to work well for NB, WB, SWB, and FB speech and noise without any bandwidth detection algorithm or bandwidth side information.

In the following sections we describe the algorithm in detail, characterize its outstanding performance on 367 diverse conditions, and provide additional insights into its successes.

2. ABC-MRT16 ALGORITHM DESCRIPTION

ABC-MRT16 is a signal processing algorithm that can process NB, WB, SWB, or FB speech and noise signals. It emulates the MRT in order to provide estimates of MRT speech intelligibility. The algorithm uses articulation index band temporal correlations following the insights found in [18]. AI band based time-frequency (T-F) patterns are key to the functioning of the algorithm. The T-F pattern of an impaired speech signal is correlated with corresponding patterns of six unimpaired word options. ABC-MRT16 departs dramatically from ABC-MRT because it then uses a model for attention that allows it to *compare the best evidence* for each of the six word options.

2.1. AI band correlations

The T-F patterns used in ABC-MRT16 are computed as follows. Time-domain samples x_t ($f_s = 48$ kHz) are Hann windowed in blocks of 512 samples (10.7 ms) with 75% overlap. After DFT, coefficients are converted to power and Stevens' Law [19] is used to produce a simple and robust first-order approximation of loudness (exponent 0.3). Each block produces a column in the matrix $\hat{\boldsymbol{X}}$. Thus the entries of $\hat{\boldsymbol{X}}$ are

$$\hat{x}_{i,k} = \left| \frac{1}{\sqrt{N}} \sum_{t=1}^{N} w_t x_{(k-1)B+t}\, e^{\frac{-j2\pi(i-1)(t-1)}{N}} \right|^{(2\times 0.3)}, \qquad w_t = \sin^2\left(\frac{\pi(t-1)}{N-1}\right),$$
$$N = 512,\ B = 128,\ i = 1 \text{ to } 215,\ k = 1 \text{ to } N_x\,, \quad (1)$$

where N_x is the number of blocks available in x_t. $\hat{\boldsymbol{X}}$ is then normalized to produce $\tilde{\boldsymbol{X}}$ where each row (each time-history at fixed frequency) has zero-mean and unit norm:

$$\tilde{x}_{i,k} = \frac{\hat{x}_{i,k} - \hat{x}_{i,\cdot}}{\sqrt{\sum_{k=1}^{N_x} (\hat{x}_{i,k} - \hat{x}_{i,\cdot})^2}}, \quad \hat{x}_{i,\cdot} = \frac{1}{N_x} \sum_{k=1}^{N_x} \hat{x}_{i,k}\,. \quad (2)$$

The resulting matrix $\tilde{\boldsymbol{X}}$ contains N_x columns, each associated with a time increment of 2.7 ms and $M = 215$ rows covering 0 to 20,063 Hz with a resolution of 93.75 Hz. A key component of ABC-MRT16 is the set of precomputed $\tilde{\boldsymbol{X}}$ patterns, one for each of the 1200 original undistorted keyword realizations (50 lists $\times$ 6 words $\times$ 4 talkers).

To apply ABC-MRT16 to a system-under-test (SUT) or condition, we pass input recordings through the SUT to produce output recordings. ABC-MRT16 then transforms an output recording to a T-F pattern $\hat{\boldsymbol{Y}}$ using (1). $\hat{\boldsymbol{Y}}$ does not require the normalization given in (2) until a later step. The next step is to evaluate $\hat{\boldsymbol{Y}}$ with respect to each of the six keyword T-F patterns $\tilde{\boldsymbol{X}}$.

Let $\tilde{\boldsymbol{X}}$ be a matrix containing an original keyword T-F pattern and $\hat{\boldsymbol{Y}}$ be a matrix containing a T-F pattern obtained from an SUT output (containing at least a keyword). $\tilde{\boldsymbol{X}}$ is M by N_x and $\hat{\boldsymbol{Y}}$ is M by N_y, with $N_x \le N_y$. First we must locate the keyword within $\hat{\boldsymbol{Y}}$. We assume that the SUT delay is approximately constant for the duration of the keyword. We use articulation bands 3 and 4 (rows 7-9, 505-795 Hz) to locate the keyword. These bands contain greater average speech power than other bands, so with no further assumptions about the noise and distortion produced by the SUT, these bands are most likely to be useful for locating the keyword. Let $\hat{\boldsymbol{y}}_i(t)$ be a column vector of N_x samples from the i^{th} row of $\hat{\boldsymbol{Y}}$:

$$\hat{\boldsymbol{y}}_i(t) = [\hat{y}_{i,t+1}, \hat{y}_{i,t+2}, \ldots, \hat{y}_{i,t+N_x}]^T, \quad i = 7 \text{ to } 9,\ \ t = 0 \text{ to } N_y - N_x. \quad (3)$$

We normalize $\hat{\boldsymbol{y}}_i(t)$ to $\tilde{\boldsymbol{y}}_i(t)$ using the process specified in (2). Let $\tilde{\boldsymbol{x}}_{\boldsymbol{i}}$ be the column vector that contains the i^{th} row of $\tilde{\boldsymbol{X}}$ and find the lag t cross-correlation at frequency i:

$$\rho_i^2(t) = \tilde{\boldsymbol{y}}_i(t)^T \tilde{\boldsymbol{x}}_{\boldsymbol{i}}, \quad i = 7 \text{ to } 9,\ \ t = 0 \text{ to } N_y - N_x. \quad (4)$$

The maximizing time shift t^* is the shift that best matches the contents of $\hat{\boldsymbol{Y}}$ with the keyword in $\tilde{\boldsymbol{X}}$:

$$t^* = \arg\max_t \left(\sum_{i=7}^{9} \rho_i^2(t) \right). \quad (5)$$

Once maximizing time shift t^* has been determined we use it to calculate correlations for the remaining frequencies. We use (3) to extract $\hat{\boldsymbol{y}}_i(t^*)$ from $\hat{\boldsymbol{Y}}$, then normalize $\hat{\boldsymbol{y}}_i(t^*)$ to $\tilde{\boldsymbol{y}}_i(t^*)$ using (2), and cross-correlate each of these vectors with the corresponding row of $\tilde{\boldsymbol{X}}$ using (4), resulting in $\rho_i^2(t^*)$. Finally we accumulate correlation results across AI bands and eliminate any negative results:

$$r_j^2 = \max\left(\sum_{i \in B_j} \rho_i^2(t^*), 0 \right), \ \ j = 1 \text{ to } 21. \quad (6)$$

B_j is the set of frequency indices that comprise the j^{th} AI band $j = 1$ to 20 as given in [1]. Note that B_{20} ends at 7 kHz

This paper was presented at and will appear in the Proceedings of the 42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP2017), New Orleans, LA, March 5-9, 2017.

as that was the upper frequency originally considered significant for speech intelligibility. Ignoring spectrum above 7 kHz means ignoring the additional speech intelligibility contributions of SWB and FB, however small. Thus we combine all remaining (up to 20 kHz) frequency samples ($i = 77$ to 215) into a single "super AI band" B_{21} for completeness. Table 2 shows the modest contribution of this band in the intelligibility estimation context.

Due to normalizations, (6) is equivalent to a single cross-correlation for each AI band. Thus we have calculated 21 AI band correlations (r_j^2) for each candidate keyword.

2.2. Attention, decisions, success rate, and intelligibility

ABC-MRT16 proceeds to emulate an intuitive hypotheses for the human execution of a forced-choice MRT trial. A listener hears a stimulus and forms an auditory precept. The listener must compare that precept with six possible explanations for the formation of that precept (the six keywords that may have been uttered) and select one. The listener is innately compelled to attend to the most convincing evidence available for making the selection and will safely ignore the least convincing evidence. In other words, attention is drawn to the AI bands that best support the decision making task.

The culmination of 2.1 is that (6) produces 21 AI band correlations (r_j^2) for each candidate keyword. We must now introduce the keyword argument $\kappa = 1$ to 6, and (6) now produces $r_j^2(\kappa)$, $j = 1$ to 21, $\kappa = 1$ to 6. For each keyword we next sort the 21 correlation values in descending order:

$$\left\{\tilde{r}_s^2(\kappa)\right\}_{s=1}^{21} = \text{sort}\left(\left\{r_j^2(\kappa)\right\}_{j=1}^{21}\right), \text{ resulting in :}$$
$$\tilde{r}_1^2(\kappa) \geq \tilde{r}_2^2(\kappa) \geq \ \ldots \ \geq \tilde{r}_{21}^2(\kappa), \ \ \kappa = 1 \text{ to } 6. \quad (7)$$

For each keyword (7) organizes the AI band evidence from most convincing (highest correlation) to least convincing (lowest correlation). Next we compete the first s_0 pieces of evidence (s_0 is a fixed value):

$$\kappa_s^* = \arg\max_{\kappa} \tilde{r}_s^2(\kappa), \ \ s = 1 \text{ to } s_0. \quad (8)$$

Eqn. (8) compares the best evidence for each of the six words to produce κ_1^*. It then compares the second best evidence for each of the six words to produce κ_2^*, continuing on through $\kappa_{s_0}^*$. Thus (unlike in ABC-MRT) these comparisons are *not aligned by AI bands*, rather they are *driven by attention.*

Once these s_0 selections have been made, they are compared with the correct word for the trial (κ_0) using the indicator function δ and the outcomes are averaged to produce a success rate c:

$$c = \frac{1}{s_0}\sum_{s=1}^{s_0} \delta(\kappa_s^* - \kappa_0), \ \delta(0) = 1, \ \delta(x) = 0 \text{ for } x \neq 0. \quad (9)$$

An MRT listener produces either success or failure on each MRT trial, but different listeners may produce different outcomes on identical trials. ABC-MRT16 is deterministic and thus will produce identical outcomes for identical inputs. In order to properly represent a *population* of listeners, ABC-MRT16 produces an estimated success *rate* for each trial. This disrupts the parallelism between a single listener and the algorithm, but it is necessary in order to acknowledge listener variation.

The success rate c can then be averaged across all available trials for a given condition or SUT to produce $\bar{c}$. Following the MRT procedure, we then correct for guessing using the affine transformation that maps $\frac{1}{6}$ (the success rate for guessing) to 0 and 1 (perfect keyword identification) to 1, thus converting estimated success rate to estimated MRT intelligibility, c':

$$c' = \frac{6}{5}\left(\bar{c} - \frac{1}{6}\right). \quad (10)$$

It is possible that the amount of evidence used to arrive at a decision (represented by s_0) will vary depending on the listener and the level of difficulty posed by a given MRT trial. But the goal of ABC-MRT16 is to produce per-condition results for an ensemble of listeners and trials. Thus we have selected a single value, $s_0 = 16$ AI bands, that best represents the aggregate behavior seen in MRT results. Note that this is the *only* optimized parameter in the ABC-MRT16 algorithm.

3. EVALUATION AND DISCUSSION

Our development and evaluation work is driven by MRT test results. We previously developed ABC-MRT using 28 conditions and evaluated its behavior across 139 conditions taken from 4 different tests. These 139 NB conditions are summarized in [16] and further details are given in [6]-[8]. This testing supported the land-mobile radio (LMR) communications needs of public safety officials, especially firefighters. MRT input recordings were mixed with high-level background noise recordings (e.g., alarms, saws, pumps, crowds), then passed through protective masks and various components of existing and emerging digital and analog LMR systems.

We developed ABC-MRT16 using that same data, augmented with 168 additional conditions described in [9] for a total of 307 conditions. These new conditions include a total of 28 different NB, WB, and FB codec modes operating at bitrates from 4.4 to 48 kbps in six different noise environments. This development process included experimentation with different treatments of frequencies above 7 kHz, different ways to model attention and estimate success rate, and optimization of only *one* parameter, s_0.

Even though there is only one optimized parameter in ABC-MRT16, we insist on evaluation with unseen data to prevent over-fitting and falsely optimistic results. Thus the results reported here also include 60 additional conditions produced by a total of five WB and SWB codec modes (12.65 and 13.2 kbps), operating in three noise environments, with frame erasure rates of 0, 5, 10 and 20%. Note that frame

This paper was presented at and will appear in the Proceedings of the 42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP2017), New Orleans, LA, March 5-9, 2017. ICASSP Proceedings are Copyright © IEEE.

erasures can impair intelligibility in a temporally localized fashion and as such they present a radically different unseen impairment for ABC-MRT16. These 60 unseen and highly unique conditions make up 16% of the 367 conditions used to evaluate ABC-MRT16 here. They were scored by crowdsourced MRT (CMRT) using 1200 trials per condition and 524 listeners. We have recently demonstrated that our CMRT protocol produces results that very closely align with laboratory MRT (LMRT) results [20]. Further our CMRT results are more repeatable than LMRT results [20]. We have no reason to consider CMRT data to be inferior to LMRT data.

The final output of ABC-MRT16 is the estimated success rate corrected for guessing c', given in (10). We measure the performance of ABC-MRT16 by comparing per-condition values of c' with per-condition subjective MRT scores ϕ. Results for ABC-MRT and ABC-MRT16 are provided in Table 1. The Pearson correlation coefficient is a normalized measure of the covariance between c' and ϕ that ranges from -1 to 1. It reports how well the *relative* scoring of ABC-MRT16 and MRT agree. RMSE is an *absolute* measure of agreement that has the same units as the MRT scores ϕ.

	ABC-MRT		ABC-MRT16	
	ρ	RMSE	ρ	RMSE
139 NB Cond. [16]	0.955	0.073	0.9710	0.061
367 Conditions	0.874	0.124	0.954	0.066

Table 1. Agreement with MRT for ABC-MRT and ABC-MRT16. 367 conditions include NB, WB, SWB, and FB.

ABC-MRT16 offers extremely high correlation and low estimation error, especially in light of the diversity of conditions represented. Fig. 1 provides a visual representation of these results. The estimation error magnitude is less that 0.05 for 57% of the conditions, less than 0.10 for 87% of the conditions, less than 0.15 for 98% of the conditions, and in every case less than 0.18.

Objective-to-subjective RMSE is usually further minimized through a first or second order polynomial mapping of objective estimates. In the case of ABC-MRT16 the first order least-squares fit between c' and ϕ is nearly null:

$$\phi \approx \hat{\phi} = \alpha c' + \beta, \text{ with } \alpha = 0.977, \text{ and } \beta = 0.031. \quad (11)$$

Table 2 shows that this fit produces almost no reduction of RMSE. These results affirm that ABC-MRT16 is capturing the desired behavior from simple first principles and no ad-hoc corrections are necessary. Thus the output of ABC-MRT16 is c', not $\hat{\phi}$. This table also shows the small advantage provided by band 21 and the large advantage provided by the attention model. Comparing the final two lines shows that without attention, adding band 21 is actually harmful.

Table 1 also shows ABC-MRT (17 AI bands and no attention model) performance for the 367 conditions. The final row of Table 2 gives performance for 20 AI bands and no attention model on those same 367 conditions. Comparing the two reveals that including AI bands 18, 19, and 20 without the attention model nearly doubles RMSE (0.124 to 0.226). But including those three bands *and* the attention model cuts RMSE nearly in half (0.124 to 0.068). These results clearly affirm the importance of attending to the most compelling evidence.

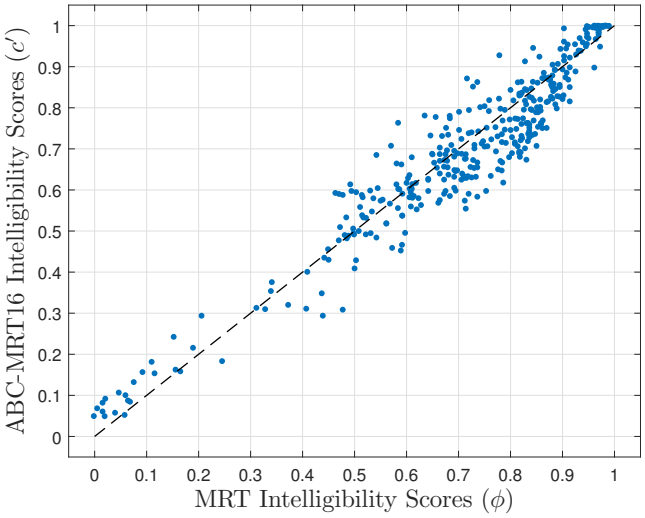

Fig. 1. ABC-MRT16 compared with MRT on 367 conditions (NB, WB, SWB, and FB).

Variant	ρ	RMSE
ABC-MRT16	0.954	0.066
With fit in (11)	0.954	0.065
Without band 21	0.953	0.068
Without attention	0.889	0.241
Without band 21 and without attention	0.889	0.226

Table 2. Agreement with MRT. ABC-MRT16 and four variants across 367 conditions (NB, WB, SWB, and FB).

4. CONCLUSIONS

We suggested an intuitive hypotheses for the human execution of an MRT trial. We cannot unravel the inner workings of the human mind to directly prove or disprove this hypothesis, but we can conclude that the simple attention model that follows from this hypothesis is critical to the remarkable success of ABC-MRT16 across many different types of conditions and bandwidths. In that sense the hypothesis (or at least our mathematical model for it) is consistent with observed data. Because of this, ABC-MRT16 works for all four standard speech bandwidths without any explicit bandwidth switching, just as human listeners do. ABC-MRT16 tools and MRT databases are available at `www.its.bldrdoc.gov/audio`.

This paper was presented at and will appear in the Proceedings of the 42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP2017), New Orleans, LA, March 5-9, 2017.

5. REFERENCES

[1] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, New Jersey: Prentice Hall, 1988.

[2] S. Voran, *Estimation of speech intelligibility and quality* in *Handbook of Signal Processing in Acoustics*. New York: Springer, 2008, vol. 2, ch. 28, pp. 483–520.

[3] P. Loizou, *Speech Enhancement, Theory and Practice*. Boca Raton, Florida: CRC Press, 2013.

[4] A. House, C. Williams, M. Hecker, and K. Kryter, "Articulation-testing methods: Consonantal differentiation with a closed-response set," *J. Acoustical Society of America*, vol. 37, no. 1, pp. 158–166, 1965.

[5] ANSI/ASA "Method for Measuring the Intelligibility of Speech over Communication Systems," S3.2-2009, 2009.

[6] D. Atkinson and A. Catellier, "Intelligibility of selected radio systems in the presence of fireground noise: Test plan and results," NTIA, Tech. Rep. TR-08-453, 2008.

[7] D. Atkinson, S. Voran, and A. Catellier, "Intelligibility of the adaptive multi-rate speech coder in emergency-response environments," NTIA, Tech. Rep. TR-13-493, 2012.

[8] D. Atkinson and A. Catellier, "Intelligibility of analog FM and updated P25 radio systems in the presence of fireground noise: Test plan and results," NTIA, Tech. Rep. TR-13-495, 2013.

[9] S. Voran and A. Catellier, "Speech codec intelligibility testing in support of mission-critical voice applications for LTE," Tech. Rep. TR-15-520, NTIA, 2015.

[10] Y. Teng, "Objective speech intelligibility assessment using speech recognition and bigram statistics with application to low bit-rate codec evaluation," Ph.D. dissertation, University of Wyoming, 2006.

[11] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, 2009.

[12] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech," *J. Acoustical Society of America*, vol. 130, no. 5, pp. 3013–3027, 2011.

[13] G. Yu, A. Brammer, K. Swan, J. Tufts, M. Cherniack, and D. Peterson, "Relationships between the modified rhyme test and objective metrics of speech intelligibility," *J. Acoustical Society of America*, vol. 127, no. 3, p. 1903, 2010.

[14] J. Dreyer, "Binaural index for speech intelligibility via bivariate autoregressive models," Ph.D. dissertation, Michigan Technological University, 2009.

[15] H. Fletcher, *The ASA Edition of Speech and Hearing in Communication*, J. Allen, Ed. Woodbury, New York: Acoustical Society of America, 1995.

[16] S. Voran, "Using articulation index band correlations to objectively estimate speech intelligibility consistent with the modified rhyme test," in *Proc. 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2013.

[17] K. Kondo, "Estimation of Japanese DRT intelligibility using articulation index band correlations," in *Proc. Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, Dec. 2014.

[18] J. Allen, *Articulation and Intelligibility*. Ft. Collins, Colorado: Morgan and Claypool, 2005.

[19] B. Moore, *An Introduction to the Psychology of Hearing*. London: Academic Press, 1992.

[20] S. Voran and A. Catellier, "Crowdsourced speech intelligibility testing that agrees with lab tests and has higher repeatability than lab tests," NTIA, Tech. Rep. (to appear).

This paper was presented at and will appear in the Proceedings of the 42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP2017), New Orleans, LA, March 5-9, 2017.