The articulation index is a Shannon channel-capacity

Jont B. Allen

ECE Dept. and The Beckman Inst.

University of IL, Urbana IL.

February 21, 2004

Running Title: ”Articulation Index and Channel Capacity”

Abstract

The articulation index (AI), denoted $\mathcal{A}(snr)$, first described by Fletcher in 1921 (Fletcher 1995), has served the speech and hearing community well, as a well known method for characterizing nonsense speech sound intelligibility. For example, the AI is a basic tool of hearing aid research, and has even been used to fit hearing aids. It is defined in terms of an average of a special function of the signal noise ratio, in frequency bands. Fletcher's formulation was extended by French and Steinberg (1947) when they established an explicit formula for the error probability in frequency bands, in terms of the band signal to noise ratio, expressed in dB. Using classic formulas from the literature, this article demonstrates that the French and Steinberg formula for $\mathcal{A}(snr)$ is essentially Shannon's *Channel capacity* $\mathcal{C}(snr)$ formula for the Gaussian channel. The similarity of the two relations justifies characterizing the AI as an *information theory* relationship. This new insight is also useful in that it helps limit the *misuse* of the AI, in situations where it may not be applicable. Such misuse includes the case of meaningful speech sounds, where the entropy is reduced by language (context) effects.

PACS: 43.71.Gv, 43.72.-p

1 Introduction

Articulation Index (AI) theory, created at Western Electric Research Labs in 1921 by Harvey Fletcher, in response to the need to characterize speech over a telephone channel, is a widely recognized method of characterizing the information bearing aspects of speech, in terms of a frequency dependent average signal to noise ratio (Fletcher 1921; Fletcher and Galt 1950; Fletcher 1995; Allen 1996). We shall show that the AI, denoted mathematically as $\mathcal{A}(snr)$, is similar to a *channel capacity*, defined as the maximum information rate that may be transmitted over a channel without error (Shannon 1948).[1]

In the following, the term *recognition* is used to mean the probability of correct identification of speech sounds, such as consonants C and vowels V. The term *articulation* is defined as the *recognition of nonsense words*. *Intelligibility* is defined as the *recognition of meaningful words*. The early pre-Bell Labs AT&T articulation tests consisted of listening to and scoring nonsense syllables, having an *a priori* distribution of 60% consonant-vowel-consonant (CVC), and 20% each of consonant-vowel (CV) and vowel-consonant (VC) sounds (Fletcher 1922; Fletcher and Steinberg 1930). These three types of speech sounds have been shown to compose 76% of all telephone speech (Fletcher 1995). This testing protocol was first used at AT&T, circa 1910, to control for speech context effects (Campbell 1910), where CV confusion matrices were first utilized to analyze speech recognition scores. The articulation may be computed from the mean of the diagonal of the confusion matrix C_{ij} between a significant group of phones in a language, namely $\sum_i C_{ii}/N$ where N is the total number of stimuli. These very basic concepts were critical to these 1910-1924 studies, well before the ideas of information theory had been formulated.

The AI formalism is based on finding a *total probability of error* e of identifying open-set nonsense speech sounds, as a function of frequency band f_k, indexed by integer $k = 1, \cdots, K$, $K = 20$, and computing the resulting frequency density function of $e_k \equiv e(f_k)$. Each band error is characterized by a band error probability $e_k(snr)$. The bandwidths are chosen to give equal average probability of error ($e_k = e_j$ for

[1] We shall use *snr* to denote the RMS voltage ratio of the signal and the noise, and *SNR* to denote *snr* expressed in dB, namely $SNR \equiv 20\log_{10}(snr)$.

any k and j), when averaged across a large speech corpus. Fletcher assumed the band errors to be independent, and using this assumption, it was found that the total probability of phone (speech sound) identification error $e(\mathit{snr})$ was equal to the product of band errors, namely

$$e(\mathit{snr}) = e_1(\mathit{snr}_1)e_2(\mathit{snr}_2)\cdots e_K(\mathit{snr}_K).$$

This measure, of the total error, was shown to be highly accurate, thereby justifying the assumption of independence. The accuracy of this formulation has been verified many times, and in many languages.

Fletcher's formulation was extended by French and Steinberg (1947) when they established an explicit formula for $e_k(\mathit{snr})$, in terms of the signal to noise ratio SNR_k, expressed in dB. This formula was based on the audible fluctuations of speech within a band and lead to a simple formula for the AI as the average SNR, averaged over the band SNRs, in dB, namely

$$\mathcal{A}(\mathit{snr}) = \sum \mathit{SNR}_k.$$

Using classic formulas from the literature, this article demonstrates that the French and Steinberg formula for $\mathcal{A}(\mathit{snr})$ is essentially Shannon's *Channel capacity* $\mathcal{C}(\mathit{snr})$ formula for the Gaussian channel. The similarity of the two relations justifies characterizing the AI as an *information theory* relationship. Like the channel capacity, the AI may be viewed as a volume, that describes the information rate (the number of bits/sec) that may be transmitted without error. This characterization should allow for an improved understanding of the AI measure, as the number of bits per second, of nonsense speech information, that may be transmitted over an auditory communication channel. This new insight is also useful in that it helps limit the *misuse* of the AI, in situations where it may not be applicable. Such misuse includes the case of meaningful speech sounds, where the entropy is greatly reduced, by context effects.

2 Modeling nonsense syllables

Listening teams typically 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 channel.

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 $c_i(\mathit{snr})$ and final $c_f(\mathit{snr})$ consonant score (each score is the probability correct of identification of that phone) was computed as $c(\mathit{snr}) = (c_i + c_f)/2$, while the vowel recognition score was $v(\mathit{snr})$. These numbers characterize the raw data. Next the data is modeled, and a *mean-CVC-syllable* score is computed from the triple product

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

Based on many thousands of trials, Fletcher found that the *average phone recognition score for nonsense syllables*, defined as

$$s \equiv (2c + v)/3, \tag{2}$$

did an excellent job of representing nonsense CVC syllable recognition, defined as

$$S_3 \equiv s^3 \approx \hat{S}. \tag{3}$$

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

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

These models fit the raw data with little error (Fletcher 1995, Figs. 175, 178, 196-218), and worked well over a large range of scores, for both filtering and noise degradations (Boothroyd and Nittrouer 1988; Rankovic 2002).[2]

The exact specifications for the tests to be modeled with these probability equations are discussed in detail in (Fletcher 1929, 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 from an independence assumption, but

[2]These formulae only apply to nonsense speech sounds, *not* meaningful words. The extension to meaningful sounds (cat, hat) has been studied by Boothroyd (Boothroyd 1968; Boothroyd and Nittrouer 1988), and more recently by Bronkhorst (Bronkhorst et al. 1993; Bronkhorst et al. 2002).

demonstrating their validity experimentally does not prove independence. To prove independence, all permutations of element *recognition* and *not-recognition* would need to be demonstrated (Bronkhorst et al. 1993).

2.1 Diphones as a speech unit:

There is a glaring problem, so obvious, that it requires some discussion. In the introduction the development was based on two probability measures given by Eq. 2 and Eq. 3. The only way that both of these equations can be simultaneously true is if

$$\lambda \equiv v/c = 1 \tag{5}$$

Fletcher did some calculations to try to estimate the value of λ, which appears to vary between 1 and 1.4, in an attempt to justify the seeming conflict. However the issue has remained largely unexplored until this day. Some speculations are in order.

Many believe that λ should be very large, reflecting the much greater energy in the vowel. If the vowel and consonant probability are correlated (e.g., if λ is independent of the *snr*), then the information in speech could well be the transition rather than the consonant and vowel, as is more commonly believed. Since the number of transitions within a /CVC/, given a fixed number of consonant and vowels units, is equal to the number of states of a /CVC/, either could be used to code the information. For example, if there were 2 C units and 3 V units in a CV, then there are 2x3=6 possible CV units. The number of transitions on the other-hand is also 2x3=6 (for each of 2 starting points, there are 3 possible outcomes). If the transitions carried the critical information, then λ could be 1, as the consonant and the vowel would share the energy on each side. If this conjecture were correct, then what is traditionally viewed as coarticulation, would actually be the information bearing signal.

3 Extensions to the frequency domain

Given the success of the average phone score Eq. 2, Fletcher extended the analysis to account for the effects of filtering the speech into bands (Fletcher 1921; Fletcher 1929). This method is now known as *articulation index theory*. A highly simplified

version of this theory is defined in the well known ANSI 3.2 AI standard. To describe this theory in more detail, we need additional definitions provided in Table 1.

Fletcher's basic idea was to vary both the signal to noise ratio *and* the bandwidth of the speech signal, in an attempt to idealize and simulate a telephone channel. Speech was passed over this simulated channel, and the *average phone articulation* $s \equiv P_c(\alpha, f_c)$ was measured. The parameter α, the gain applied to the speech, was 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 α. For the wideband case, the consonant and vowel articulations $c(\alpha)$ and $v(\alpha)$, and thus $s(\alpha)$, are functions of the speech level α. The *average phone articulation error* is $e(\alpha) = 1 - s(\alpha)$.

The speech was filtered by complementary lowpass and highpass filters, having a cutoff frequency proportional to the 3 dB crossover frequency f_c Hz. The low frequency filter cutoff was typically $f_c/1.2$ and the high frequency cutoff is $1.2f_c$. Because the skirts of the filters were very steep, the crossover frequency of the filters were more than 30 dB below the 3dB frequency f_c. This is an important detail that is commonly not appreciated.

The articulation for the low band is defined as $s_{_L}(\alpha, f_c)$, and $s_{_H}(\alpha, f_c)$ for the high band. The nonsense syllable articulation, and word and sentence intelligibility, are defined as $S(\alpha)$, $W(\alpha)$ 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 detailed relations were first published by French and Steinberg (1947), but first derived by Fletcher in 1921.

The key Fletcher insight 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 $s_{_H}(\alpha, f_c)$ for nonsense speech sounds, he sought a nonlinear transformation of probability $\mathcal{A}$, now called the *articulation index*, which would render the articulations additive, namely

$$\mathcal{A}(s) = \mathcal{A}(s_{_L}) + \mathcal{A}(s_{_H}). \tag{6}$$

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

$$\log(1-s) = \log(1-s_{_L}) + \log(1-s_{_H}), \tag{7}$$

or in terms of error probabilities

$$e = e_{_L} e_{_H}, \tag{8}$$

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

$$\mathcal{A}(s) = \frac{\log(1-s)}{\log(e_{min})}. \tag{9}$$

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 most of the AT&T measurements $s_{max} = 0.986$ (i.e., 98.6% was the maximum articulation), corresponding to $e_{min} = 0.015$ (i.e., 1.5% was the minimum articulation error) (Rankovic and Allen 2000, MM-3373, Sept. 14, 1931, J.C. Steinberg), (Fletcher 1995, Page 281) and Galt's notebooks (Rankovic and Allen 2000).

Fletcher's simple two-band example illustrates Eq. 8: 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

$$e = 0.1 \times 0.2 = 0.02, \tag{10}$$

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

$$e = e_1 e_2 \cdots e_k \cdots e_K, \tag{11}$$

where $e = 1 - s$ is the wideband average articulation error and $e_k \equiv 1 - s_k$ is the average articulation error in each of the K bands. Formula 11 is the basis of the *articulation index*. The K band case has never been formally or directly tested, but was verified by working out many examples. The number K =20 was a compromise that may have depended on both computational cost and theoretical considerations. Fewer bands were insufficiently accurate. Since there were no computers, more than

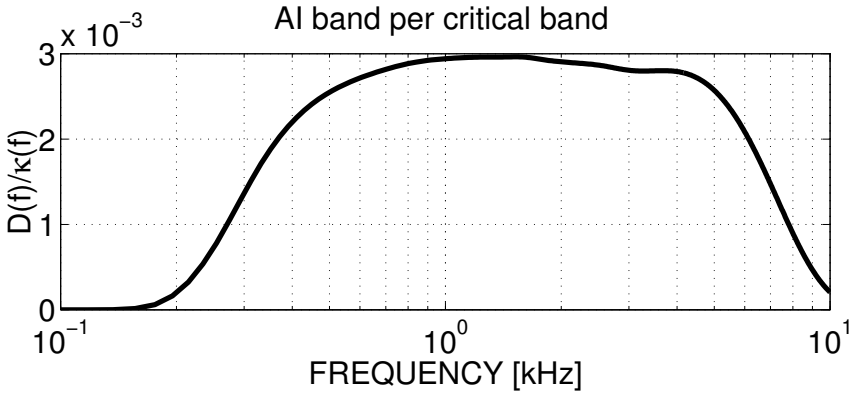

Figure 1: *This figure shows the ratio of the articulation index density (also called the* speech importance function*), and the* critical bandwidth *(also called the* equivalent rectangular bandwidth *or ERB), which is a measure of the cochlear filter bandwidth. The critical bandwidth was derived from the ratio of the RMS level of a tone, adjusted to its detection threshold level, to the spectral level of a noise. Note that ratio has units of bandwidth. From the figure we conclude that the information density of speech used in the AT&T tests, per cochlear critical band, is approximately uniform.*

20 bands was prohibitive with respect to computation. More important perhaps, more bands would result in the bands being unrealisticly narrow. Each of $K = 20$ articulation bands corresponds to approximately 1 mm along the basilar membrane (Fletcher 1995), resulting in each articulation band corresponding to about 2 cochlear critical bands, which Fletcher estimated as being about 0.5 mm.

The details of the AI calculations are outlined in the classic 1947 French and Steinberg paper. Each of the K bands was chosen to have an equal contribution to the articulation (This represents a maximum entropy partitioning). 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 roughly constant (Fletcher 1948; Fletcher 1950; Allen 1994; Allen 1996), as shown in Fig. 1.[3] This figure was calculated using the data of Table 63 on page 333 of Fletcher 1953 book (Fletcher 1995), and dividing it by the ERB estimate (Figure 121 of 1953 book) given in (Fletcher 1938a; Fletcher 1938b). The raw data of Table 63 were smoothed by a polynomial interpolation of the log of the tabulated data. The critical ratio data are given in table 16, page 101, and were interpolated using splines to the same frequency base. The figure was

[3]This result was discovered by Galt. He tried several times to publish his observation, but the paper was rejected by JASA. This was a source of considerable anguish to Galt, as may be observed in his many notebooks (Rankovic and Allen 2000).

computed from the ratio of the frequency density of articulation per critical bandwidth, and the final curve was normalized to have unit area. The results of Fig. 1 depends critically on the natural distribution of sounds used in the speech tests. When biased (unnatural and non-typical) initial distributions are chosen, the importance function (articulation density over frequency) will be different (Duggirala et al. 1988).

3.1 French and Steinberg (1947)

In 1947 French and Steinberg derived an expression relating Fletcher's band error e_k (the k^{th} band probability of error) to the band signal to noise ratio SNR_k, in dB. Steinberg had worked closely with Fletcher (Fletcher and Steinberg 1930) on the articulation index in the late 1920's. It is therefore not surprising that some 17 year later (i.e., in 1947), a deep insight for e_k had evolved. They understood that the speech information was contained in the natural level fluctuations in speech energy, measured in half-octave bands, in 1/8 second intervals. Dunn and White (Dunn and White 1940) had shown that speech energy, in bands, was linear on a log scale over a 30 dB range of intensity. French and Steinberg state their critical assumption as follows:[4]

> When speech, which is constantly fluctuating in intensity, is reproduced at a sufficiently low level, only the occasional portions of highest intensity will be heard, but if the level of reproduction is raised sufficiently even the portions of lowest intensity will become audible. Thus the similarity in slope of the straight line portions of the W curves and the speech distribution curve suggests that W is equal to the fraction of the intervals of speech in a band that can be heard.

The phrase "sufficiently low level" could be exchanged with the phrase "sufficiently high noise," and retain the same meaning. The variable W is the contribution of the AI in a band.

Definition of *SNR* The standard method for calculating a perceptually relevant signal to noise ratio was specified in 1940 (Dunn and White 1940). In each articulation

[4]Page 106 (French and Steinberg 1947).

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 should be chosen to be consistent with a cochlear analysis of the speech signal.

To characterize the observed fluctuations, each band snr_k was 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 justification of this formula is that when the snr_k is less than 1 within each cochlear critical band, the speech is undetectable, and when snr_k is greater than 30, the noise has no affect.

The factor 30 comes from the fact that speech has a 30 dB dynamic range in a given articulation band (French and Steinberg 1947, Fig. 4, page 95). Between 0 and 30 dB, SNR_k is proportional to $\log(\mathit{snr}_k)$, and thus may be normalized to vary lineally between zero and one.

In the terminology of the present paper, the final formula for the band error is[5]

$$e_k = e_{min}^{\mathit{SNR}_k/K}. \tag{14}$$

Merging Eq. 11 and Eq. 14

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

provides a formula for the total error in terms of the average SNR, defined as

$$\mathcal{A} \equiv \overline{\mathit{SNR}} = \frac{1}{K} \sum_k \mathit{SNR}_k. \tag{16}$$

[5]Both Fletcher, and French and Steinberg, worked almost exclusively in dB units, which were more convenient in those times. Now that the computer is available to compute these formulas, it seems better to rework the equations in a more meaningful notation.

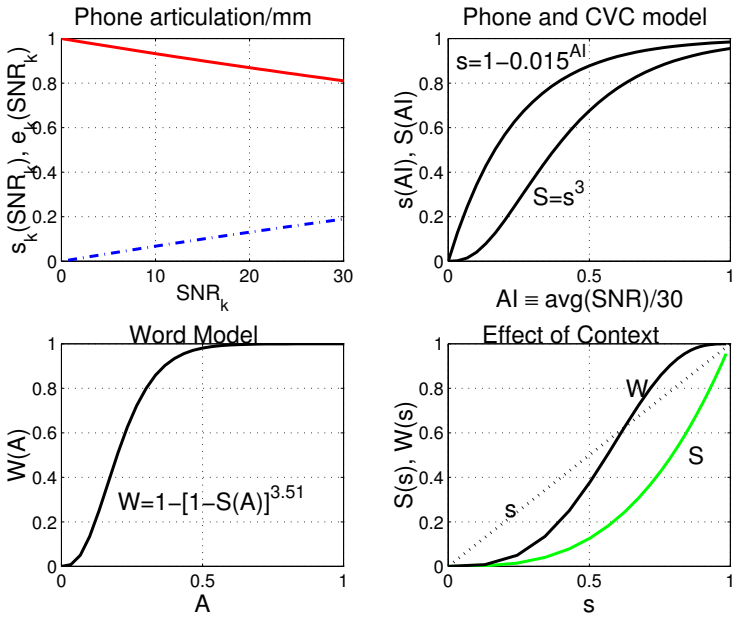

Figure 2: This figure shows a typical set of results for the French and Steinberg AI model, as defined in (Allen 1994). See the text for a detailed description of each panel.

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

$$s = 1 - e_{min}^{\mathcal{A}}. \tag{17}$$

Note that as $snr_k \to 30$ dB in every band, $\mathcal{A} \to 1$ and $s \to s_{max}$. When $snr_k \to 0$ dB in all the bands, $\mathcal{A} \to 0$ and $s \to 0$. This formula for $s(\mathcal{A})$ has been verified many times, for a wide variety of conditions. However it is not perfect (Allen 2004).

Figure 2 shows typical results of articulations in a band [$s_k(SNR_k)$], for phones [$s(\mathcal{A})$], CVCs [$S(\mathcal{A})$], high context words [$W(\mathcal{A})$] (Boothroyd and Nittrouer 1988, Fig. 7) with $j = 3.51$, and the effects of two types of context (Allen 1996; Allen 2004). The upper left panel shows $s_k(SNR_k)$ (dashed curve) and $e_k(SNR_k)$ (solid curve) for band k. As SNR_k varies from 0 to 30 dB, the band articulation goes from 0 to just under 20%, corresponding to an error between 1 and 80%. The product of 20 such bands, when subtracted from 1, gives the average wideband articulation s, as shown in the upper right panel. The CVC syllable error is then $S = s^3$. Since $s < 1$, the cube must be less than s, namely $s^3 < s$. We conclude that the human speech code is an example of a spread-spectrum channel, with about 4.3 bits/phone, with about 5 phones/s (e.g., ≈21.5 bits/sec), spread over about 7 kHz of bandwidth. The channel capacity is about 70 kB/s (7 kHz and snr^2=1000).

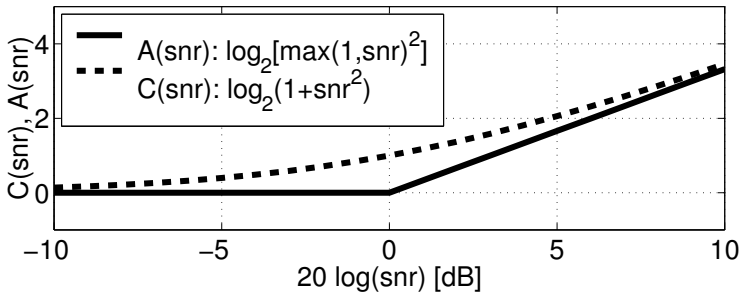

Figure 3: *Plot of* $\log(1 + snr^2)$ *and* $\log[\max(1, snr)^2]$ *versus SNR=* $20 * \log(snr)$.

In the lower left panel, a typical example for a meaningful word score is provided. Boothroyd has shown that meaningful word scores W may be related to the nonsense syllable score S with a "degree-of-freedom" model of the form

$$W(S) \equiv 1 - [1 - S(\mathcal{A})]^j.$$

where j is a number greater than 1. In the example of Fig. 2, $j = 3.51$. Using this example, the above relation may be written as following

$$1 - W(S) \equiv [1 - S(\mathcal{A})][1 - c_x]^{j-1},$$

where c_x is the probability correct due to context alone (Boothroyd and Nittrouer 1988). The value of $j - 1$ characterizes context effect of real words, namely it depends on the occupancy of words in nonsense CVC syllable space. One may view such word models as being similar to error correcting codes, where a nearest neighbor is taken as the best choice when an error is heard. The precise details are unclear about how this might function, however the work of Bronkhorst has enlightened us considerably on this question (Bronkhorst et al. 1993). In the lower right panel, the model relations between $S(s)$ and $W(s)$ are shown graphically.

4 The AI and the Channel Capacity

When computing the AI measure, it is important to note that this band average is taken over dB values ($\sum_k SNR_k$) rather than the linear values $\sum_k snr_k$. This is a subtle and significant observation that has been overlooked in previous 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

$$\mathcal{A} \equiv \frac{1}{K} \sum_k SNR_k \propto \log \left(\prod_k snr_k \right)^{1/K} . \tag{18}$$

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, where each frequency band is treated as a dimension capable of transmitting Nyquist rate samples, having a number of levels proportional to the linear signal to noise ratio.

The Shannon Gaussian channel capacity formula

$$\mathcal{C} = \int_{-\infty}^{\infty} \log_2[1 + snr^2(f)] df \tag{19}$$

is a measure of a Gaussian channel's maximum capacity for carrying information, is very similar to Eq. 16, if we take the limit to zero filter bandwidth.

From Fig. 3, we see that $\mathcal{A}(snr)$ is a straight–line approximation to to the Shannon channel capacity formula $\mathcal{C}(snr)$. The figure shows the two functions

$$C(snr) \equiv \log_2[1 + snr^2] \tag{20}$$

and

$$A(snr) \equiv \log_2[\max(1, snr^2)], \tag{21}$$

which represent the integrands of Eq. 19 and Eq. 18 respectively.

We conclude that the articulation index, first proposed by Fletcher in an internal Western Electric report in 1921, is in fact functionally a channel capacity very similar to that for the Gaussian channel. This is important because it helps us understand the meaning and limitations of the AI, by placing it in a much firmer footing. It also provides us with an important, practical example of Shannon's channel capacity. Shannon's formulation might result in a slight improvement to the use of the max function in the traditional AI formulation. In retrospect, the function $1 + snr$ appears to fit the raw data given in the French and Steinberg paper as a smooth curve, better than their final $max(1, snr)$ formulation.

5 Discussion

We have reviewed the early speech articulation work done at AT&T research labs starting in 1921. Harvey Fletcher was the first to clearly demonstrate that probability theory could be used to account for the errors made in recognizing nonsense speech sounds. Two types of rules were found. Both were based on the assumption of "independence."

The *first* class, called *sequential processing*, results from *products of probability correct*. An example is the relationship between the syllable and the phone score, as shown in Eq. 1 or Eq. 3. Sequential processing reflects the fact that any error in a chain reduces the score. When one unit is wrong, the whole is wrong. The *second class*, called *parallel processing*, results from the *products of probabilities of errors*. An example is given by Eq. 11, which corresponds to across-frequency listening. In parallel processing, a single channel, can dramatically increase the articulation. In parallel processing, channels having no information (chance error close to 1) do not contribute to the final score.

A non-rigorous analysis leads to the conclusion that consonants and vowel articulations must be approximately equal. This follows from the fact that Eq. 1 is well modeled by Eq. 3, as defined by Eq. 2. Real words are a simple extension of these principles, as shown by Boothroyd in 1988 (Eq. 3.1), which is an example of parallel processing. Here word context is modeled as $j-1$ independent parallel channels (i.e., a product of errors is used for this model).

Using a concept of *linearity of articulation* Eq. 6, Fletcher found the speech information density across frequency, which turned out to be constant on a cochlear filter bandwidth (critical band) scale. This analysis lead to demonstrating that across frequency band the errors are consistent with independence Eq. 11. Thus across frequency, speech information appears to be an example of parallel processing.

Using a specific definition of the signal to noise ratio, we call *SNR*, a formula for the articulation was established Eq. 17. This formula depends on the average signal to noise ratio, and this formula is very close to Shannon's channel capacity formula for the Gaussian channel Eq. 19.

When the Fletcher AI theory was developed, the computer had not yet been in-

vented, thus many of the variables were expressed in log units (i.e., dB) to simplify the computations. In my view, this use of log units, justified in those early days for computational reasons, obscures the fundamentals. Thus we have reviewed AI theory using a modern notation.

The AI and the channel capacity measure are nearly identical (Fig. 3). I suspect that if we start viewing AI as a channel capacity, we will use it more effectively. Also there may be other extensions that have been developed in the *Theory of Communications*, that could be applied back to speech communication.

Given what we know today, it would be better to compute the *SNR* based on a cochlear filter bank, and what we know about loudness integration times. Cochlear filter bandwidths are presently an uncertain quantity of human hearing (Allen 1996; Shera et al. 2002; Oxenham and Shera 2003). An averaging time constant of 200 ms, following the filter bank, corresponds to the integration of loudness over time (Munson 1947).

A Appendix: Historical context

The early idea of a channel capacity, first proposed by R.V.L. Hartley (Hartley 1928), was to count the number of intensity levels in units of noise variance (Wozencraft and Jacobs 1965). This idea has its historical roots in the psychophysical literature, and is a conceptually related to "counting JNDs." The internal noise variance of the auditory system determines the number of physiological levels (i.e., the number of intensity JNDs). Thurstone (Thurstone 1927) is given credit for the first developing the idea of precisely relating psychological variables to the JND (for a detailed review see (Torgerson 1967)). Fletcher also performed a very interesting full analysis of the number of joint intensity and frequency JNDs in an early analysis (Fletcher 1923a; Fletcher 1923b), prior to Thurstone's famous work of 1927. (The 1923 paper of Fletcher's contains an extensive bibliography of the early literature.) Allen and Neely first determined the relationship between ΔL and L for the case of loudness L (Allen and Neely 1997).

It is interesting and relevant that R.V.L. Hartley (a Rhodes scholar, well versed in psychophysical concepts) also proposed the decibel, which was also based on the intensity JND ΔI (Hartley 1929; Hartley 1919). Prior to 1924, loudness was assumed

to be proportional to the log of intensity, which was called *Fechner's Law*. Fechner's analysis leading to Fechner's law was soon to be proved to be wrong by Riesz and by Fletcher and Munson (Allen and Neely 1997). The expression

$$\ln(1 + snr^2) = \ln\left(\frac{I + \Delta I}{I}\right) \approx \frac{\Delta I}{I} - \frac{1}{2}\left(\frac{\Delta I}{I}\right)^2 + \cdots, \tag{22}$$

(the approximation holding when the ratio $\Delta I/I$ is small) where ΔI and I are the JND and intensity respectively, is closely related to counting JNDs. It has been shown, by George A. Miller (Miller 1947), that noise is close to the first JND level if its presence changes the input stimulus by 0.43 dB, corresponding to a Weber fraction of 0.1. That is

$$10\log_{10}(1 + \Delta I/I) = 0.43 \approx \frac{10}{\ln(10)}\frac{\Delta I}{I},$$

thus $\Delta I/I \approx 0.1$.

The function $\log_2(1 + snr^2)$ is related to the number of JNDs (in bits) (French and Steinberg 1947; Fletcher and Galt 1950; Allen and Neely 1997). The product of the number of articulation bands times the number of JNDs determines a volume, just as the channel capacity determines an information based volume. In my view, the similarity between the AI and the Shannon channel capacity, is striking.

It is difficult to understand why these two great thinkers, Fletcher and Shannon, did not connect. From 1935 to 1950 Fletcher was the head of the Bell Labs Physical Research Laboratory. Shannon worked at Bell Labs from 1941 to 1956. The chain of management was Fletcher (Dir. of Physical Research), Bode (Dir.), Schelkunoff or Dietzold (Dept. Head), Shannon.

Endnotes

1. We shall use *snr* to denote the RMS voltage ratio of the signal and the noise, and *SNR* to denote *snr* expressed in dB, namely $SNR \equiv 20\log_{10}(snr)$.
2. These formulae only apply to nonsense speech sounds, *not* meaningful words. The extension to meaningful sounds (cat, hat) has been studied by Boothroyd (Boothroyd 1968; Boothroyd and Nittrouer 1988), and more recently by Bronkhorst (Bronkhorst et al. 1993; Bronkhorst et al. 2002).
3. This result was discovered by Galt. He tried several times to publish his observation, but the paper was rejected by JASA. This was a source of considerable anguish to Galt, as may be observed in his many notebooks (Rankovic and Allen 2000).
4. Page 106 (French and Steinberg 1947).
5. Both Fletcher, and French and Steinberg, worked almost exclusively in dB units, which were more convenient in those times. Now that the computer is available to compute these formulas, it seems better to rework the equations in a more meaningful notation.

References

Allen, J. (2004). "Articulation and intelligibility," in Juang, B. H., editor, *Lectures in Speech and Audio Processing*. Morgan and Claypool, 3401 Buckskin Trail, La-Porte, CO 80535. TO APPEAR.

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

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

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

Boothroyd, A. (1968). "Statistical theory of the speech discrimination score," *J. Acoust. Soc. Am.* **43**(2):362–367.

Boothroyd, A. and Nittrouer, S. (1988). "Mathematical treatment of context effects in phoneme and word recognition," *J. Acoust. Soc. Am.* **84**(1):101–114.

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

Bronkhorst, A. W., Brand, T., and Wagener, K. (2002). "Evaluation of context effects in sentence regonition," *J. Acoust. Soc. Am.* **111**(6):2874–2886.

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

Duggirala, V., Studebaker, G. A., Pavlovic, C. V., and Sherbecoe, R. L. (1988). "Frequency importance functions for a feature recognition test material," *J. Acoust. Soc. Am.* **83**(6):2372–2382.

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

Fletcher, H. (1921). "An empirical theory of telephone quality," *AT&T Internal Memorandum* **101**(6).

Fletcher, H. (1922). "The nature of speech and its interpretation," *J. Franklin Inst.* **193**(6):729–747.

Fletcher, H. (1923a). "Physical measurements of audition and their bearing on the theory of hearing," *J. Franklin Inst.* **196**(3):289–326.

Fletcher, H. (1923b). "Physical measurements of audition and their bearing on the theory of hearing," *Bell System Tech. Jol.* **ii**(4):145–180.

Fletcher, H. (1929). *Speech and Hearing*. D. Van Nostrand Company, Inc., New York.

Fletcher, H. (1938a). "Loudness, masking and their relation to the hearing process and the problem of noise measurement," *J. Acoust. Soc. Am.* **9**:275–293.

Fletcher, H. (1938b). "The mechanism of hearing as revealed through experiments on the masking effect of thermal noise," *Proc. Nat. Acad. Sci.* **24**:265–274.

Fletcher, H. (1948). "Perception of speech and its relation to telephony," *Science* **108**:682.

Fletcher, H. (1950). "A method of calculating hearing loss for speech from an audiogram," *J. Acoust. Soc. Am.* **22**:1–5.

Fletcher, H. (1995). "Speech and hearing in communication," in Allen, J. B., editor, *The ASA edition of Speech and Hearing in Communication*. Acoustical Society of America, New York.

Fletcher, H. and Galt, R. (1950). "Perception of speech and its relation to telephony," *J. Acoust. Soc. Am.* **22**:89–151.

Fletcher, H. and Steinberg, J. (1930). "Articulation testing methods," *J. Acoust. Soc. Am.* **1**(2):17–113.

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

Hartley, R. (1919). "The function of phase difference in the binaural location of pure tones," *Phy. Rev.* **13**:373–385.

Hartley, R. (1928). "Transmission of information," *Bell System Tech. Jol.* **3**(7):535–563.

Hartley, R. (1929). ""TU" becomes "DECIBEL"," *Telephone Engineering* **33**:40.

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

Munson, W. (1947). "The growth of auditory sensation," *J. Acoust. Soc. Am.* **19**:584–591.

Oxenham, A. J. and Shera, C. A. (2003). "Estimates of human cochlear tuning at low levels using forward and simultaneous masking," *J. Assoc. Res. Otolaryngol.* **4**(4):541–554.

Rankovic, C. and Allen, J. (2000). *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.

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

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

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

Thurstone, L. (1927). "A law of comparitive judgment," *Psychol. Rev.* **34**:273–286.

Torgerson, W. S. (1967). *Theory and Methods of Scaling*. John Wiley and Sons, Inc., New York.

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

Table 1: *Table of definitions required for the articulation index experiments.*

DEFINITIONS FOR THE AI

SYMBOL	DEFINITION
α	Gain applied to the speech
$c(\alpha) \equiv P_c(\text{consonant}\|\alpha)$	consonant articulation
$v(\alpha) \equiv P_c(\text{vowel}\|\alpha)$	vowel articulation
$s(\alpha) = [2c(\alpha) + v(\alpha)]/3$	Average phone articulation for CVC's
$e(\alpha) = 1 - s(\alpha)$	Phone articulation error
f_c	Highpass and lowpass cutoff frequency
$s_{_L}(\alpha, f_c)$	s for lowpass filtered speech
$s_{_H}(\alpha, f_c)$	s for highpass filtered speech
$S(\alpha)$	Nonsense syllable (CVC) articulation
$W(\alpha)$	Word intelligibility
$\mathcal{I}(\alpha)$	Sentence intelligibility

Figure Captions

Fig. 1. This figure shows the ratio of the articulation index density (also called the *speech importance function*), and the *critical bandwidth* (also called the *equivalent rectangular bandwidth* or ERB), which is a measure of the cochlear filter bandwidth. The critical bandwidth was derived from the ratio of the RMS level of a tone, adjusted to its detection threshold level, to the spectral level of a noise. Note that ratio has units of bandwidth. From the figure we conclude that the information density of speech used in the AT&T tests, per cochlear critical band, is approximately uniform.

Fig. 2. This figure shows a typical set of results for the French and Steinberg AI model, as defined in (Allen 1994). See the text for a detailed description of each panel.

Fig. 3. Plot of $\log(1 + snr^2)$ and $\log[\max(1, snr)^2]$ versus $SNR = 20 * \log(snr)$.

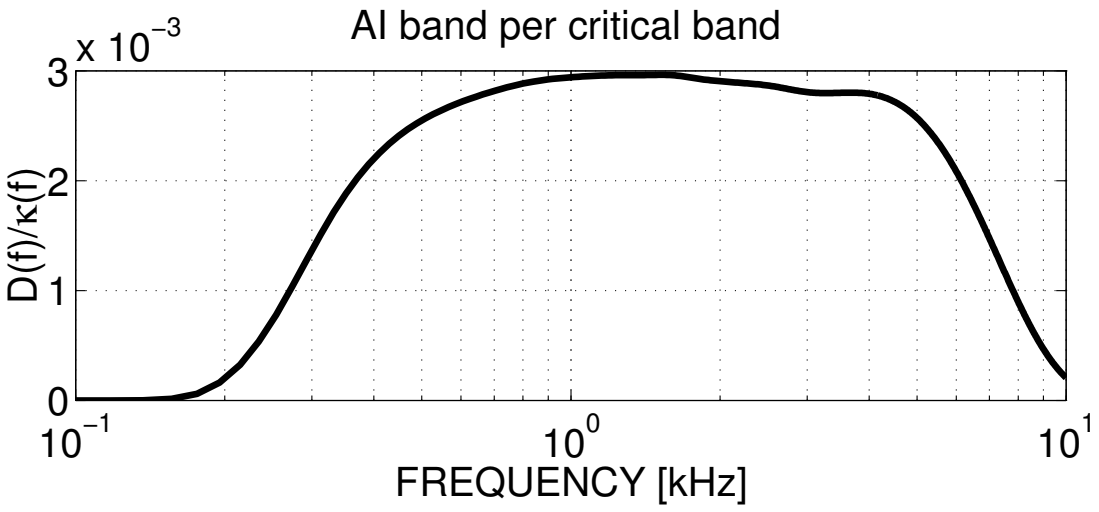
AI band per critical band
x 10^−3
3
2
1
0
D(f)/κ(f)
10^−1
10^0
10^1
FREQUENCY [kHz]

Figure 1

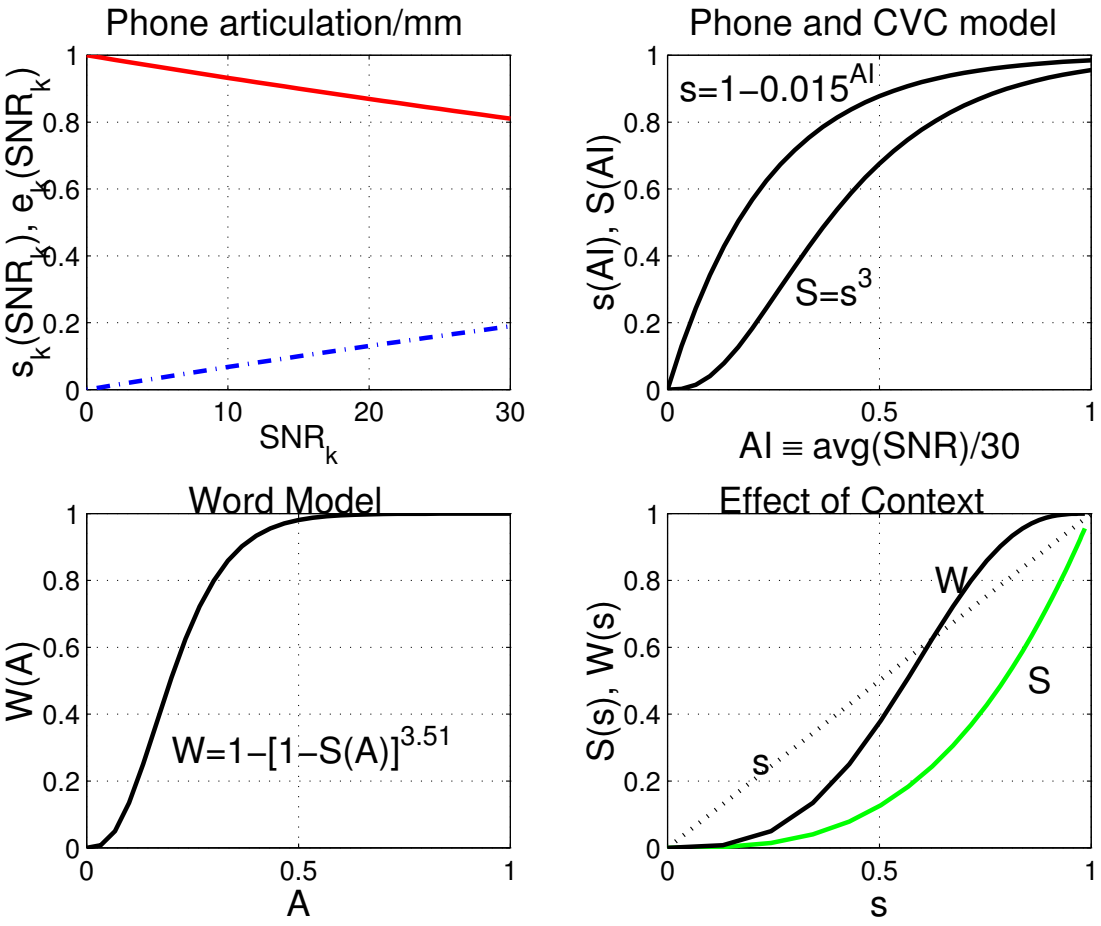
Phone articulation/mm
$s_k(SNR_k)$, $e_k(SNR_k)$
SNR_k
Phone and CVC model
s=1−0.015^AI
S=s^3
s(AI), S(AI)
AI ≡ avg(SNR)/30
Word Model
W=1−[1−S(A)]^3.51
W(A)
A
Effect of Context
W
S
s
S(s), W(s)
s

Figure 2

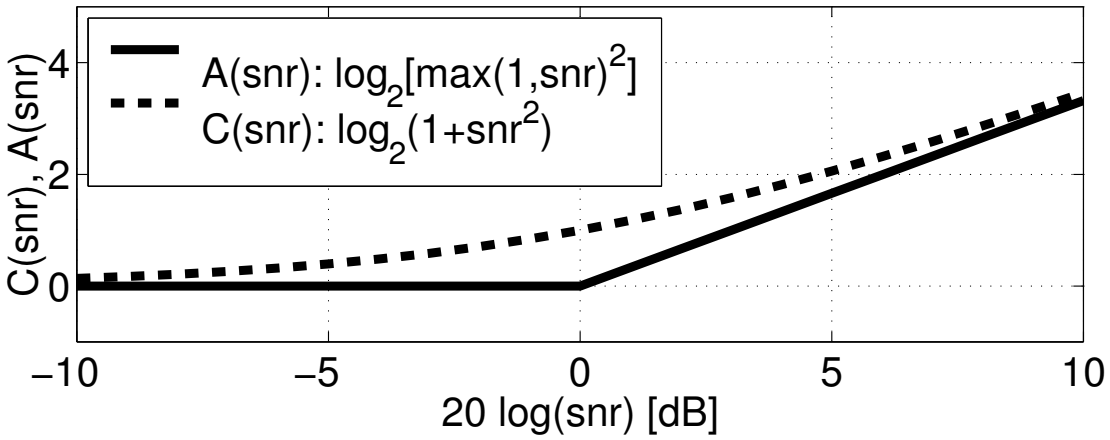
A(snr): log2[max(1,snr)2]
C(snr): log2(1+snr2)
C(snr), A(snr)
4
2
0
−10
−5
0
5
10
20 log(snr) [dB]

Figure 3