# The Articulation Index is a Shannon Channel Capacity

Jont B. Allen

ECE Dept. and The Beckman Inst.
University of IL, Urbana IL.

## 1 Introduction

*Articulation Index theory*, created at Western Electric Research Labs by Harvey Fletcher in 1921, is a widely recognized method of characterizing the information bearing frequency regions of speech, [**?** ]. We shall show that the AI [denoted mathematically as $\mathcal{A}(snr)$] is similar to a *channel capacity*, an important concept from Shannon's information theory, defining the maximum amount of information that may be transmitted on a channel without error [**?** ].

The term *articulation*, in this context, is defined as the recognition of nonsense words. *Intelligibility* is defined as the recognition of meaningful words. Bell Labs articulation testing consisted of playing nonsense syllables, composed of 60% CVC, and 20% each of CV and VC sounds. These three types of speech sounds have been shown to compose 76% of all telephone speech [**?** ]. The use of balanced *nonsense* sounds maximizes the *entropy* of the corpus. This was an important methodology, first used around 1910 to control for context effects [**?** ], which were recognized as having a powerful influence on the recognition score. The speech corpus was held constant during these tests, to guarantee that the source entropy was constant. Even thought information theory had not yet been formally proposed, these very basic concepts were clear.

The team consisted of 10 members, with 1 member acting as a *caller*. Three types of linear distortions were used, lowpass filtering, highpass filtering, and a variable *snr*. The sounds were typically varied in level to change the signal to noise ratio *snr*, to simulate the level variations of the telephone network.

The test consisted of the caller repeating context neutral *zero predictability* (ZP) sentences, such as "The first group is *na'v*." and "Can you hear *pōch*." All the initial consonants, vowels, and final consonants were scored, and several statistical measures were computed. For CVCs, the average of the initial $c_i(snr)$ and final $c_f(snr)$ consonant score (each score is the probability correct of identification of the nonsense phone) was computed as $c(snr) = (c_i + c_f)/2$, while the vowel recognition score was $v(snr)$. These numbers characterize the raw data. Next the data is modeled, and a *mean-CVC-syllable* score is computed from the triple product

$$\hat{S}(snr) = cvc. \quad (1)$$

Based on thousands of trials, they found that the *average nonsense phone recognition score*, defined as

$$s \equiv (2c + v)/3, \quad (2)$$

did a good job of representing nonsense CVC syllable recognition, defined as

$$S_3 \equiv s^3 \approx \hat{S}. \quad (3)$$

Similarly, nonsense CV and VC phone recognitions were well represented by

$$S_2 \equiv s^2 \approx (cv + vc)/2. \quad (4)$$

From a great number of measurements it was found that these models did a good job of characterizing the raw data [**?** , Figs. 175, 178, 196-218]. These few simple models worked well over a large range of scores, for both filtering and noise degradations [**?** ].

Note that these formulae only apply to nonsense speech sounds, *not* meaningful words. The exact specifications for the tests to be modeled with these probability equations are discussed in detail in [**?** , Page 259-262]. The above models are necessary but not sufficient to prove that the phones may be modeled as being independent. Namely the above models follow given independence, but demonstrating their validity experimentally does not guarantee independence. To prove independence, all permutations of element *recognition* and *not-recognition* would need to be demonstrated [**?** ].

## 2 Extensions to the frequency domain

Given the success of the average phone score Eq. 2, Fletcher immediately extended the analysis to account for the effects of filtering the speech into bands [**? ?** ]. This method later became known as *articulation index* theory, which many years later developed into the well known ANSI 3.2 AI standard. To describe this theory in full, we need more definitions.

The basic idea was to vary the signal to noise ratio *and* the bandwidth of the speech signal, in an attempt to idealize and simulate a telephone channel. Speech would be passed over this simulated channel, and the phone articulation $s \equiv P_c(\alpha, f_c)$ measured. The parameter $\alpha$ is the gain applied to the speech, used to vary the *snr*. The signal to noise ratio depends on the noise spectral level (the power in a 1 Hz bandwidth, as a function of frequency), and $\alpha$. The consonant and vowel articulation [$c(\alpha)$ and $v(\alpha)$] and $s(\alpha)$ are functions of the speech level. The *mean phone articulation error* is $e(\alpha) = 1 - s(\alpha)$.

The speech was filtered by complementary lowpass and highpass filters, having a cutoff frequency of $f_c$ Hz. The articulation for the low band is $s_L(\alpha, f_c)$, while for the high band is $s_H(\alpha, f_c)$. The nonsense syllable, word, and sentence intelligibility are $S(\alpha)$, $W(\alpha)$ and $I(\alpha)$, respectively.

*Formulation of the AI.* Once the functions $s(\alpha)$, $s_{_L}(\alpha, f_c)$ and $s_{_H}(\alpha, f_c)$ are known, it is possible to find relations between them. These relations, first derived by Fletcher in 1921, were first published by French and Steinberg (1947).

The key insight Fletcher had was to find a linearizing transformation of the results. Given the wideband articulation $s(\alpha)$, and the banded articulations $s_{_L}(\alpha, f_c)$ and $s_{_H}(\alpha, f_c)$, he sought a nonlinear transformation of probability $\mathcal{A}$, now called the *articulation index*, which would render the articulations additive, namely

$$\mathcal{A}(s) = \mathcal{A}(s_{_L}) + \mathcal{A}(s_{_H}). \tag{5}$$

This formulation payed off handsomely.

The function $\mathcal{A}(s)$ was determined empirically. It was found that the data for the nonsense sounds closely follows the relationship

$$\log(1 - s) = \log(1 - s_{_L}) + \log(1 - s_{_H}), \tag{6}$$

or in terms of error probabilities

$$e = e_{_L} e_{_H}, \tag{7}$$

where $e = 1 - s$, $e_{_L} = 1 - e_{_L}$ and $e_{_H} = 1 - s_{_H}$. These findings require $\mathcal{A}(s)$ of the form

$$\mathcal{A}(s) = \frac{\log(1 - s)}{\log(e_{min})}. \tag{8}$$

This normalization parameter $e_{min} = 1 - s_{max}$ is the minimum error, while $s_{max}$ is the maximum value of $s$, given ideal conditions (i.e., no noise and full speech bandwidth). For much of the the Bell Labs work $s_{max} = 0.986$ (i.e., 98.6% was the maximum articulation), corresponding to $e_{min} = 0.015$ (i.e., 1.5% was the minimum articulation error) [**?** , MM-3373, Sept. 14, 1931, J.C. Steinberg], [**?** , Page 281] and Galt's notebooks [**?** ].

Fletcher's simple two-band example illustrates Eq. 7: If we have 100 spoken sounds, and 10 errors are made while listening to the low band, and 20 errors are made while listening to the high band, then

$$e = 0.1 \times 0.2 = 0.02, \tag{9}$$

namely two errors will be made when listening to the full band. Thus the wideband articulation is 98% since $s = 1 - 0.02 = 0.98$, and the wideband nonsense CVC syllable error would be $S = s^3 = 0.941$.

In 1921 Fletcher, based on results of J.Q. Stewart, generalized the two-band case to $K = 20$ bands:

$$e = e_1 e_2 \cdots e_k \cdots e_K, \tag{10}$$

where $e = 1 - s$ is the wideband average error and $e_k \equiv 1 - s_k$ is the average error in one of K bands. Formula 10 is the basis of the *articulation index*. The $K$ band case has never been formally tested, but was verified by working out many examples.

The number of bands $K = 20$ was an empirically choice that was determined after many years of experimental testing. The number 20 was a compromise that probably depended on the computation cost as much as anything. Since there were no computers, too many bands was prohibitive with respect to computation. Fewer bands were insufficiently accurate.

Each of the bands was chosen to have an equal contribution to the articulation (This represents a maximum entropy partition). Eventually they found that articulation bands, defined as having equal articulation, were proportional to cochlear critical bands. Each of the $K = 20$ articulation bands corresponds to approximately 1 mm along the basilar membrane [**?** ]. When the articulation is normalized by the critical ratio, as a function of the cochlear tonotopic axis, it was found that the articulation density per critical band, is constant [**? ?** ]. This property depends critically on the initial maximum entropy distribution of sounds used in the testing.

## 3 French and Steinberg (1947)

In 1947 French and Steinberg provided an important extension of the formula for the band errors by relating $e_k$ (the $k^{th}$ band probability of error) to the band signal to noise ratio $\mathit{SNR}_k$ (in dB), by the relation

$$e_k = e_{min}^{\mathit{SNR}_k/K}, \tag{11}$$

which is the same as Eq. (10a) of the 1947 French and Steinberg paper, where $\mathit{SNR}_k$ is the normalized signal to noise ratio, defined next.

In each articulation band the signal and noise power is measured, and the long term ratio is computed as

$$\mathit{snr}_k \equiv \frac{1}{\sigma_n(\omega_k)} \left[ \frac{1}{T} \sum_{t=1}^{T} \sigma_s^2(\omega_k, t) \right]^{1/2}, \tag{12}$$

where $\sigma_s(\omega_k, t)$ is the short-term RMS of a speech frame and $\sigma_n(\omega_k)$ is the noise RMS, at frequency band $k$. The time duration of the frame impacts the definition of the *snr*, and this parameter must be chosen to be consistent with a cochlear analysis of the speech signal. It seems that the best way to established this critical duration is to use a cochlear filter bank, which is presently an uncertain quantity of human hearing [**? ?** ]. The standard method for calculating a perceptually relevant signal to noise ratio was specified in 1940 [**?** ].

Each band $\mathit{snr}_k$ is converted to dB, and then limited and normalized to a range of 0 to 30, defined as

$$\mathit{SNR}_k \equiv \begin{cases} 0 & 20\log_{10}(\mathit{snr}_k) < 0 \\ 20\log_{10}(\mathit{snr}_k)/30 & 0 < 20\log_{10}(\mathit{snr}_k) < 30 \\ 1 & 30 < 20\log_{10}(\mathit{snr}_k). \end{cases} \tag{13}$$

The factor 30 comes from the fact that speech has a 30 dB dynamic range in a given articulation band [**?** , Fig. 4, page 95].

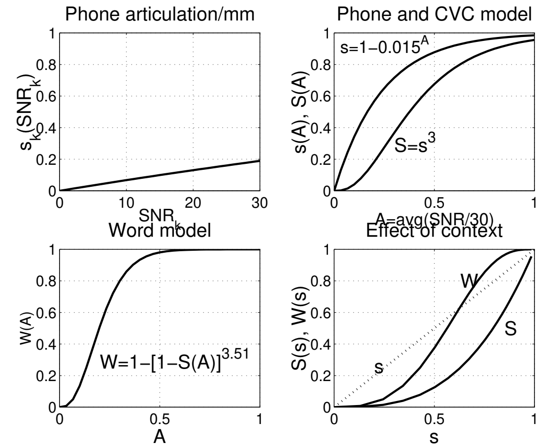


**Fig. 1.** *Typical results for the French and Steinberg AI model, as defined by in [**?** ].*

The basic idea of this formula is that when the $snr_k$ is less than 0 dB within each cochlear critical band, the speech is undetectable. When $snr_k$ is greater than 30 dB, the noise has no effect. Between 0 and 30 dB $SNR_k$ is proportional to $\log(snr_k)$.

Merging the formula for the total error Eq. 10 with that for the band errors $SNR_k$ Eq. 13, the total error is related to the average SNR

$$\mathcal{A} \equiv \overline{SNR} = \frac{1}{K} \sum_k SNR_k \tag{14}$$

since

$$e = e_1 e_2 \cdots e_K = e_{min}^{\overline{SNR}} = e_{min}^{\mathcal{A}}. \tag{15}$$

The final articulation index formula, relating the articulation $s = 1 - e$ to the articulation index $\mathcal{A} \equiv \overline{SNR}$, is therefore

$$s = 1 - e_{min}^{\mathcal{A}}. \tag{16}$$

Note that as $snr_k \rightarrow 30$ dB in every band, $\mathcal{A} \rightarrow 1$ and $s \rightarrow s_{max}$. When $snr_k \rightarrow 0$ dB in all the bands, $\mathcal{A} \rightarrow 0$ and $s \rightarrow 0$. This formula for $s(\mathcal{A})$ has been verified many times, for a wide variety of conditions [**?** ]. However it is not perfect [**?** ]. Figure 1 shows typical results of articulations in a band [$s_k(SNR_k)$], for phones [$s(\mathcal{A})$], CVCs [$S(\mathcal{A})$], words [$W(\mathcal{A})$], and the effects of two types of context. For details, see [**? ?** ].

### 3.1 The AI and the Channel Capacity

It is interesting that this band average is taken over the dB values $\sum_k SNR_k$ rather than the linear values $\sum_k snr_k$. This is a subtle and significant fact that has been

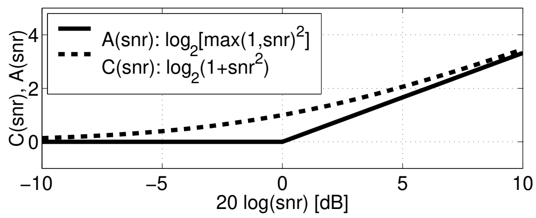


**Fig. 2.** *Plot of* $\log(1 + snr^2)$ *and* $\log[\max(1, snr)^2]$ *versus SNR=* $20 * log(snr)$.

overlooked in discussions of the AI. The average over $SNR_k$, which are in log units, is proportional to the log of the *geometric mean* of $snr_k$, namely

$$\mathcal{A} \equiv \frac{1}{K}\sum_k SNR_k \propto \log\left(\prod_k snr_k\right)^{1/K}. \tag{17}$$

The geometric mean of the *snr* is used in information theory as a measure an abstract volume, representing the amount of information that can be transmitted by a channel. For example, if the integral is replaced by a finite sum, then the Shannon Gaussian channel capacity formula

$$C = \int_{-\infty}^{\infty} \log_2[1 + snr^2(f)]df, \tag{18}$$

which is a measure of a Gaussian channel's maximum capacity for carrying information, is very similar to Eq. 14. From Fig. 2, we see that $\mathcal{A}(snr)$ is a straight–line approximation to to the Shannon channel capacity formula $C(snr)$. The figure shows the two functions $C(snr) \equiv \log_2[1 + snr^2]$ and $A(snr) \equiv \log_2[\max(1, snr)^2]$, which is $\propto \mathcal{A}(snr)$.

The early idea of a channel capacity, as proposed by R. V. L. Hartley, was to count the number of intensity levels in units of noise variance [**? ?** ]. This is a concept related to counting JNDs. It is interesting and relevant that Hartley, a Rhodes scholar well versed in psychophysical concepts, also proposed the decibel, which was also based on the intensity JND [**? ?** ]. The expression

$$\log(1 + snr^2) = \log\left(\frac{I + \Delta I}{I}\right) \approx \frac{\Delta I}{I}, \tag{19}$$

(the approximation holding when the ratio $\Delta I/I$ is small) where $\Delta I$ and $I$ are the JND and intensity respectively, is closely related to counting JNDs. It has been shown, by George A. Miller [**?** ], that noise is close to the first JND level if its presence changes the input stimulus by 1 dB, that is when $10\log_{10}(1 + \Delta I/I) = 1$, or $\Delta I/I = 1/10$. Hence, the function $\log_2(1 + snr^2)$ is related to the number of JNDs, in bits [**? ? ?** ]. The product of the number of articulation bands times the number of JNDs determines a volume, just as the channel capacity determines a volume.

## References


[] J. B. Allen. How do humans process and recognize speech? *IEEE Transactions on speech and audio*, 2(4):567–577, October 1994.

[] J. B. Allen. Harvey Fletcher's role in the creation of communication acoustics. *J. Acoust. Soc. Am.*, 99(4):1825–1839, April 1996.

[] J. B. Allen and S. T. Neely. Modeling the relation between the intensity JND and loudness for pure tones and wide–band noise. *J. Acoust. Soc. Am.*, 102(6):3628–3646, December 1997.

[] J.B. Allen. Articulation and intelligibility. In B. H. Juang, editor, *Lectures in Speech and Audio Processing*, chapter IV, page xxx. Morgan and Claypool, 3401 Buckskin Trail, LaPorte, CO 80535, 2004. To appear.

[] A. W. Bronkhorst, A. J. Bosman, and G. F. Smoorenburg. A model for context effects in speech recognition. *J. Acoust. Soc. Am.*, 93(1):499–509, January 1993.

[] G. A. Campbell. Telephonic intelligibility. *Phil. Mag.*, 19(6):152–9, January 1910.

[] H. K. Dunn and S. D. White. Statistical measurements on conversational speech. *Journal of the Acoustical Society of America*, 11:278–288, 1940.

[] Harvey Fletcher. An empirical theory of telephone quality. *AT&T Internal Memorandum*, 101(6), October 1921. First document about the Articulation Index. Case 211031, report 21839. The only know copy is in Fletcher's archives, Brigham Young University, Provo UT. (UA029, Box 19 #2, MSS1233) [**?** ].

[] Harvey Fletcher. *Speech and Hearing*. D. Van Nostrand Company, Inc., New York, 1929.

[] Harvey Fletcher. Speech and hearing in communication. In Jont B. Allen, editor, *The ASA edition of Speech and Hearing in Communication*. Acoustical Society of America, New York, 1995.

[] Harvey Fletcher and R.H. Galt. Perception of speech and its relation to telephony. *J. Acoust. Soc. Am.*, 22:89–151, March 1950.

[] N. R. French and J. C. Steinberg. Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.*, 19:90–119, 1947.

[] R.V.L Hartley. The function of phase difference in the binaural location of pure tones. *Phy. Rev.*, 13:373–385, June 1919.

[] R.V.L. Hartley. Transmission of information. *Bell System Tech. Jol.*, 3(7):535–563, July 1928.

[] R.V.L. Hartley. "TU" becomes "DECIBEL". *Telephone Engineering*, 33:40, January 1929.

[] G. A. Miller. Sensitivity to changes in the intensity of white noise and its relation to masking and loudness. *J. Acoust. Soc. Am.*, 19:609–619, 1947.

[] C. M. Rankovic. Articulation index predictions for hearing-impaired listeners with and without cochlear dead regions (l). *J. Acoust. Soc. Am.*, 111(6):2545–2548, June 2002.

[] C.M. Rankovic and J.B. Allen. *Study of Speech and hearing at Bell Telephone Laboratories: The Fletcher Years;* CDROM *containing Correspondence Files (1917–1933), Internal reports and several of the many Lab Notebooks of R. Galt*. Acoustical Society of America, Suite 1N01, 2 Huntington Quadrangle, Melville, New York, 2000.

[] C. E. Shannon. The mathematical theory of communication. *Bell System Tech. Jol.*, 27:379–423 (parts I, II), 623–656 (part III), 1948.

[] C. A. Shera, J. J. Guinan, and A. J. Oxenham. Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proc. Natl. Acad. Sci. USA*, 99:3318–2232, 2002.

[] J. M. Wozencraft and I. M. Jacobs. *Principles of Communication Engineering*. John Wiley, New York, 1965.