Articulation and Intelligibility

Jont B. Allen
Room 2061.
University of IL,
Beckman Inst.

November 29, 2004

1 Introduction

As discussed by George Miller in his book 1951 book *Language and Communication*, speech intelligibility depends on a code in that it is incredibly robust to mangling distortions. Can we analyze speech intelligibility using the scientific method? Miller's analysis shows an analysis that is both scientific and artful. His scientific analysis critically depends on the use of statistics, information theory, and psychophysical methods.

The science of intelligibility is a science of the error analysis of human speech communication. The goal in science is always to make a mathematical model. Data is collect, and tested against the model. This article is a review of what is known about modeling *human speech recognition* (HSR).[1] A model is proposed, and the data are tested against the model.

As developed in Miller's book, information and communication theory form the basic underlying scientific basis for understanding the speech code. One of the basic tools of information theory is the *channel*, the mathematical correlate to a pair of wire. The channel is the fundamental building blocks of any theory of speech and language communication.

The art of studying speech intelligibility is to restrict the problem in such a way that progress may be made. Intelligibility depends on *visual* as well as auditory cues. In this review our discussion will be limited to that subset of intelligibility which is unique to acoustic speech, *without visual input*. Thus we assume that we wish to model the *auditory channel*, in the absents of the *visual channel*.

In many of the early studies of human speech recognition, the many effects of language *context* were removed, by testing with nonsense sounds. When listening to meaningful words and sentences, people report what they understand, leaving many errors they hear unreported. When listening to nonsense speech, having limited context constraints, people report what they actually hear. Thus to meaningfully study the decoding of speech sounds, one must carefully control for *context*

[1] See Table 1 tab:Abbreviations for each abbreviation.

effects. One does this by removing the *context channel*, by the use of nonsense sounds.

ABBREVIATIONS

ASR	**Automatic Speech Recognition**
HSR	**Human Speech Recognition**
CV	**consonant-vowel** (ex. “pa, at, be, c”)
CVC	**consonant-vowel-consonant** (ex. “cat, poz, hup”)
snr	**Signal to Noise Ratio** (*linear units*) Eq. 14 eq:snr
SNR	$\mathbf{20 \log_{10}}(snr)$ [dB units]
AI	**Articulation Index**
$\mathcal{A}_k$	Specific AI [*dB/dB units*] Eq. 15 eq:SNR
PI(*snr*)	**Performance Intensity function**
AM	**Articulation Matrix**
VOT	**Voice onset time**
ZP	**Zero predictability**
LP	**Low predictability**
HP	**High predictability**
ERP	**Event Related (scalp) Potential**

Table 1: *Table of abbreviations.*

The complete elimination of all context channels is an impossible task, and even undesirable. English nonsense sounds are a distinct subset of all sounds, and even a small subset of the sounds of language. For example, *tonal* sounds are common in many languages such as Chinese, but are absent in English. Thus the subset of English nonsense sounds, while rich enough to encode English, is a distinct subset of human vocalization. The best we can do is attempt to characterize these more subtle context channels, not eliminate them, as we attempt to saddle this untamable “context beast.”

Intelligibility is the identification of meaningful speech, while *articulation* is the identification of *nonsense* speech sounds.[2] Understanding intelligibility scores requires models of syntax (structure), lexicality (vocabulary), grammar (form) and semantics (meaning). To understand articulation scores, only models of phonology are required, by design, by carefully controlling for as many of these context channels as possible.

The word articulation is a tricky term, as it has strong meanings in both the production (physical) and perceptual (psychophysical) domains. An *articulatory feature* is a speech production concept, whereas the *articulation matrix* is a perceptual concept. This is quite unfortunate that this word has these two very different, yet related, meanings. We have inherited these terms from the long past, and thus must deal with this confusion. One way to deal with this problem of terminology would be to create new terms. However this would just obscure and confuse the situation further, so I have avoided that approach. It is better to be aware of,

[2] See Table 2 tab:Defi nitions for important defi nitions.

understand, and carefully parse the two meanings.

Is there any information left in speech after these major information channels, visual and context, have been removed? Emphatically, *yes!* In fact humans do quite well in identifying basic speech sounds well below 0 dB SNR, without these powerful information side channels. The reasons for this *natural robustness* in HSR are becoming increasingly clear, and is the topic of this review.

1.1 Problem statement

Articulation has been studied since the turn of the 20th century by teams of physicists, mathematicians and engineers at Western Electric Engineering, and later at The Bell Laboratories. A key science, information theory, has been frequently emphasized, yet rarely applied to the field of speech recognition. As a result, the source of robustness in HSR is poorly understood. For example, many believe that robustness follows from context. An example of this view may be found in Flanagan's classic text *Speech analysis Synthesis and Perception* (Flanagan, 1965, p. 238)

> *Items such as syllables, words, phrases, and sometimes even sentences, may therefore have a perceptual unit. In such an event, efforts to explain perception in terms of sequential identification of smaller segments would not be successful.*

Discuss this quote with Jim Flanagan.

This summary of human speech perception says that larger units such as words and maybe even sentences, may be the perceptual units (the events), and that attempting to work with smaller units would not work. Nothing could be further from the truth, as we shall see. Speech is first detected in white masking noise at about -25 dB SNR, and basic sound classes are easily discriminated at -20 dB SNR.

Only data and experiment can resolve this fundamental question "what is the smallest unit that make up the building blocks of oral speech perception?' The resolution of this question may be found by comparing intelligibility and articulation data. The robustness of nonsense speech has been measured with a confusion matrix (Campbell, 1910; Miller and Nicely, 1955), which we denote the *Articulation Matrix* (AM). Many important issues regarding AM data remain unstudied, as extensively discussed in this review.

This brings us to Miller's unsolved problem, the decoding of nonsense speech sounds, which have been mangled by filtering and noise. What are the remaining information channels that need to be accounted for, and how can we model them? This review will explore this question in some detail. We begin in Section 1.2 with some definitions and an overview of the robustness problem. We then proceed in Section 2 with a literature review of articulation testing and the articulation index (AI), and the work of George Miller, who first controlled for the *articulation test entropy* $\mathcal{H}$ with the use of close set testing. This leads us to an AI analysis of Miller and Nicely's consonant confusion data. In Section 3 we look at the nature of the context channel. From this higher ground we model how oral speech is coded and processed by the auditory system.

TERM:	DEFINITION:
phone	A consonant (C) or vowel (V) speech sound
syllable	A sequence of C's and V's, denoted {C,V}
word	A *meaningful* syllable
phoneme	Any equivalent set of phones which leave a word meaning invariant
allophones	All the phone variants for a given phoneme
recognition	Probability measure $\boldsymbol{P_c}$ of correct phone identification
articulation	Recognition of nonsense syllables ({C,V})
intelligibility	Recognition of words (i.e., meaningful speech)
confusion matrix	Table of identification frequencies $N_{sh} \equiv N_{h\|s}$
articulation matrix	A *confusion matrix* based on nonsense sounds
relative robustness	Ratio of the conditional entropies for two conditions to be compared
event	A perceptual feature. Multiple events define a phone.
trial	A single presentation of a set of events
state	A values of a set of events at some instant of time
state machine	A machine (program) that transforms from one state to another
noiseless state machine	A deterministic state machine
p_n	Probability of event n, of N possible events
information density	$\boldsymbol{I_n \equiv \log_2(1/p_n)}, \qquad \boldsymbol{n = 1, \cdots, N}$
entropy	Average information: $\boldsymbol{\mathcal{H} \equiv \sum_{n=1}^{N} p_n I_n}$
conditional entropy	A measure of context: high entropy $\Longrightarrow$ low context
context	Coordinated combinations of events within a trial
message	Specific information transmitted by a trial (e.g., a syllable)
AI	Articulation index $\boldsymbol{AI \equiv \frac{1}{K} \sum k = 1^{K} AI_k}$

Table 2: *Table of definitions.*

1.2 Basic Definitions and Abbreviations

Tables 1 and 2 provide key abbreviations and definitions use throughout the paper. While it is important to carefully define all the terms, this section could be a distraction to the flow of the discussion. Rather than bury these definitions in an Appendix, I have placed them here, but with the warning that the reader should not get bogged down with the definitions at first. I suggest you first skim over this section, to familiarize yourself with its content. Then proceed with the remainder of the material, coming back when an important idea or definition is unclear. Refer to the Tables as a quick guide, and the text once the basic ideas of the model are established. It is essential you understand the definition of *articulation*, the *event*, and why the term *phone* is used rather than the more popular term *phoneme*. A qualitative understand of *entropy* is also required. All of the required terms are now carefully defined.

The phone vs. phoneme: The *phone* is any basic speech sound, such as a consonant or vowel, or a cluster of these units. It must be carefully distinguished from the *phoneme* which is notoriously difficult to define because all of these definitions incorporate *minimal meaning*. A definition (see Table 2) has been carefully chosen to be common, if not widely accepted, but perhaps not agreed upon by everyone.

We shall argue that meaning is irrelevant to the speech robustness problem. During WW-II, people were trained to transcribe languages that they did not understand, and they did this with agility and fidelity. Fletcher AI theory (1921–1950) was based on nonsense CV, VC and CVC syllables. Miller and Nicely's classic study (1955) used isolated consonants, which by themselves have no meaning.

Thus one may proceed with the study of human speech recognition, without the concept of meaning, and therefore the phoneme. This view has a nice parallel with Shannon's (1948) theory of information, which specifically rejected meaning as relevant.

It is difficult to argue strongly for the importance of the phoneme, whose definition depends on meaning, if meaning plays little or no role in peripheral language processing (the robust identification of unit phones).

A *Syllables* is one or more phones. A *word* is a syllable with meaning (it is found in a dictionary).

Recognition is the probability for correct average identification, denoted P_c. *Recognition error* is given by

$$E_{\%} \equiv 100 * (1 - P_c),$$

is typically quoted in percent,[3] and is the sum over all the individual sound confusions defined by AM $P_{h|s}$, where $h \neq s$. The recognition (P_c) (and thus the corresponding recognition error $E_{\%}$) is a function of the signal to noise ratio *snr*. When the recognition is measured as a function of the signal to noise ratio, it is frequently called a *performance-intensity* (PI) function, defined by $P_c(snr)$.

[3]The symbol $\equiv$ is read "equivalence" and means that the quantity on the left is defined by the quantity on the right.

Intelligibility is the recognition of meaningful sounds (typically words), while *articulation* is the recognition of *meaningless speech sounds* (e.g., nonsense phones) (Fletcher, 1929, Page 255).

Robustness: An important, but again difficult concept to define, is that of *robustness*, the main subject of this review. The first property of robustness must be that it is a relative measure. Second we would like a measure that is defined in terms of bits. Specific examples of the use of the relative robustness can help us to further nail down the full definition: An important example is the robustness of one sound versus another (i.e., /pa/ vs. /ma/). A second is the robustness of one group of sounds, say the nasals, against another group of sounds, say the fricatives.

In each example there are two cases we wish to compare, and we would like a measure that tells us which is more robust. The candidate measure for the first example of two sounds is the conditional entropy, defined as

$$\mathcal{H}(h|s) = -\sum_h P_h \log_2(P_{h|s}),$$

which is just the entropy of row s of the articulation matrix $P_{s,h}$.

This measure is in bits, as required, for each spoken sound s. This measure has the unfortunate property that it becomes smaller as the sound becomes more certain, which is backward from a robustness measure. If we define the relative robustness as the ratio of two conditional entropies, for the two different sounds, then we have a measure that increases as the score increases. For example, the robustness of spoken sound s_2 relative to that of s_1 would be

$$\mathcal{R}(s_2/s_1) = \frac{\sum_h P_h \log_2(P_{h|s_1})}{\sum_h P_h \log_2(P_{h|s_2})}.$$

This measure would increase if s_2 is more robust (has a smaller conditional entropy) than s_1.

As a second example lets take the relative robustness of intelligibility vs. articulation (i.e., the effect of context). In this case the robustness due to intelligibility would be taken to be

$$\mathcal{R}(\mathcal{I}/\mathcal{A}) = \frac{\sum_{s,h} P_{s,h}(\mathcal{A}) \log_2(P_{h|s}(\mathcal{A}))}{\sum_{s,h} P_{s,h}(\mathcal{I}) \log_2(P_{h|s}(\mathcal{I})}),$$

where $P(\mathcal{I})$ is with context (intelligibility) and $P(\mathcal{A})$ is with no context (articulation).

If we wish to compare the relative robustness of ASR and HSR, the robustness would then be

$$\mathcal{R}(\text{HSR}/\text{ASR}) = \frac{\sum_{s,h} P_{s,h}(\text{ASR}) \log_2(P_{h|s}(\text{ASR}))}{\sum_{s,h} P_{s,h}(\text{HSR}) \log_2(P_{h|s}(\text{HSR}))}.$$

For these last examples it makes sense to restrict comparisons to cases which have the same maximum entropy, namely for which the corpus is the same size.

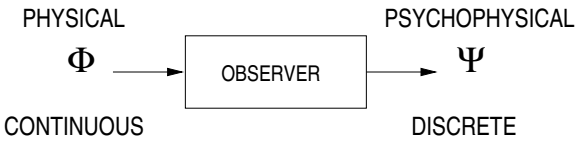

Figure 1: *The basic model of an observer with the physical variables Φ on the left and the psychophysical variables Ψ on the right. As an example of a pair of variables is acoustic intensity as the Φ or physical intensity and and loudness, the Ψ or psychoacoustic correlate of intensity. In the case of speech perception we treat physical variables as analog (continuous) and psychophysical variables as discrete, as in the case of events.*

Events: In the speech perception literature the terms *articulatory feature perceptual feature* and *distinctive feature* are commonly used, even interchangeably. For example, *voicing*, *nasality*, and the *place* of a constriction in the vocal tract, that occur when forming a speech sound, constitute typical articulatory features. The term voicing is frequently spoken of as a perceptual feature. It seems wise to choose a new word to represent the *perceptual correlates* of speech features. We use the word *event* to deal with such meaning.

The basic model of psychophysics and the observer is shown in Fig. 1 fig:PhiPsi . As for the case of intensity and loudness, we need a language for relating *perceptual features* (Ψ variables) to *articulatory features* (Φ variables). Thus we speak of the *event* when referring to the Ψ correlate of an speech Φ feature. For example, it might turn out that the Ψ-event corresponding to the Φ-feature *voicing* is determined by quantizing the so-called Φ voice onset time (VOT) to some fixed time range of values. For example, a Φ-VOT between 0 and 30 ms would be Ψ-VOICED, while Φ-VOTs greater than 30 ms would be Ψ-UNVOICED. This may not a particularly good example of an event, since we have yet to isolate one.

Make the point that the event is a binary random variable, measured in the real world as a probability.

The *event* must be measured by experimental outcomes, expressed as a probability, rather than assumed *a priori*, as in the case of *distinctive features*. The articulation matrix $\mathcal{A}(snr)$ is the measure of these experimental outcomes. We do an experiment where we repeat the presentation of the stimulus many times, and we then define a probability measure of the underlying binary *event*, in terms of the frequency of its observation, based on a large number of subjects and a large number of talkers.

Each presentation and reception is called a *trial*. This idea is a formal one, as described by books on communication theory (Wozencraft and Jacobs, 1965, Chapter 2) and probability theory (Papoulis, 1965, Section 2-2). The definitions, of a *trial* and an *event*, as defined in this mathematical literature, are ideally suited to our purpose.

When groups of events are mathematically bound together at an instant of time, the group is called the *state* of the system. As an example, think of the events that define the states of a phone. A machine (think computer program) is typically pictured as a box that transforms an input state into an output state. Such a program is call a *state machine*. When the state machine is deterministic, it is called a *noiseless state machine*. During training (the learning phase), the state is not de-

terministic, but the learning mode is considered to be an exception, for the purpose of modeling the state machine. We view the auditory brain as a state machine decoding the events coming out of many event processors, having inputs from the cochlea. This model structure represents the "front end" of the HSR system. This model is based on experimental observations, not fanciful dreams. Any model can be wrong, but it must be rejected based on experimental outcomes, and a better model.

SNR: The snr plays a very important role in the theory of HSR because it is the underlying variable in the articulation index measure. The detection of any signal is ultimately limited by detector noise. This leads to the concept of an internal noise, specified as a function of frequency. It is the *internal* signal to noise ratio *snr*(f), a Ψ variable, that ultimately determines our perceptual performance (French and Steinberg, 1947; Allen and Neely, 1997). This quantity must be inferred from external measurements.

An example is instructive: The external *snr* of a pure tone, in wide band noise, is not perceptually meaningful since a relevant noise bandwidth must be used when calculating the detection threshold. This bandwidth, called the *critical bandwidth*.[4] is cochlear in origin, since the internal *snr*(f) depends on cochlear filtering. The discovery of the cochlear critical bandwidth marked the recognition of this fact (Fletcher and Munson, 1937; Fletcher, 1938; French and Steinberg, 1947; Allen, 2001).

Exactly the same principle applies to the detection of speech. The detection threshold for speech sounds are determined by the same cochlear critical bandwidth. In the case of speech however, unlike the tonal case, the peak to RMS ratio of the speech in the band becomes a key factor, when estimating the speech detection threshold. These basic issues of speech detection and articulation were well understood by Fletcher and his colleagues Wegel, Steinberg and Munson, and were repeatedly described in their many early papers. These points will be carefully reviewed in the next section of this review, *Articulation*.

Two different notations for the signal to noise ratio shall be used, σ_s/σ_n, denoted *snr*, and $10\log_{10}(\sigma_s^2/\sigma_n^2)$, denoted *SNR*, in dB. Each of these measures will be indexed by k to indicate frequency, indexed by critical band. Thus *snr*$_k \equiv \sigma_{s,k}/\sigma_{n,k}$ in frequency band k.

Context and entropy: The concept of *context* in language is ubiquitous. Context results from a time–correlated sequence of speech units, leading to the higher probability of predicting a word, given the preceding words. Mathematically, this can be expressed as

$$P_c(x_n|\mathcal{C} = x_1 x_2 \ldots x_N) \geq P_c(x_n), \qquad n \in \{1, 2, \ldots, N\}. \tag{1}$$

[4] In many of the early papers the level of a tone in noise above threshold, expressed in dB-SL, was commonly denoted by the variable Z (French and Steinberg, 1947, Eq. 2, page 97). This defi nition explicitly accounts for the critical bandwidth of the ear.

where x_n are speech units and $\mathcal{C}$ is the conditioning context. If x_n are random unrelated units (i.e., words or phonemes), then the sequence $x_1, x_2, \ldots x_N$ does not change the score of x_n; namely the conditional recognition of x_n is the same as that in isolated speech.

It is critically important to control for context effects when studying speech recognition. Real words have greater context than randomly ordered meaningless speech sounds, which ideally, would have none. Meaningful HP sentences have greater context than nonsense ZP sentences. One classic way of modeling context is with Markoff models (Shannon, 1948; Shannon, 1951).

By *redundancy* we mean the repetition of events within a trial.[5] Interestingly, sometimes redundancy requires context to recognize the redundancy, as in the example *Sierra Mountains*.[6]

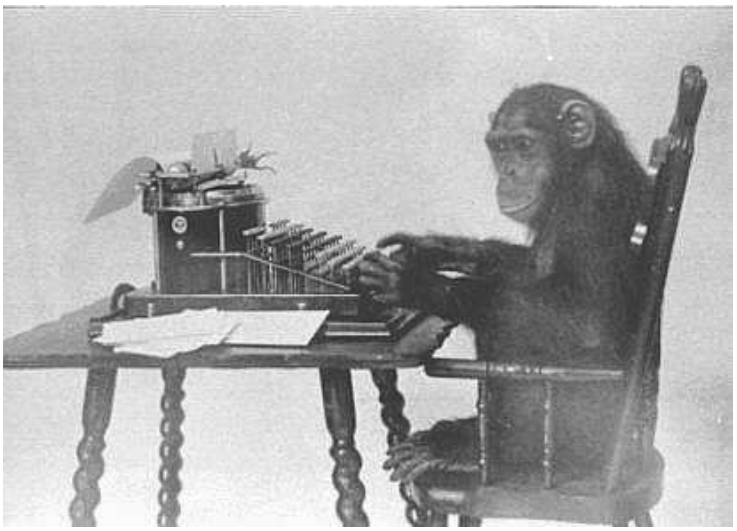

Figure 2: *Who is this monkey thinking of, an what does she want to say? How many messages can a monkey type? This picture epitomizes the concept of* Entropy*. It is highly likely, yet not impossible, that through random typing, this monkey will produce a work of Shakespeare, corresponding to a astronomically small entropy. Even a single sentences virtually impossible.*

The *information density* I_n is defined as the log base 2 of the reciprocal probability p_n. The log base 2, is a simple transformation that gives units of *bits*. The important concept here is reciprocal probability, so that a rare event (small probability) is defined as having large information. The concept of information requires a *set of outcomes*. Thus p_n requires an index n labeling N possible outcomes, while p_n measures the relative frequency (parts of a whole) of these outcomes, which obey the constraint that $\sum_n p_n = 1$.

Entropy is the average amount of information, as computed by taking a weighted average of the information density. When all the outcomes are equal (i.e., $p_n =$

[5] This term has been mathematically defi ned by Shannon in his classic paper (Shannon, 1948, Page 24).

[6] The word Sierra (redundancy) means mountain in Spanish (a language context).

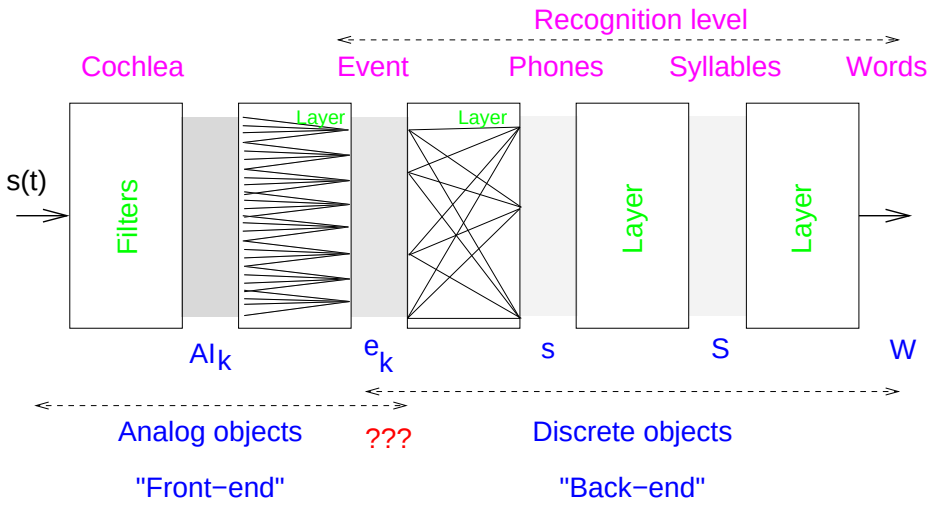

Figure 3: *Model block diagram summary of speech recognition by humans. At the top of each block is a label that attempts to identify the physical operation, or a unit being recognized. The labels below the boxes indicate the probability measure defined at that level. See the text for the discussion of objects, at the very bottom. The speech* $s(t)$ *enters on the left and is processed by the cochlea (first block), breaking the signal into a filtered continuum of band-passed responses. The output of the cochlea is characterized by the specific* AI_k*, a normalized snr, expressed in dB units. The second box represents the work of the early auditory brain, which is responsible for the identification of events in the speech signal, such as onset transients and the detection of measures which define things like the VOT. The third block puts these basic features together defining phones. The remaining blocks account for context processing.*

$1/N$) the entropy $\mathcal{H}$ is maximum, and the information is minimum (Cover and Thomas, 1991). Figure 2 fig:MonkeyTyping is an epitome of entropy. How many monkeys would it take to produce a work of Shakespeare? The entropy of such a document is very low. The number of monkeys required to produce such a document is astronomical.

1.3 Modeling HSR

It is important to develop a model of human speech recognition (HSR) which summarizes what we know in a succinct manner. A model is presented in Fig. 3 fig:RecMod which shows the structural relations between the various quantitative probabilistic measures of recognition.

It is widely accepted, typically with no justification, that HSR is modeled by a *front-end* driving a *back-end*. These terms have been used loosely in the past, and they can have different meanings in different fields (e.g., speech, psychology and physiology). We shall define the front-end as the acoustics processing and *event extraction*, and the back-end as the *context processing*. An excellent quantitative justification for doing this is provided by the work of Boothroyd (1968) and later Bronkhorst (Bronkhorst et al., 1993), who defined a front-end and a back-end in a

mathematical model of context processing. Bronkhorst *et al.* integrated Fletcher's AI model and generalized Boothroyd's context models to include all possible combinations of recognition errors, thereby quantitatively extending context models. They also derived model weighting coefficients from first principles, using a lexicon. In 1968 Boothroyd modeled the effect of word recognition, given phone scores, as a contextual constraint, and made empirical models to account for this context effect (Boothroyd, 1968; Boothroyd and Nittrouer, 1988; Boothroyd, 1993).

Their is a long–standing unanswered question: *Is there feedback from the back–end to the front–end?* The HSR model shown in Fig. 3 fig:RecMod , assumes that *events* are extracted from the cochlear output in frequency regions (up to, say, the *auditory cortex*), and then these discrete events are integrated by a *noiseless state machine* representing the cerebral cortex. One of the most important issues developed here is that *front–end* phone feature recognition analysis appears to be independent of the *back-end* context analysis. Thus in the model shown in Fig. 3 fig:RecMod there is no feedback.

Furthermore, in this model, *all* of the recognition errors in HSR are a result of event extraction labeling errors, depicted by the second box of Fig. 3 fig:RecMod , modeled by the articulation-band errors e_k. In other words, *sound recognition errors are modeled as a noise in the conversion from analog objects to discrete objects*. I will argue that much of this event processing is implemented as *parallel processing*,[7] which is equivalent to assuming that the event recognitions are independent.

As shown in the figure, the input speech signal is continuous, while the output stream is discrete. Somewhere within the auditory brain discrete decisions are made. A critical aspect of our understanding is to identify at what point and at what level this conversion from continuous to discrete takes place. I will argue that this conversion is early, at the event level. Once these decisions have been made, the processing is modeled as a *noiseless state machine* (i.e., a state machine having no stochastic elements).

When testing either HSR or ASR systems, *it is critical to control for language context effects*. This was one of the first lessons learned by Fletcher *et al.*, that context is a powerful effect, since the score is strongly affected by context.

The HSR model of Fig. 3 fig:RecMod is a "bottom–up," divide and conquer strategy. Humans recognize speech based on a hierarchy of context layers. Humans have an intrinsic robustness to noise and filtering. In fact, the experimental evidence shows that this *robustness* does not seem to interact with semantic context (language), as reflected by the absents of feedback in the model block diagram.

The auditory system has many parallels to vision. In vision, features are first extracted, such as edges in an image. As in vision, entropy decreases as we integrate the features and place them in layers of context. This view is summarized in Fig. 3 fig:RecMod as a *feed-forward* process. We recognize events, phones, phonemes, and perhaps even words, without access to high level language context. For designers of ASR systems, this is important and good news, because of its

[7] The idea behind *parallel processing* will be properly defined in Sec. 2.3 sec:CompositionLaws .

simplicity.

As early as 1963 Miller and Isard made a strong case against the use of Markoff models in speech recognition, using an argument based on robustness, in an apparent reaction to the use of language (context) models (i.e., in ASR applications this amounts to hidden Markov models, or HMM) for solving the robustness problem. While language context is key in reducing many types of errors, for both ASR and HSR, the front-end robustness problem remains. While it is widely believed that there is much room for improvement in such language models (Miller, 2001), it now seems clear that even major context processing improvements will not solve the ASR noise robustness problem. We know this from an analysis of data from the literature which shows that humans attain their inherent robustness to background noise early in the process, independent of and *before*, language context effects.

To obtain equal HSR performance between meaningful 5-7 word sentences and randomized word order sentences, required degrading the *SNR* of the meaningful sentences by 6 to 10 dB, a negligible change for HSR. They argue that a word-randomizing transformation would have a major performance degradation on a Markoff driven ASR system, which heavily depends on word order (see Sec. 1.3.1 sec:Miller62).

This leads to the following dilemma for ASR, as predicted in 1963 by Miller: Both ASR's front-end phone error rate *and* its back-end context processing are significantly worse than those of HSR. Language models can never achieve the desired goals of solving the robustness problem because it is the front-end that accounts for the errors causing the robustness issues. Thus we must deal directly with the front end problems of talker variation, noise and spectral modifications, *independent* of context effects, as HSR does. This view is not to depreciate the importance of the back end, rather is an attempt to clarify that improved context processing cannot solve the robustness problem.

1.3.1 Context Models

An Example of a Context Effect: A detailed example of the utility of context in HSR was demonstrated by Miller (1962). This example stands out because of the early use of ideas from information theory to control for the entropy of the source, with the goal of modulating human performance via context. The experiment was simple, yet it provides an insight into the workings of context in HSR. In this experiment 5 groups of 5 words each make up the test set. This is a *closed–set*[8] listening task with the number of words and the signal to noise ratio varied. There are 4 conditions. For test condition 1 the subjects are shown 1 of the 5 lists, and they hear a word from that list. For the other 3 conditions the subjects are shown 1 list of all the 25 words. The probability correct $P_c(snr)$ was measured for each of the 4 conditions:

1. 5 words

[8] A *closed–set* test is one with a limited number of outcomes that are known *a priori* to the subjects.

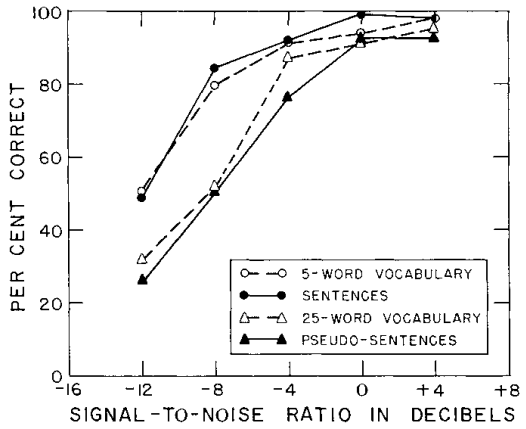

Figure 4: *This figure, from Miller (1962), summarizes the results of a 4-way experiment, performed as a function of the signal to noise ratio. Test 1 (open circles, dashed line) shows $P_c(SNR)$ for 5 word vocabularies, with no context. In test 2 (closed circles, solid line) 5 word sentences were made from the 5, 5 word lists. As an example "Don brought his black socks." The word "Don" was one of 5 possibilities [Don, He, Red, Slim, Who]. For tests 1 and 2, $P_c(snr)$ is the same. Test 3 (open triangles, dashed line) was to test using the larger corpus of one of the 25 words, spoken in isolation. Test 4 (closed triangles, solid line) was to generate "pseudo-sentences" by reversing the order of the sentences of test 3. Going from 5 to 25 isolated words (test 1 to 3) causes a 4 dB SNR reduction in performance at the 50% correct level. Presenting the 25 words as pseudo-sentences, that make no sense (test 4), has no effect on $P_c(SNR)$. However adding a grammar (test 2) to a 25 word test returns the score to the 5 word test. In summary, increasing the test size from 5 to 25 words reduces performance by 4 dB. Making 5 word meaningful sentences out of the 25 words restores the performance to the 5 word low entropy case.*

2. 5 word grammatically correct sentences, chosen from the 25 words
3. 25 words
4. non grammatical sentences chosen from the 25 words.

As described in the caption of Fig. 4 fi g:Miller62Fig1 , in condition (1) 5 word lists are used in each block of trials. The lists are randomized. The subject hears 1 of 5 words, degraded by noise, and is asked to pick the word from the list. In condition (3) the number of words is increased from 5 to 25, causing a reduction of 4 dB in performance (at the 50% level). These two conditions (1 and 3) were previously studied in a classic paper (Miller et al., 1951) which observed that the size of the set of CVCs has a large impact on the score, namely $P_c(SNR)$ depends on the entropy of the task. In condition (2), the effect of context is measured. By placing the 25 words in a context having a grammar, the scores returned to the 5 isolated word level (condition 1). When sentences having no grammar (pseudo-sentences) were used (condition 4), generated by reversing the meaningful sentences of condition 2, the score remains equal to the 25 isolated word case of condition 3.

Thus the grammar in experiment (2) improves the score to the isolated word level (1), but not beyond. It probably does this by providing an improved framework

for remembering the words. Without the grammatical framework, the subjects become confused and treat the pseudo-sentences as 25 random words (Miller and Isard, 1963).

One may quantify context by measuring the change in the *SNR* at the 50% correct point on the $P_c(SNR)$ curve, in units of *bits/dB*. In this experiment the difference in entropies (entropy of 25 words less the entropy of 5 words) is $\log_2(25) - \log_2(5) = 2.32$ bits, which means that a change of 2.32 bits corresponds to a change in the *SNR*of 4 dB. Thus the trading relation is 0.58 bits/dB (or 1.7 dB/bit). It is generally useful to compare the bits/dB in this manner.

Outline: The paper is organized as follows: Sections 2.1 and 2.2 summarizes important results from the 30 years of work (1921-1950) by Fletcher and his colleagues, which resulted in *Articulation Index theory*, 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*, which is an important concept from information theory defining the maximum amount of information that may be transmitted on a channel. Section 2.4 sec:MillerEtAl summarizes the speech work of George Miller. Miller showed the importance of source entropy (randomness) in speech perception. He did this by controlling for both the cardinality (size of the test corpus) and the signal to noise ratio of the speech samples. Section 2.8 sec:Validation discusses the validation and Sec. 2.9 sec:Criticisms criticisms of articulation index theory. Section 3 sec:Intelligibility discusses the importance of context on recognition, summarizing key results. For continuity, research results are presented in chronological order.

Figure 5: *The Acousticon LT was invented in about 1905.*

2 Articulation

In 1908 Lord Rayleigh reported on his speech perception studies using the "Acousticon LT," a commercial electronic sound system, produced in 1905. As shown in Fig. 5 fig:acousticon ,[9] it consisted of a microphone and 4 loudspeakers, and was

[9] http://dept.kent.edu/hearingaidmuseum/AcousticonLTImage.html