USING ARTICULATION INDEX BAND CORRELATIONS TO OBJECTIVELY ESTIMATE SPEECH INTELLIGIBILITY CONSISTENT WITH THE MODIFIED RHYME TEST

Stephen Voran

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

ABSTRACT

We present an objective estimator of speech intelligibility that follows the paradigm of the Modified Rhyme Test (MRT). For each input, the estimator uses temporal correlations within articulation index bands to select one of six possible words from a list. The rate of successful word identification becomes the measure of speech intelligibility, as in the MRT. The estimator is called Articulation Band Correlation MRT (ABC-MRT). It consumes a tiny fraction of the resources required by MRT testing. ABC-MRT has been tested on a wide range of impaired speech recordings unseen during development. The resulting Pearson correlations between ABC-MRT and MRT results range from .95 to .99. These values exceed those of the other estimators tested.

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

1. SPEECH INTELLIGIBILITY

Testing the intelligibility of a speech signal is an important and time-honored problem. Numerous techniques have been developed over the years, and these often provide satisfying and repeatable results within specific limited application areas. Overviews of this field can be found in many places including [1]–[3].

1.1. Human Evaluation of Speech Intelligibility

The most direct approach to evaluate speech intelligibility is based on human listening. Carefully prepared speech material is played or read to screened listeners in highly controlled environments. Listeners then respond by answering questions or repeating what was heard. Analysis of those responses leads to conclusions regarding the intelligibility of the speech, within the specific context of the test methodology.

A key factor in these tests is controlling the amount of context available to the listeners. One approach is rhyme testing, and a specific form called the Modified Rhyme Test (MRT) [4], is standardized in [5]. The U.S. National Fire Protection Association specifies the MRT for critical communications testing and our colleagues have completed four large MRTs to support the communications needs of public safety officials, especially firefighters [6]–[8]. Additional details for the four tests are given in Section 3.

MRT speech materials include 50 lists, each containing six English language words with the phonetic pattern consonant-vowel-consonant. The six words differ only in the leading or trailing consonant. A trial consists of the presentation of one word in a carrier phrase (e.g., "Please select the word kit.") The listener then selects what was heard from six options (e.g., "kit," "bit", "fit," "hit," "wit," and "sit") on a graphical interface. The rate of correct word identification leads to a measure of speech intelligibility.

Human speech intelligibility tests can provide useful results if test protocols are fully specified and carefully followed. But these tests take time, they require specialized facilities, and they always include the variabilities inherent in human perception and behavior.

1.2. Objective Estimation of Speech Intelligibility

Signal processing algorithms can be used to analyze speech signals and estimate intelligibility. This approach is fast and perfectly repeatable (objective) but the results are only estimates of what human testing would produce. Seminal work by Harvey Fletcher in the 1920s determined how different frequency bands contribute to speech intelligibility, resulting in the idea of articulation bands and the intelligibility estimator called articulation index (AI) [9]. Many other approaches can be found in [1]–[3],[10],[11]. In [12] existing estimators are successfully tuned to track MRT scores resulting from stationary additive noise and clipping.

Automatic speech recognition (ASR) provides a natural route to objective speech intelligibility estimation. When speech becomes impaired, ASR performance suffers. If ASR errors are consistent with human errors, then ASR performance can serve as a speech intelligibility estimate. In [13] conventional ASR techniques were adapted to successfully approximate intelligibility ratings for a database of five speech coders with ten bit error rates. We have located only one prior attempt to emulate MRTs using ASR [14]. This work addresses additive noise and reverberation. The ASR incorporates multiple bivariate autoregressive models but it falls far short at matching MRT results. The ASR in [14] requires an artificial SNR advantage of 24 to 45 dB in order to match MRT results and thus cannot be used in any practical application.

The set of relevant factors that influence speech intelligibility continues to evolve and objective intelligibility estimation for combinations of these factors remains a challenge. One example is the mobile-to-mobile telecommunications scenario where speech may be impaired by non-stationary noises at the transmit and receive locations, imperfect transducers, noise reduction algorithms, digital coders, and packet losses.

2. ARTICULATION BAND CORRELATION MRT (ABC-MRT)

We have developed a very effective objective speech intelligibility estimator that follows the paradigm of the MRT. The core is a highly specialized ASR algorithm. Many ASR tools are already available and common goals for these are large vocabularies, speaker independence, and robustness to impaired speech. The MRT application is distinctly different: the vocabulary is tiny, speakers are known a

priori, linguistic context is zero, and the word list structure limits the usefulness of the phonemic context. Finally, we are not seeking maximal robustness to impairments, we are seeking a failure characteristic that matches that of humans across the full range from zero intelligibility to perfect intelligibility.

These unique requirements motivated us to design a simple ASR algorithm to emulate the MRT task. It uses basic properties of human audition and speech perception to select one of six words. AI bands provide an organization of the speech spectrum that is highly applicable to speech recognition. Following the insights offered in [15] we use articulation index band temporal correlations to select words, resulting in ABC-MRT.

ABC-MRT creates an AI band based time-frequency (T-F) pattern for an impaired speech signal and then correlates that pattern with the corresponding patterns of the six unimpaired word options to make a selection. ABC-MRT addresses narrowband speech, consistent with the MRTs available to us. The required steps are outlined below.

2.1. AI-based T-F Patterns

Given a sequence of time-domain samples x_t ($f_s = 48$ kHz) the steps for computing the corresponding T-F pattern are as follows. Apply the Hann window to blocks of 512 samples (10.7 ms) and use a 128 sample increment (2.7 ms) between blocks (75% overlap). Compute the DFT of the windowed samples, convert the results to power (exponent 2), then use Stevens' Law [16] to approximate loudness (exponent 0.3). Each result becomes a column in the matrix $\hat{\boldsymbol{X}}$. More formally

$$\hat{x}_{i,k} = \left| \frac{1}{\sqrt{N}} \sum_{t=1}^{N} w_t x_{(k-1)B+t}\, e^{\frac{-j2\pi(i-1)(t-1)}{N}} \right|^{(2\times 0.3)},$$

$$w_t = \sin^2\left(\frac{\pi(t-1)}{N-1}\right),$$

$$N = 512,\ B = 128,\ i = 1 \text{ to } 42,\ k = 1 \text{ to } N_x\ , \qquad (1)$$

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

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

This normalization removes relations between frequency components, but it maintains time-histories for each frequency and is integral to the correlation operations that follow. The resulting matrix $\tilde{\boldsymbol{X}}$ contains $M = 42$ rows covering 0 to 3844 Hz with a resolution of 93.75 Hz and these will be aggregated later to cover 17 AI bands (rows 1, 2, and 3 are unused). $\tilde{\boldsymbol{X}}$ contains N_x columns, each associated with a time increment of 2.7 ms.

ABC-MRT uses all six words from all 50 MRT lists. Each of these 300 words was read (in the carrier phrase) by two female and two male talkers and recorded, resulting in 1200 recordings. For each recording we isolated the MRT keyword, then created and stored a T-F pattern using steps given above.

To apply ABC-MRT to a system-under-test (SUT), pass the 1200 input recordings through the SUT to produce 1200 output recordings. The SUT may introduce delay, so the recording operation must be timed so that each output recording captures at least the entire keyword. Next transform each output recording to a T-F pattern using (1). The normalization in (2) is not required because a local (temporal) normalization is applied later. Each resulting pattern $\hat{\boldsymbol{Y}}$ must be compared with the patterns for six candidate words. Next we present the process for one such comparison.

2.2. Comparing T-F patterns

Let $\tilde{\boldsymbol{X}}$ be a matrix containing an original word T-F pattern and $\hat{\boldsymbol{Y}}$ be a matrix containing a T-F pattern obtained from one 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 $N_x \le N_y$. The first step of the comparison process is to locate the keyword within $\hat{\boldsymbol{Y}}$. Our approach assumes that the SUT delay is approximately constant for the duration of the keyword.

Use articulation bands 3 and 4 (rows 7-9, 505-795 Hz) to locate the keyword. On average, these bands contain greater speech power than other bands, so if we make no assumptions about the noise and distortion produced by the SUT, then these bands are most likely to be useful for locating the keyword. Define $\hat{\boldsymbol{y}}_i(t)$ to be a column vector containing N_x samples from the i^{th} row of $\hat{\boldsymbol{Y}}$:

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

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 $\tilde{\boldsymbol{X}}$. Find the lag t cross-correlation at frequency i:

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

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

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

Once t^* has been determined, calculate correlations for the other frequencies of interest, $i = 4$ to 42, as follows. Use (3) to extract $\hat{\boldsymbol{y}}_i(t^*)$ from $\hat{\boldsymbol{Y}}$, 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 $\rho_i^2(t^*)$. Then accumulate correlation values across AI bands and eliminate any negative results:

$$r_j^2 = \max\left(\sum_{i \in B_j} \rho_i^2(t^*), 0 \right),\ \ j = 1 \text{ to } 17, \qquad (6)$$

where B_j is the set of frequency indices that comprise the j^{th} AI band given in [1]. Due to normalizations, (6) is equivalent to a single cross-correlation for each AI band.

2.3. Word Selection

The T-F pattern $\hat{\boldsymbol{Y}}$ is based on the SUT output. It contains a known keyword taken from a list of six keywords. Thus $\hat{\boldsymbol{Y}}$ must be compared with six T-F patterns $\tilde{\boldsymbol{X}}$ as described in 2.2. The result is 17 values of r_j^2 for each candidate keyword. Introduce the keyword argument $\kappa = 1$ to 6 to indicate which keyword is under consideration. The result of (6) becomes $r_j^2(\kappa)$, $j = 1$ to 17, $\kappa = 1$ to 6.

Next make a word selection based on each of the 17 AI bands. The success rate *across AI bands and lists* leads to the ABC-MRT measure of intelligibility for the SUT. This is in loose analogy to the MRT where the success rate *across subjects and lists* becomes the measure of intelligibility. In each AI band, select the keyword associated with the highest correlation:

$$\hat{w}_j = \arg\max_{\kappa} \left(r_j^2(\kappa) \right),\ j = 1 \text{ to } 17. \tag{7}$$

2.4. Intelligibility Estimate

Compare the keyword selections $\hat{w}_j$ with the known correct keyword w^* associated with $\hat{\boldsymbol{Y}}$:

$$\hat{w}_j = w^* \Rightarrow c_j = 1, \text{ otherwise } c_j = 0,\ j = 1 \text{ to } 17. \tag{8}$$

Average the success flags c_j across the 17 AI bands to produce $\bar{c}$, then average $\bar{c}$ across all 1200 trials to produce $\bar{\bar{c}}$. In the MRT, the intelligibility result is formed from the success rate via an affine transformation that maps $\frac{1}{6}$ (the guessing rate) to 0 and 1 (perfect keyword identification) to 1. Apply that same transformation to $\bar{\bar{c}}$ to produce c':

$$c' = \frac{6}{5}\left(\bar{\bar{c}} - \frac{1}{6} \right). \tag{9}$$

3. RESULTS

We have access to speech files and scores from four MRTs [6]-[8] that were conducted to support the land-mobile radio (LMR) communications needs of public safety officials, especially firefighters. For these tests, MRT input recordings were mixed with high-level background noise recordings (e.g., alarms, saws, pumps, crowds), passed through self-contained breathing apparatus (SCBA) masks and passed through various components of analog and digital LMR systems and proposed future systems. Many different combinations of these factors were tested, and Table 1 provides a high-level summary. The tests cover 139 conditions and 119 of these are unique. Five conditions from Test 2 were repeated in Test 3, 12 conditions from Test 2 were repeated in Test 4, and three conditions from Test 3 were repeated in Test 4.

Subjects performed the MRT tasks in the presence of pink noise (13 to 19 dB below the speech level) to model receive location noise. For consistency, that same noise was added to each recording at the correct level before ABC-MRT processing.

We used only Test 4 to develop ABC-MRT. Tests 1, 2, and 3 were held back as unseen testing data. To best align ABC-MRT results with MRT results from Test 4, use the transformation:

$$\hat{\phi} = \alpha c' + \beta, \text{ with } \alpha = 0.865, \text{ and } \beta = 0.119. \tag{10}$$

The coefficients were selected to minimize the RMS error (RMSE) between the ABC-MRT intelligibility estimate $\hat{\phi}$ and MRT results ϕ using only Test 4. These are the *only optimized coefficients* used in ABC-MRT. Otherwise ABC-MRT is completely motivated by very simple models for the human audition and word-selection tasks in the MRT. Note that (10) reduces large c' values very slightly (1.0 maps 0.984). More significantly, (10) boosts low c' values (0 maps to 0.119). This boosts ABC-MRT word identification performance in difficult conditions so that it better matches the average MRT subject in Test 4.

Test Number	1	2	3	4
Number of Subjects	30	32	20	15
Number of Conditions	30	56	25	28
Analog FM LMR in Hardware	✓			
Analog FM LMR in Software		✓	✓	✓
MBE Speech Coding	✓			
MBE Speech Coding with Noise Reduction		✓	✓	✓
AMR Speech Coding				✓
Impaired Radio Channels		✓	✓	
Amplifier Overload		✓		
SCBA Masks	✓	✓	✓	✓
Quiet Environment	✓	✓	✓	✓
Background Noise	✓	✓	✓	✓
Lowest per-condition MRT result	.00	.33	.53	.02
Highest per-condition MRT result	.89	.91	.92	.84

Table 1: Summary of factors and results for four MRTs. MBE is Multi-Band Excitation and AMR is Adaptive Multi-Rate.

We measure the performance of ABC-MRT by comparing $\hat{\phi}$ with ϕ across the four tests. The Pearson correlation coefficient is a normalized measure of the covariance between $\hat{\phi}$ and ϕ that ranges from -1 to 1. As such it reports how well the *relative* scoring of ABC-MRT and MRT agree. RMSE is an *absolute* measure of agreement that has the same units as ϕ. Results are provided in Table 2.

Test Number	1	2	3	4
Pearson Correlation Coefficient	.985	.947	.965	.950
RMSE	.121	.086	.130	.059

Table 2: Agreement between four MRTs and ABC-MRT.

The correlation values in Table 2 are quite high. Tests 1 and 3 are unseen testing data and have higher correlations than the development data of Test 4. The lowest correlation (Test 2) is only slightly different from the Test 4 correlation. We read this as affirmation that ABC-MRT correlation is not unduly related to the development process or any specific characteristics of Test 4.

On the other hand, RMSE values show a preference for Test 4. This is due to (10) which fits $\hat{\phi}$ to ϕ to minimize RMSE on Test 4 (but has no effect on correlation.) But these RMSE values must be viewed in the proper context. There are 20 conditions that overlap between tests. RMSE (MRT to MRT) for those conditions is 0.115 and RMSE values in Table 2 are never much greater than that baseline value.

The segregation of development data and testing data enforced above is very important to prevent over-fitting and falsely optimistic results. Given the high cost of MRT data, however, we also want to use every available MRT result to offer the research community the most useful tool. Toward that end we also performed the fit in (10) across all four tests. The resulting coefficients are $\alpha = 1.109$ and $\beta = 0.050$. This fit boosts high-end values significantly (0.857 maps to 1.0) and it boosts low-end values to a lesser degree (0 maps to 0.050). Using these values the correlation between $\hat{\phi}$ and ϕ calculated across all four tests is 0.955 and RMSE is 0.073. This result is shown graphically in Fig. 1.

Table 3 reports these results and includes analogous results (allowing a single affine fit to all four MRTs) for seven other estima-

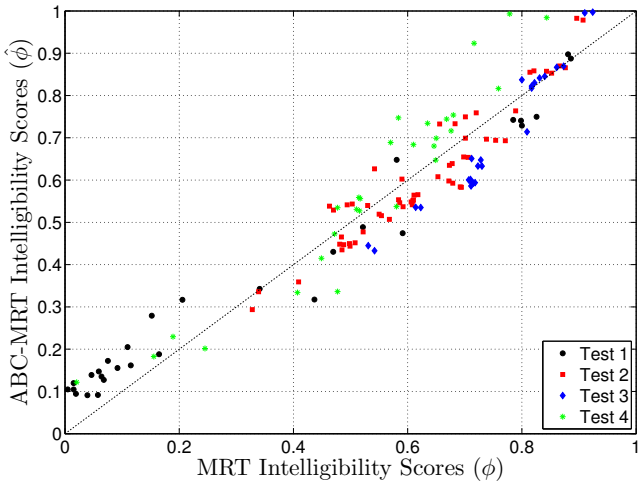

Figure 1: ABC-MRT compared with four MRTs.

tors. ABCa-MRT is similar to ABC-MRT but it uses a detailed auditory model to form T-F patterns. The model includes basilar membrane filtering, rectification, and envelope filtering and was used to produce AIgrams in [17] . ABCa-MRT gives slightly better estimates than ABC-MRT but its computational complexity is more than ten times that of ABC-MRT. PESQ is a very effective speech quality estimator that also shows some applicability for intelligibility estimation. The final five estimators are described in [10] and our implementations were taken from [3]. Each estimator has demonstrated effectiveness in specific application areas but our tests apply them outside those areas. In spite of this, Normalized Covariance Measure shows good results.

Estimator	Correlation	RMSE
ABC-MRT	.955	.073
ABCa-MRT	.963	.066
PESQ	.836	.135
Normalized Covariance Measure	.926	.093
CSII, mid level	.740	.165
I3	.682	.174
modified I3, for sentences	.551	.205
modified I3, for consonants	.742	.165

Table 3: Agreement between MRTs and eight estimators. CSII is the Coherence Speech Intelligibility Index and I3 is a 3-level version of CSII.

ABC-MRT provides good estimates of MRT intelligibility results. We are very encouraged by these first results, especially in light of the simplicity of ABC-MRT, the fact that it uses only two optimized parameter values, and the breadth of the testing to date. But there remain countless additional speech impairment scenarios of interest that should be studied. In addition, the extension of ABC-MRT to wideband speech is straightforward but verification requires wideband MRTs. We encourage other researchers to build on our work. ABC-MRT tools and MRT databases are available at `www.its.bldrdoc.gov/audio`.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

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

[17] M. Régnier and J. Allen, "A method to identify noise-robust perceptual features: Application for consonant /t/," *J. Acoustical Society of America*, vol. 123, no. 5, pp. 2801–2814, 2008.