How Do Humans Process and Recognize Speech?

Jont B. Allen, *Fellow, IEEE*

***Abstract*—Until the performance of automatic speech recognition (ASR) hardware surpasses human performance in accuracy and robustness, we stand to gain by understanding the basic principles behind human speech recognition (HSR). This problem was studied exhaustively at Bell Labs between the years of 1918 and 1950 by Harvey Fletcher and his colleagues. The motivation for these studies was to quantify the quality of speech sounds in the telephone plant to both improve speech intelligibility and preference. To do this he and his group studied the effects of filtering and noise on speech recognition accuracy for nonsense consonant-vowel-consonant (CVC) syllables, words, and sentences. Fletcher used the term "articulation" as the probability of correct recognition for nonsense sounds, and "intelligibility" as the probability of correction recognition for words (sounds having meaning). In 1919, Fletcher found a way to transform articulation data for filtered speech into an additive density function $D(f)$ and found a formula that accurately predicts the average articulation. The area under $D(f)$ is called the "articulation index." Fletcher then went on to find relationships between the recognition errors for the nonsense speech sounds, words, and sentences. This work has recently been reviewed and partially replicated by Boothroyd and by Bronkhorst, *et al.* Taken as a whole, these studies tell us a great deal about how humans process and recognize speech sounds.**

I. Introduction

SPEECH recognition by machine is a critical core technology for the "information" age. Existing machine recognition systems do not work the way humans work. This is because automatic speech recognition (ASR) machines use spectral templates, while humans work with *partial recognition* information across frequency, probably in the form of speech features that are local in frequency (e.g., formants). It has been shown, for example, that forcing partial recognition errors in one frequency region does not affect the partial recognition at other frequencies (i.e., the partial recognition errors across frequency are independent). To extract the features spread across frequency requires frequency-local signal processing, namely *independent feature-processing channels.* It seems to be this local feature-processing, uncoupled across frequency, that makes human speech recognition (HSR) robust to noise and reverberation.

The partial recognition (i.e., extracted features) are then integrated into sound units (phones), and the phones are then grouped into syllables, then words, and so forth. This model of HSR was devised by Harvey Fletcher in about 1918 in the form of an empirical probabilistic analysis of speech recognition scores obtained from a series of listening experiments. In these experiments, the context in spoken speech was selectively removed, the speech was altered by filtering and additive noise, and the phone, syllable, word, and sentence recognition scores were estimated. This analysis was first published in its full glory in 1953, four years after Fletcher's retirement from Bell Labs and 35 years after its conception. We believe that very few people have attempted to understand this publication in any depth.[1]

In the machine case, when using spectral templates, the errors across frequency are not independent. As a result, when presented with noise, filtering, reverberation, multiple speakers, and other degradations, machine systems are not robust because degradatons at one frequency affect the entire template. Thus, to reach the important goal of robust machine recognition, we need a better understand of the partial recognition of speech processing, as found in HSR.

In this paper, we will describe articulation testing, the experimental results, and Fletcher's quantitative analysis methods (the model of HSR) that he developed at Bell Labs in 1919. We will interpret and discuss the experimental data and Fletcher's model in terms of our present day knowledge of the auditory system. We then will relate this updated model to the problem of robust ASR.

II. Theory of HSR

A. History

Beginning about 1910, Campbell and Crandall from AT&T and Western Electric Engineering initiated a series of experiments to explore the nature of human speech perception. After 1918, these experiments were continued by Fletcher and his colleagues at The Bell Telephone Laboratories (Western Electric Engineering until 1925) [15]. These studies led to a speech recognition measure called the *articulation index*, which accurately characterizes speech intelligibility under conditions of filtering and noise. Fletcher used the word *articulation* in this perceptual context to mean the probability of correctly identifying nonsense speech sounds. The word *intelligibility* is defined as the probability of correctly identifying meaningful speech sounds, such as words.

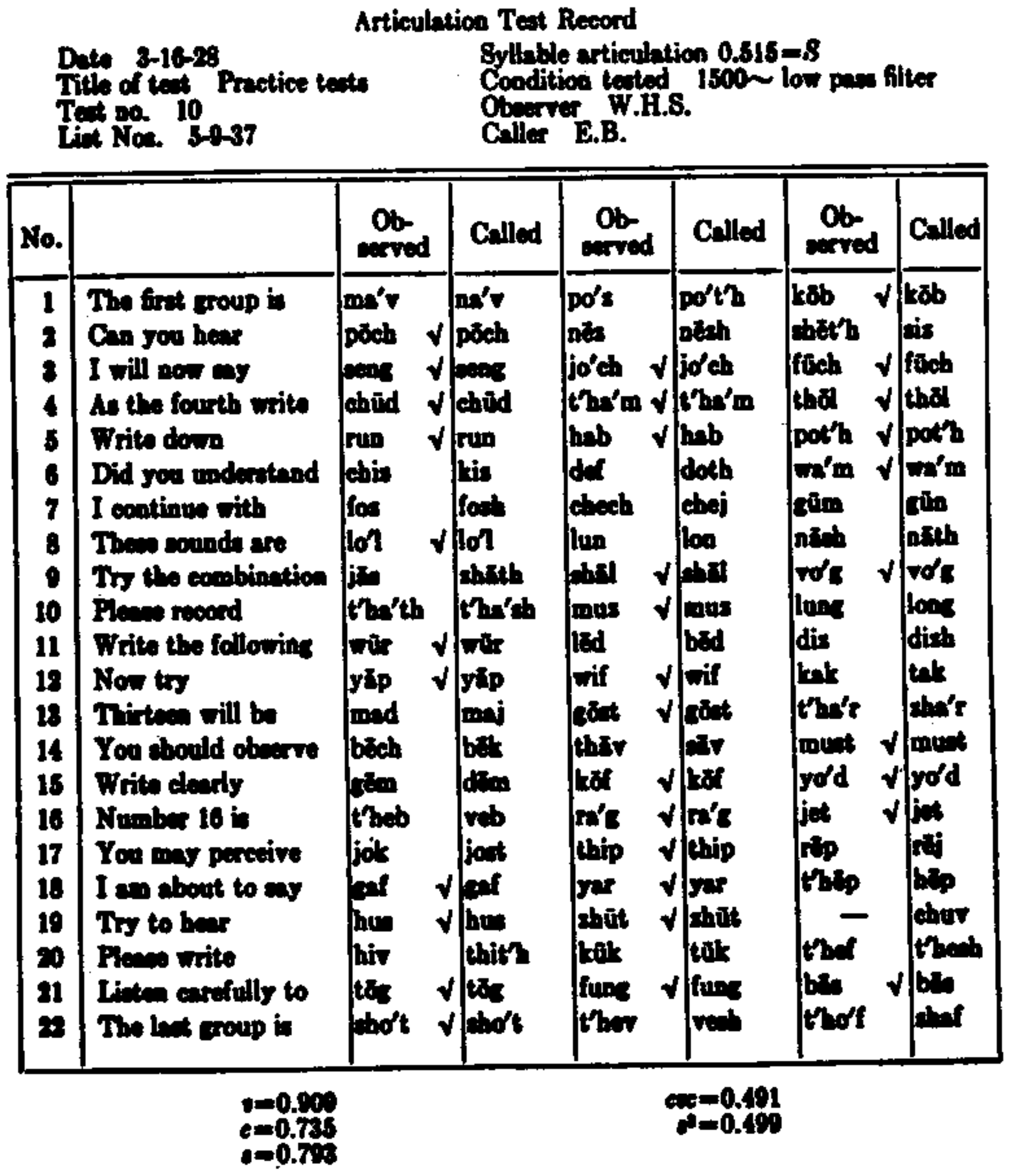
The original experiments in 1910–1918 began with normal conversational speech over a modified telephone channel. The subjects were asked to listen to distorted speech sounds and repeat what they heard. Fletcher refined the measurement methods, and by 1919 (see Fig. 1) they were using nonsense sounds (e.g., "yif," "vou," and "moush"), as described in a 1921 internal AT&T report by Fletcher entitled "An empirical theory of telephone quality." These nonsense words were

Manuscript received November 16, 1993; revised April 13, 1994.

The author is with the Acoustics Research Department, AT&T Bell Labs, Murray Hill, NJ 07974 USA.

IEEE Log Number 9403968

[1] We know of no one who has a working knowledge of the details in this difficult paper.



Articulation Test Record

Date 3-16-28
Title of test Practice tests
Test no. 10
List Nos. 5-9-37

Syllable articulation 0.515 = S
Condition tested 1500~ low pass filter
Observer W.H.S.
Caller E.B.

No.		Observed	Called	Observed	Called	Observed	Called
1	The first group is	ma'v	na'v	po's	po't'h	kōb ✓	kōb
2	Can you hear	pōch ✓	pōch	nēs	nēsh	shēt'h	sis
3	I will now say	seng ✓	seng	jo'ch ✓	jo'ch	fūch ✓	fūch
4	As the fourth write	chūd ✓	chūd	t'ha'm ✓	t'ha'm	thōl ✓	thōl
5	Write down	run ✓	run	hab ✓	hab	pot'h ✓	pot'h
6	Did you understand	chis	kis	def	doth	wa'm ✓	wa'm
7	I continue with	fos	fosh	chech	chej	gūm	gūn
8	These sounds are	lo'l ✓	lo'l	lun	lon	nāsh	nāth
9	Try the combination	jās	shāth	shāl ✓	shāl	vo'g ✓	vo'g
10	Please record	t'ha'th	t'ha'sh	mus ✓	mus	lung	long
11	Write the following	wūr ✓	wūr	lēd	bēd	dis	dish
12	Now try	yāp ✓	yāp	wif ✓	wif	kak	tak
13	Thirteen will be	mad	maj	gōst ✓	gōst	t'ha'r	sha'r
14	You should observe	bēch	bēk	thāv	sāv	must ✓	must
15	Write clearly	gēm	dēm	kōf ✓	kōf	yo'd ✓	yo'd
16	Number 16 is	t'heb	veb	ra'g ✓	ra'g	jet ✓	jet
17	You may perceive	jok	jost	thip ✓	thip	rēp	rēj
18	I am about to say	gaf ✓	gaf	yar ✓	yar	t'hēp	hēp
19	Try to hear	hus ✓	hus	shūt ✓	shūt	—	chuv
20	Please write	hiv	thit'h	kūk	tūk	t'hef	t'hesh
21	Listen carefully to	tōg ✓	tōg	fung ✓	fung	bās ✓	bās
22	The last group is	sho't ✓	sho't	t'hev	vesh	t'ho'f	shaf

$v = 0.909$
$c = 0.735$
$s = 0.793$

$cvc = 0.491$
$s^3 = 0.499$

Fig. 1. This shows an example of the articulation score sheet used in 1928. This example corresponds to a 1.5 kHz low-pass filter applied to the speech. The first column is the carrier phrase, followed by three sets of "called" and "observed" CVC's. At the top is the overall CVC score ($S = 0.515$ or 51.5% correct). At the bottom, the vowel score is shown as $v = 0.909$, namely, 90.9% of the vowels were correctly received. Similarly for the consonants, $c = 0.735$, corresponding to 75.3% correct. The average phone score, $s = (2c + v)/3$, was then computed (79.3%) along with an estimate of the CVC syllable articulation score S, assuming independent C and Vunits, as $cvc = 0.491$. Finally, $s^3 = 0.499$, corresponding to an average independent-phone model computed from the average phone articulation s.

composed of three sounds (triphones): two consonants and one vowel; and two sounds (diphones): one consonant and one vowel. The class of triphones they used are referred to as nonsense "CVC" (consonant-vowel-consonant) syllables, which account for 34% of all syllables spoken on the telephone, while the CV and VC diphones each account for 20% of the syllables. The three types of syllables account for a total of 74% (34+20+20) (p. 94 of [13]). This subset of nonsense sounds was viewed as an idealized subset of the language. The identification task on this subset was viewed as an idealized speech recognition task.

As shown in Fig. 1, the CVC's were scored for errors, both in terms of the C's and V's, and as syllable units. The errors were then described in terms of modifications to the telephone channel (channel distortion), which in turn was characterized in terms of channel bandwidth and signal-to-noise ratio. Human phone error rates for nonsense CVC's, under the best conditions, are about 1.5% error (98.5% correct) [19], [15], [13].

Fig. 2 shows an interpretation of the talker-listener experiment, modeled as an information-theoretic M-symbol noisy channel, where M is the number of phones in the language. During World War II, while Fletcher was applying the articulation index theory to pilot-navigator communications as chairman of National Defense Research Committee,[2] Shannon was studying the entropy of language and applying it to cryptography. These studies were fundamental to the development of information theory. Shannon came to Bell Labs in 1941. It is interesting that some time before 1945 and the time that Fletcher retired in 1949, Fletcher was in Shannon's line of management.[3]

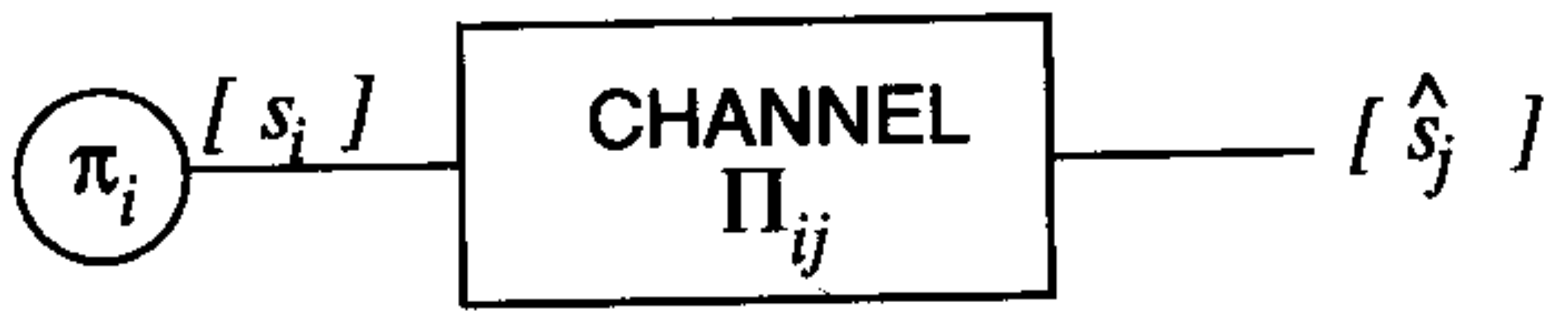


Fig. 2. The human acoustic channel is represented here as a hypothetical information-theoretic M-symbol noisy channel. Input speech sounds spoken by an average speaker are received (heard) by an average listener. An input set of spoken phones $[s_i]$, labeled by i, are spoken with probability π_i. Errors occur in transmission as described by a transition probability matrix Π having elements Π_{ij}. The set of received phones is $[\hat{s}_j]$, indexed by j.

B. Context Entropy

Context is a powerful force in speech recognition and has many *levels*. A famous example of this is demonstrated by the difference between the question "How do humans recognize speech?" and the question "How do humans wreck a nice beach?" These two sentences can be spoken so that only context can distinguish them. In a situation with the proper context however, the two could never be confused.

Context plays an important role in the functional relationship between nonsense syllable articulation and word intelligibility. This type of context results from the lower *entropy* of words relative to nonsense syllables. If one could compute the entropy of nonsense syllables $\mathcal{H}_S$, and of words $\mathcal{H}_W$, then

$$\mathcal{H}_S > \mathcal{H}_W.$$

For example, given 20 possible phones, then a triplet of CVC syllables would have a maximum entropy of $3\log_2(20) = 13$ bits (assuming a uniform distribution of the phones). It follows that the entropy of CVC words must be less than 13 bits, since "words" are the subset of CVC's that have meaning.

Context can be used across levels. Due to the meaning conveyed by a word or sentence, the listener can compensate for missing phone information. For example, the sentence *Cn yu udrstnd the sntnc*? is easily understood because of context even though all the vowels are missing. Thus, for the same channel distortion, words and sentences have higher recognition scores than those of syllables, due to their lower entropy.

Fletcher clearly understood this concept and about 1918 introduced nonsense syllables into telephone speech testing to remove the strong influence of context entropy. In doing this he greatly simplified the problem he was working on. One may study many context questions by working with full

[2] He received a citation signed by President Truman for this work.

[3] In 1945, the line of management was Fletcher, Director of Physical Research; Bode, Director; Schelkunoff, Head; and Shannon. During WWII, Shannon worked on entropy of language, which was a field closely related to Fletcher's Articulation Index. However, there is no indication that Fletcher or his work influenced Shannon's thinking in any way.

TABLE I
TABLE OF GENERAL DEFINITIONS

NAME	DEFINITION
Recognition	probability of correct sound identification
Recognition error	1-recognition
Context	indicates a subset of sounds have a prescibed meaning measured in terms of *entropy/phone*
Intelligibility	recognition of sounds having *context* (i.e., words)
Context entropy	Reduction in entropy due to contextual information is
Articulation	the recognition having no context (e.g., nonsense syllables)
Acoustic entropy	Entropy of the sounds evaluated from the confusion matrix

text or phonetic transcriptions. However, to study the difficult problem of acoustic transcription, Fletcher (wisely) decided to systematically remove various levels of context.

By 1929, Fletcher and Steinberg [21] had found a functional relation between the nonsense CVC recognition rates and word and sentence recognition rates (see also [2]–[4] and p. 226 of [10]). These relations demonstrate that *word and sentence error rates can be predicted from the nonsense* $\{C, V\}$ *error rates for a given speech corpus.* Thus Fletcher and Steinberg were the first to indirectly quantify the use of entropy in spoken language [21], [10], although they did not use these terms.

Using sentences with context complicates the intelligibility testing, decreases the efficiency of the testing, and increases the variability of the results, because context greatly confounds the measurement of phone errors.

C. Definitions

Fletcher's basic approach was to treat the speech units in terms of their empirical probabilities. As shown in Table I, he defined *intelligibility* as the empirical probability of correct recognition when context is present (e.g., recognition of words) and *articulation* as the empirical probability of correct recognition when context is *not* present (e.g., recognition of nonsense syllables) [13]. The *articulation error* is 1-articulation. A *phone* is defined as a specific sound in the language. We use square brackets to specify the phones, such as $[s_i]$, where $i = 1, \ldots, M$, assuming M phones. The phones used are defined in p. 95 of [13].

III. THE ARTICULATION EXPERIMENT

After about 1918, the databases were formed from statistically balanced nonsense CVC, CV, and VC syllable source lists. The syllable source lists were spoken and the listeners recorded what they heard. The list was chosen so that the set was closed, namely, all sounds that were recorded were also included in the source list. Articulations c and v (see Table II) were computed for the phones, and articulation S was computed for the syllables, as shown in Fig. 1. Many crews of listeners were trained to do the task, and their performance was monitored over years. Crew performance was found to stabilize within a few months.

We will assume that $[s_i]$ is the "called" phone, $[\hat{s}_j]$ is the "received" phone, M is the number of phones in the corpus, N_i the number of times $[s_i]$ is called, and $N_{ij}(\alpha)$ is the member times that $[s_i]$ is called and $[\hat{s}_j]$ is received. The speech gain α was used to vary the speech signal-to-noise ratio (SNR), and thus the errors, in a natural manner. Using standard matrix notation, i is the row index, and j is the column index. Each row corresponds to a different called phone, while each column corresponds to a different received phone. Because the set was closed, the frequency matrix N_{ij} is square.

From these definitions the i^{th} row sum $N_i = \sum_{j=1}^{M} N_{ij}(\alpha)$ is the number of times the i^{th} phone is called. The empirical probability of calling each phone is given by

$$\pi_i = N_i \Big/ \sum_{i=1}^{M} N_i. \qquad (1)$$

The lists were designed to make the called probabilities approximately equal, namely to make $\pi_i \approx 1/M$.

The empirical probability of hearing $[\hat{s}_j]$ after calling $[s_j]$ is

$$\Pi_{ij}(\alpha) = \frac{N_{ij}(\alpha)}{N_i}. \qquad (2)$$

The matrix $N(\alpha)$ defined by elements $N_{ij}(\alpha)$ is called the *frequency matrix.* The matrix $\Pi(\alpha)$ defined by elements $\Pi_{ij}(\alpha)$ is called the *confusion matrix,* or alternatively the *transition matrix.* If Π were measured for sounds with no context, we could also call it the *articulation matrix.*

Since every called phone gives rise to a received phone, the row sum is unity, namely

$$\sum_{j=1}^{M} \Pi_{ij} = 1. \qquad (3)$$

When the conditions are ideal ($\alpha = 1$), the matrix Π is close to diagonal, meaning that phone recognition is nearly perfect. When the signal-to-noise ratio was very poor ($\alpha = 0$), recognition goes to chance levels. In this case $\Pi_{ij} = 1/M$, since the observers understood that the sounds had equal probability, namely that $\pi_i = 1/M$.

A. Results

Fletcher found that the CVC syllable articulation $S(\alpha)$ (the probability of correct identification of the CVC syllable) is accurately predicted from the phone articulations $c(\alpha)$ and $v(\alpha)$ by the relation

$$S(\alpha) = c^2 v. \qquad (4)$$

This formula reflects the fact that the three sound-units are heard as independent sounds, and that to correctly identify the syllable, all three sound-units must be correctly identified.

For example, suppose the CVC is *wif.* If a listener responds with *mif,* that would be one error of the first C sound-unit and the syllable would be scored *incorrect.* To the extent that the above formula for S is accurate, it implies that coarticulations of the speech sounds are transformed by the auditory system into independent units at a early stage, before context is used, since context was not present during testing.

B. Fletcher's Measure of Recognition

To unify the computations across the various data bases, an average $\{C, V\}$ phone articulation s was computed from the composition of $\{C, V\}$ units in the data base. For example, for nonsense CVC's, Fletcher defined the empirical recognition probability of an average phone articulation as $s(\alpha) = (2c + v)/3$. In terms of (2)

$$s(\alpha) = \frac{1}{M}\sum_{i=1}^{M} \Pi_{ii}(\alpha) \tag{5}$$

which is the average of the diagonal elements of the transition matrix. When α=1, they found that $s\alpha = S_{\max} < 1$. In the case of proper (unbiased) guessing when the speech gain $\alpha = 0$, s should be $1/M$. Because of (3), the articulation error $e = 1 - s$ may also be computed from the sum of all the off-diagonal elements as

$$e(\alpha) = \frac{1}{M}\sum_{i=1}^{M}\sum_{j'=1}^{M} \Pi_{ij}(\alpha) \tag{6}$$

where the prime on the index j indicates that the sum does not include the diagonal terms along $j = i$. For the case of guessing ($\alpha = 0$), $e = 1 - 1/M$.

After some manipulation it was shown that (4) may be approximated by

$$S \approx s^3 \tag{7}$$

with only a small error [17], [15], (pp. 283–285 of [13]). This approximation depends on the experimental variation of the relative constant-vowel probability ratio $\lambda = c/v$. As shown in Fig. 3, when Fletcher plotted the syllable articulation S against the cube of the average phone articulation error s^3, he found an almost perfect agreement. A systematic error of less than 0.04 was found, as shown in the upper panel of that figure. It is not clear if or how they corrected for guessing as s becomes small (as $\alpha \to 0$).

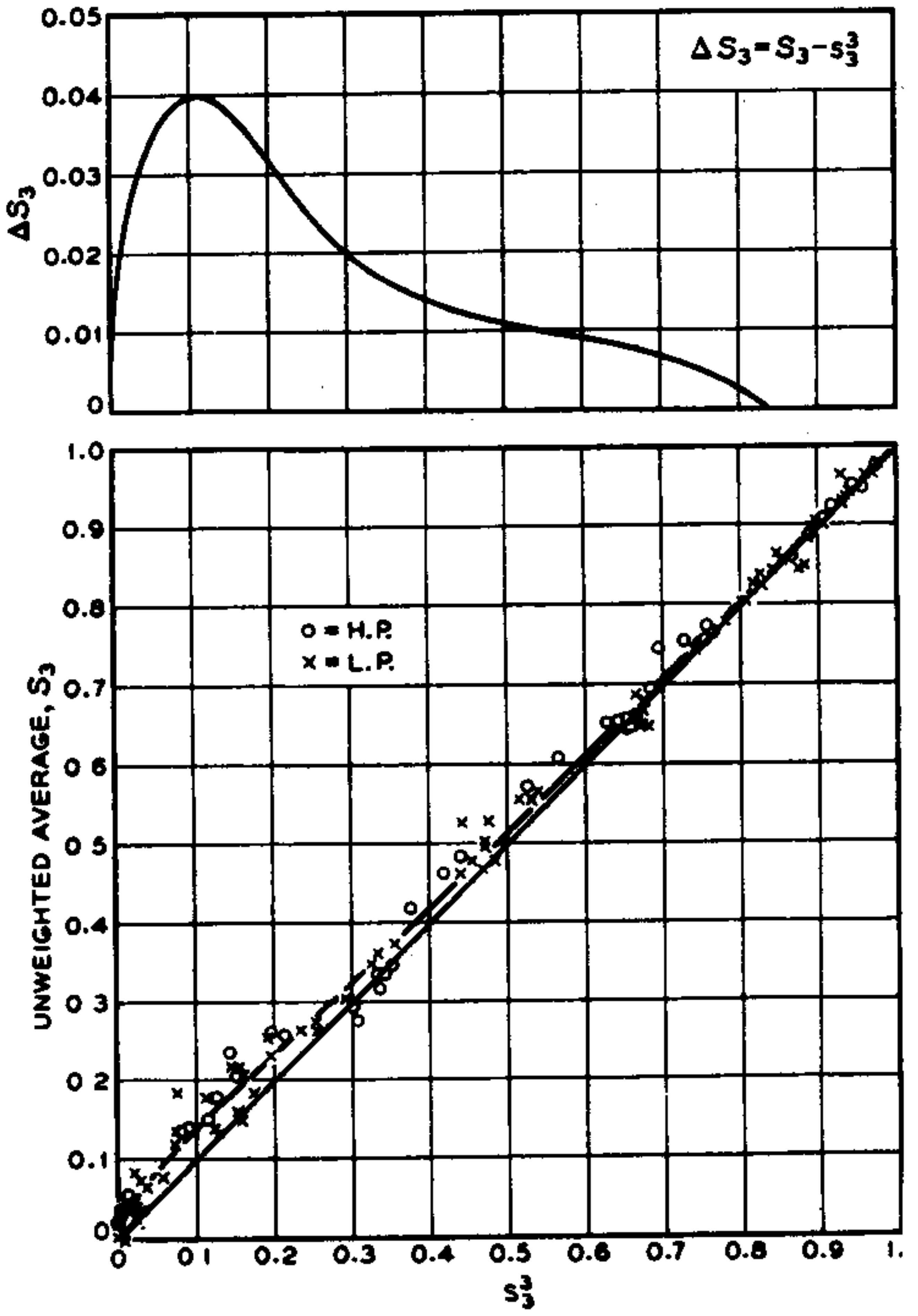


Fig. 3. This figure, which has been reproduced from p. 285 of [13], shows the relation between S (lower panel) and the error $\Delta S = S - s^3$ between S and s^3 for nonsense CVC's that have been low- and high-pass filtered. The maximum error between S and s^3 is about 0.04. It is not clear if these results were corrected for guessing (i.e., $s(\mathcal{A} = 0) = 1/M$). If the error were zero, it would mean that each CVC syllable was, on the average, perceived as three independent phone units. For further discussion, see Section II–B.

Other Measures of Recognition: Other measures of recognition besides s and S are interesting. Working with the $\{C, V\}$ sounds, it is possible to compute the *acoustic entropy* of the speech corpus. For example, we may define the *conditional entropy* given that phone $[s_i]$ was called, treating filtering, noise, and the listener as a "channel," as shown in Fig. 2. The idea here is to measure the quality of the channel, using the estimated transition matrix probabilities as a function of the physical attributes of the physical channel, namely filtering, the speech gain, and, in the case of hearing impaired listeners, the listener's hearing loss. This conditional entropy is given by the row sum

$$\mathcal{H}(\hat{s}\,|\,[s_i]) = -\sum_{j=1}^{M} \Pi_{ij}\log_2(\Pi_{ij}). \tag{8}$$

This measure is in *bits* and lies between 0 (perfect recognition) and $\log_2(M)$ (guessing). The conditional entropy of the received phone given the called phone is the expected value of (8), namely the weighted column sum over called phones

$$\mathcal{H}(\hat{s}\,|\,s) = \sum_{i=1}^{M} \pi_i \mathcal{H}(\hat{s}_j\,|\,[s_j]). \tag{9}$$

This quantity is between 0 (perfect recognition) and $\log_2(M)$ (guessing).

Coarticulation and HSR: It has long been observed that the production of a speech sound is strongly dependent on the preceding and following sound. This interdependence is called *coarticulation.* Many speech researchers believe that the coarticulation is what makes the speech recognition problem difficult.

Coarticulation is a "production" concept. From the results of the syllable articulation experiments (4), one must conclude that humans decode syllables as *independent* phone units over time. Fletcher [15], as shown in Fig. 3, and more recently Bronkhorst *et al.* [8] found that under conditions of low noise, *the phones are perceived independently.*

One might conclude that the problem of robust HSR can be split into two problems. First is the problem of decoding phones from the acoustic wave form. Second is the problem of utilizing context (e.g., entropy) to edit corrections and fill in missing information.

This leads us to the next obvious and important question, namely, "How do we decode phones?" Important insight and a major simplification of this complex problem comes from Fletcher's *articulation index.*

IV. The Articulation Index

The tight relation between the syllable and phone articulation (4) emphasizes the fundamental importance of the phone articulation to human speech recognition. Due to its basic nature, Fletcher was soon studying the phone articulation s for various channel frequency responses and channel noise [15]. To do this he used low-pass and high-pass filters on the speech.[4] In these studies Fletcher soon found that the *partial articulations* (the articulations for each band) did not sum to the wide band articulation. He then showed that a nonlinear transformation of the partial articulations would make them additive [9]. As described below, the nonlinearly transformed articulation defines an *articulation index density* $D(f)$ over frequency f. Integration, or summation, over this density gives the *articulation index* $\mathcal{A}$. The articulation index can be viewed as a fundamental internal variable of speech recognition. All recognition conditions are determined once $\mathcal{A}$ is determined for a given context entropy. (Another fundamental internal variable is the context entropy $\mathcal{H}$.)

As shown in Table II, we designate the articulations of the low- and high-pass filtered sounds as $s_L(f_c, \alpha)$ and $s_H(f_c, \alpha)$. The cut-off frequency of the filters is given by f_c, and the parameter α is the gain applied to the speech. By varying the speech level, the signal-to-noise ratio of the speech was varied. As shown in Fig. 4, s_L approaches 0.985 for f_c above 8 kHz, and 0 for f_c below 100 Hz, and s_H is 0 for f_c above 8 kHz and 0.985 for f_c below 100 Hz. Both functions are monotonic with frequency.

Fletcher [1], [9] showed that $s_L + s_H$ does not sum to s. (To motivate the discussion, we now know that for nonsense CVC's, $s = s_L + s_H - s_L s_H$, as will be shown below.) To get around this "problem" he proposed finding an invertible nonlinear transformation $\mathcal{A}(s)$ of the articulation s which he called the *articulation index* that would make the two articulation bands add to one. In other words, he wanted a transformation $\mathcal{A}(s)$ such that

$$\mathcal{A}(s_L(f_c, \alpha)) + \mathcal{A}(s_H(f_c, \alpha)) = \mathcal{A}(s(\alpha)) \qquad (10)$$

for all values of the filter cutoff frequency f_c and the speech gain α. There was, of course, no guarantee that such a transformation should exist, but his intuition suggested that it would. Since we are dealing with transformations of probabilities, the additivity condition is basically an independence argument.

He determined this transformation by finding the cutoff frequency $f_c = f_c^*$ such that

$$s_L(f_c^*, \alpha) = s_H(f_c^*, \alpha) \qquad (11)$$

which is the frequency where the curves s_L and s_H cross in Fig. 4. He then argued that the two transformed articulations must be equal at f_c^*, and therefore must each be 1/2, namely

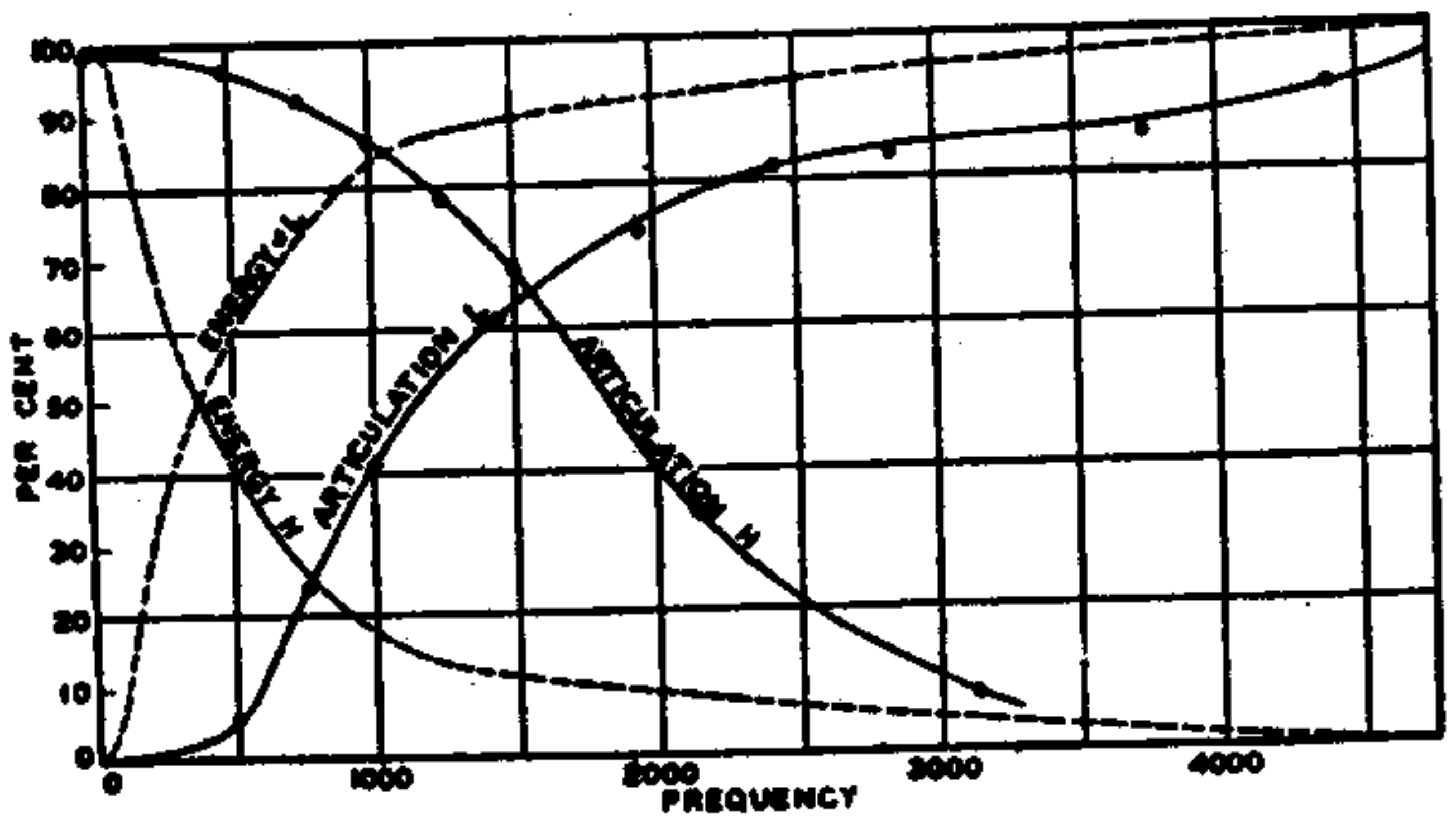


Fig. 4. This figure is reproduced from [9] and p. 280 of [10]. Speech was low- and high-pass filtered with very sharp filters having a cutoff frequency defined by the abscissa. Two things were measured for each filtered speech sound, the RMS level and the articulation. The speech energy for the two filter outputs is shown by the dashed lines and the articulations are shown by the solid lines. The curve labeled *"Articulation H"* is the same as s_H, and the curve labeled *"Articulation L"* is the same as s_L. Note how the energy curves cross at the 50% point, as they should for two sharp filters. Note how the articulation curves do not cross at 50% but at 65%. Also, the frequency of the crossover is very different for energy and articulation. The equal energy point is at 450 Hz, while the equal articulation point is at 1550 Hz.

$$\mathcal{A}(s_L(f_c^*, \alpha)) = 0.5\mathcal{A}(s(\alpha)). \qquad (12)$$

By repeating this procedure as a function of the speech gain α, he could empirically determine $\mathcal{A}(s)$, since the articulation $s(\alpha)$ is a function of the speech level.

A. What They Found

Under the conditions that the word corpus consisted of nonsense CVC's (the maximum entropy condition), Fletcher found that the nonlinear transformation that gives articulation additivity is

$$\mathcal{A}(s) = \frac{\log_{10}(1 - s)}{\log_{10}(1 - s_{\max})}. \qquad (13)$$

The constant $s_{\max} = 0.985$ is the maximum articulation and $e_{\min} = 1 - s_{\max} = 0.015$ is the corresponding minimum articulation error (p. 282 of [13]). If we solve (13) for s we find

$$s(\mathcal{A}) = 1 - e_{\min}^{\mathcal{A}}. \qquad (14)$$

Note that when $\mathcal{A} = 0$, $s = 0$, and when $\mathcal{A} = 1$, $s = s_{\max}$. This equation can also be written in terms of the articulation error $e = 1 - s$, which gives

$$e(\mathcal{A}) = e_{\min}^{\mathcal{A}}. \qquad (15)$$

B. The Independent-Channel Model

Fletcher (following Stewart) then went on to show that the phones are processed in independent *articulation bands* (frequency channels), and that these independent estimates of the speech sounds in each frequency band are "optimally" merged, as given by the following two-band example [15], [13]:

If 10 errors out of 100 spoken sounds are made when only band 1 is used, and 20 errors are made when only band 2 is

[4] It is interesting that George Campbell invented the lattice filter to do these experiments.

TABLE II
Table of Specific Symbol Definitions

SYMBOL	DEFINITION
α	gain applied to the speech
f_c	filter high- and low-pass cut-off frequency
$c(\alpha)$	consonant articulation
$v(\alpha)$	vowel articulation
$s(\alpha) = (2c + v)/3$	average phone articulation for CVC's
$e(\alpha) = 1 - s$	phone articulation error
$s_L(f_c, \alpha)$	s for low-pass filtered speech
$s_H(f_c, \alpha)$	s for high-pass filtered speech
$S(\alpha)$	nonsense syllable (CVC) articulation
$W(\alpha)$	word intelligibility
$I(\alpha)$	sentence intelligibility

used, then when both bands 1 and 2 are used simultaneously, the error is $e = 0.1 \times 0.2 = 0.02$, or two errors will be made.

For the two band example, using (10) and (13), we find

$$\log(1 - s) = \log(1 - s_L) + \log(1 - s_H) \quad (16)$$

which becomes

$$1 - s = (1 - s_L)(1 - s_H) \quad (17)$$

or in terms of the articulation error $e = 1 - s$

$$e = e_L e_H. \quad (18)$$

This equation is true for every value of f_c. The definition of $\Pi(\alpha)$ may be extended to the low- and high-pass filtered speech case as $\Pi_L(f_c, \alpha)$ and $\Pi_H(f_c, \alpha)$.

Equation (18) says that the articulation errors due to low-pass filtering are independent of the articulation errors due to high-pass filtering. I interpret this equation to mean that *we are listening to independent sets of phone features in the two bands* and processing them independently, up to the point where they are fused to produce the phone estimates. The term *feature* implies the recognition of partial information. This model does not tell us how across-channel conflicts are resolved.

C. The Articulation Index Density

As a result of the additivity required by (10), the nonlinear transformation $\mathcal{A}(s)$ transforms $s(f_c, \alpha)$ into an integral over an articulation index density $D(f)$. This follows if we let each term of (10) correspond to an integral of $D(f)$ over frequency, namely

$$\mathcal{A}(s_L(f_c)) = \int_0^{f_c} D(f) df \quad (19)$$

$$\mathcal{A}(s_H(f_c)) = \int_{f_c}^{\infty} D(f)\, df \quad (20)$$

$$\mathcal{A}(s) = \int_0^{\infty} D(f)\, df. \quad (21)$$

The density $D(f)$ may then be uniquely determined from

$$D(f_c) = \frac{\partial}{\partial f_c} \mathcal{A}(s_L(f_c)). \quad (22)$$

From these studies Fletcher was able to derive the *density over frequency of the phone articulation index* $D(f)$. This was first done in 1921. Frequencies where $D(f)$ is large carry the greatest speech information. Thus, D is called the *importance function*; it is shown in Fig. 177 and is tabulated in Table 63 on p. 333 in Fletcher's 1953 book [13], [14].

D. The Multichannel Model

Given the concept of the articulation index density, it follows that (18) may be generalized to a multichannel articulation band model, namely

$$e = e_1 e_2 \cdots e_K \quad (23)$$

where K is the number of independent articulation bands.

This model (23) was first proposed by J. Q. Stewart in 1921, but was developed by Fletcher (p. 281 of [13]). Thus, it seems proper to call it the *Fletcher-Stewart multiindependent channel* (MIC) *model of phone perception.* It is easy to show that the relation between the k^{th} band error e_k and the density is given by

$$e_k = e_{\min}^{D_k} \quad (24)$$

where

$$D_k = \int_{f_k}^{f_{k+1}} D(f)\, df. \quad (25)$$

The frequency limits f_k were chosen so that all the D_k's were equal, which means that under optimum listening conditions (α near 1), $D_k = 1/K$.

The number of bands K is frequently taken to be 20, which makes each band correspond to 1 mm along the basilar membrane. Since Fletcher identified a critical band to be about 0.5 mm along the basilar membrane, one articulation band represents two critical bands. However, I believe that the number K was chosen for convenience, and should not be taken as a particularly significant number. It has been reported, for example, that 10 bands is too few, and 30 bands gives no improvement in accuracy over 20 bands.

It was first observed by Galt, working with Fletcher, that the equal spacing of the articulation index density function *approximately corresponds to equal spacing along the basilar membrane since (19) is very similar to the cochlear map function, which is the relation between normalized place X on the basilar membrane and characteristic frequency F(in Hz) along the basilar membrane.* The normalized place variable is defined as $X = (L - x)/L$, where x is in mm from the stapes, and $L = 35$ mm is the length of the basilar membrane. It is frequently expressed as a percent (p. 293 of [13]). From the Greenwood human cochlear

$$F(X) = 165(10^{2.1X} - 0.88) \quad \text{(Hz)} \quad (26)$$

we know that the distance along the basilar membrane, between 300 Hz and 8 kHz, is 20 mm. Thus, there is about one articulation band/mm corresponding to about 4000/35=114 hair cells or about 1140 neurons.

The slope of the cochlear map dF/dX was found to be proportional to the *critical ratio* $\kappa(f)$ [1], [11]–[13]. The critical ratio $\kappa(f)$ is an important psychophysical measure of the relative bandwidths of our cochlear filters [1], [13]. The ratio

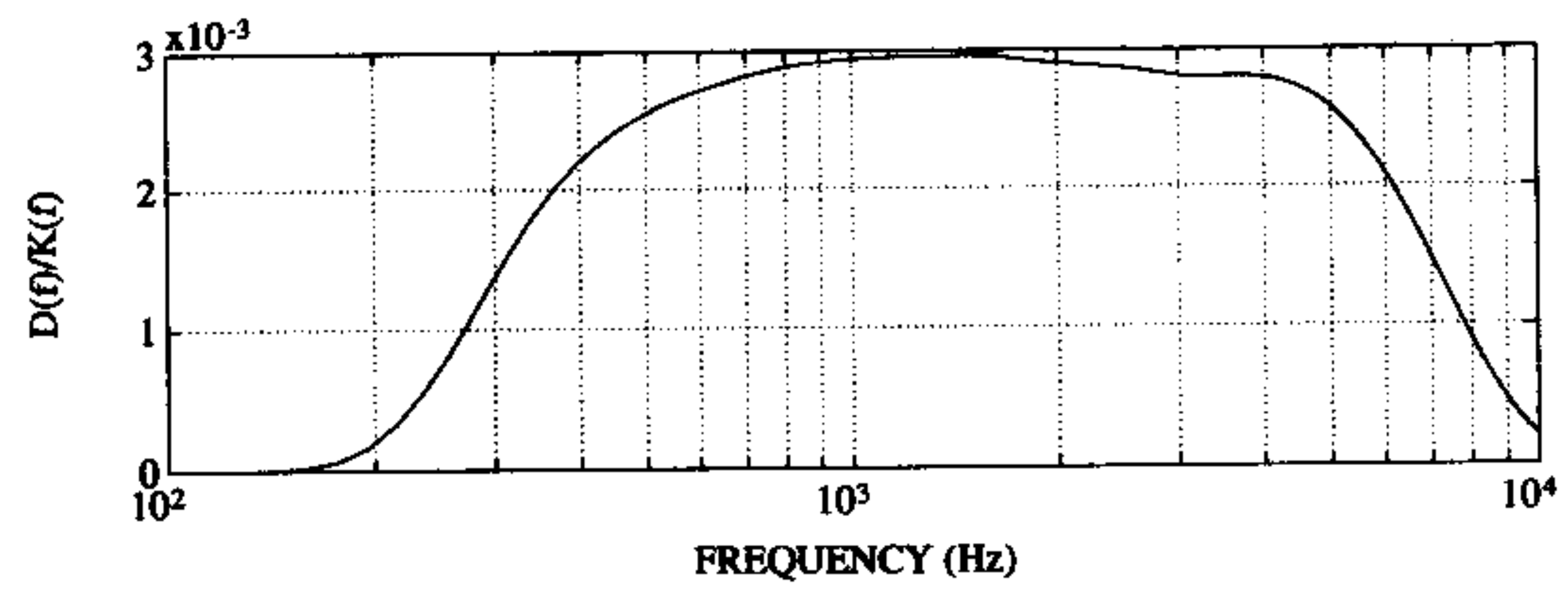


Fig. 5. Figure showing the ratio of $D(f)/\kappa(f)$. This ratio is a measure of the phone articulation per critical band. The ratio has been normalized so that its area is 1.

of $D(f)$, the articulation index density, and $\kappa(f)$, the critical ratio, is a measure of the relative speech articulation/mm or per critical band. As shown in Fig. 5, this ratio is approximately uniform over the speech band. These results have been scaled so that $\int D(f)/\kappa(f) df = 1$.

A Physical Measure of the Articulation Errors: [21] and [19] went on to show that the signal-to-noise ratio expressed in dB, in each critical band, normalized to 30 dB, determines the band articulation index D_k in band k corresponding to the band articulation error e_k. This relation is given by

$$D_k(\alpha) = \frac{1}{K}\mathrm{SNR}_k(\alpha)/30. \tag{27}$$

The band signal-to-noise ratio SNR_k is set to zero when it becomes negative, and is set to 30 when it is greater than 30. Thus the articulation index depends on the signal-to-noise ratios in each band rather than the speech energy spectrum.

Articulation Bands Versus Critical Bands: It is important to clarify the difference between articulation bands and critical bands. Critical bands represent filtered and neurally encoded signal intensity, and are related to the *partial loudness* of the speech as a function of the position along the basilar membrane X. The partial loudness is sometimes called the *neural excitation pattern* [1]. Articulation bands, on the other hand, are a measure of *partial phone recognition*, as a function of a tonotopic axis similar, but to equal, to X. Thus a very important transformation has taken place between the critical band signal and the articulation band signal, namely the *neural representation of signal intensity has been transformed into a measure of partial recognition.* We must not assume that this is a trivial transformation. If it were, robust ASR would have been achieved many years ago. It is worth remembering that Fletcher discovered both the critical band and the articulation band, and nowhere did he suggest equating them.

E. The Intelligibility of Words

When words are used instead of nonsense syllables, the syllable errors must be further transformed to determine the word intelligibility $W(\mathcal{A})$. This case represents a decrease in the speech entropy. These methods were partially worked out by Fletcher and Steinberg [9], [18], [15]. Boothroyd described similar results using the empirical expression

$$W(\mathcal{A}) = 1 - (1 - S(\mathcal{A}))^j \tag{28}$$

where the constant $j > 1$ depends on the entropy of the word corpus and may be empirically determined [2]–[4].

F. Summary

In summary, for speech having a signal-to-noise ratio of SNR_k dB, where k labels the frequency bands

$$D_k(\alpha) = \frac{1}{K}\mathrm{SNR}_k(\alpha)/30 \tag{29}$$

$$\mathcal{A}(\alpha) = \sum_{k=1}^{K} D_k(\alpha) \tag{30}$$

$$s(\mathcal{A}) = 1 - e_{\min}^{\mathcal{A}} \tag{31}$$

$$S(\mathcal{A}) = s^3 \tag{32}$$

$$W(\mathcal{A}) = 1 - (1 - S(\mathcal{A}))^j. \tag{33}$$

When $\mathrm{SNR}_k(\alpha)/30$ is less than 0 it is set to 0, and when it is greater than 1, it is set to 1. The constant j is greater than 1.

This model has been tested with hundreds of combinations of channel parameters and is impressively accurate over a large range of channel conditions [15], [19].

V. The Recognition Chain

Fletcher's analysis defines a heuristic model of human speech recognition as a layered hierarchy, as shown in Fig. 6. The acoustic signal enters the cochlea and is broken into frequency bands (critical bands) which define the signal-to-noise ratios SNR_k, where k labels the cochlear frequency channel. There are about 4000 inner hair cells along the basilar membrane corresponding to a heavily overlapped set of cochlear filters. These outputs are then processed by the first "layer" which defines the phone features represented by the partial articulation errors e_k, as given by (29) and (24). Usually, $K = 20$ of these bands are assumed, corresponding to 1 mm each along the basilar membrane, or one or two critical bands. The next layer defines a phone space, measured as articulations s found from (31). There are about 20 phones per $\{C, V\}$ unit. The phones are then transformed into syllable units having articulation $S(a)$ (32) and then into words with intelligibility $W(s)$ (33). The approximate number of nonsense CVC's is about 8000. The number of CVC words is much less, and could be estimated using data from the tables of Chapter 5 of [13]. A plot of typical values for these articulation measures is shown in Fig. 7.

The fact that we are able to recognize nonsense words and sentences without difficulty makes it unlikely that feedback is common or significant between the deeper layers and the outer layers. Furthermore, the delay involved in any feedback mechanism could create serious "real-time" problems. However, this interesting question is open at this time, since we really have very little evidence to guide us.

A. Implications for Modern Machine Speech Recognition

Fletcher's articulation index studies and models have important implications to ASR. Typical ASR systems start with a "front-end" that transforms the speech signal into a "feature vector" which is then processed by a "back-end" classifier. These systems frequently place a heavy emphasis on word and language models as a method of increasing the recognition scores.

Because of confusion and misunderstanding based on coarticulation arguments, only a small amount of research has been

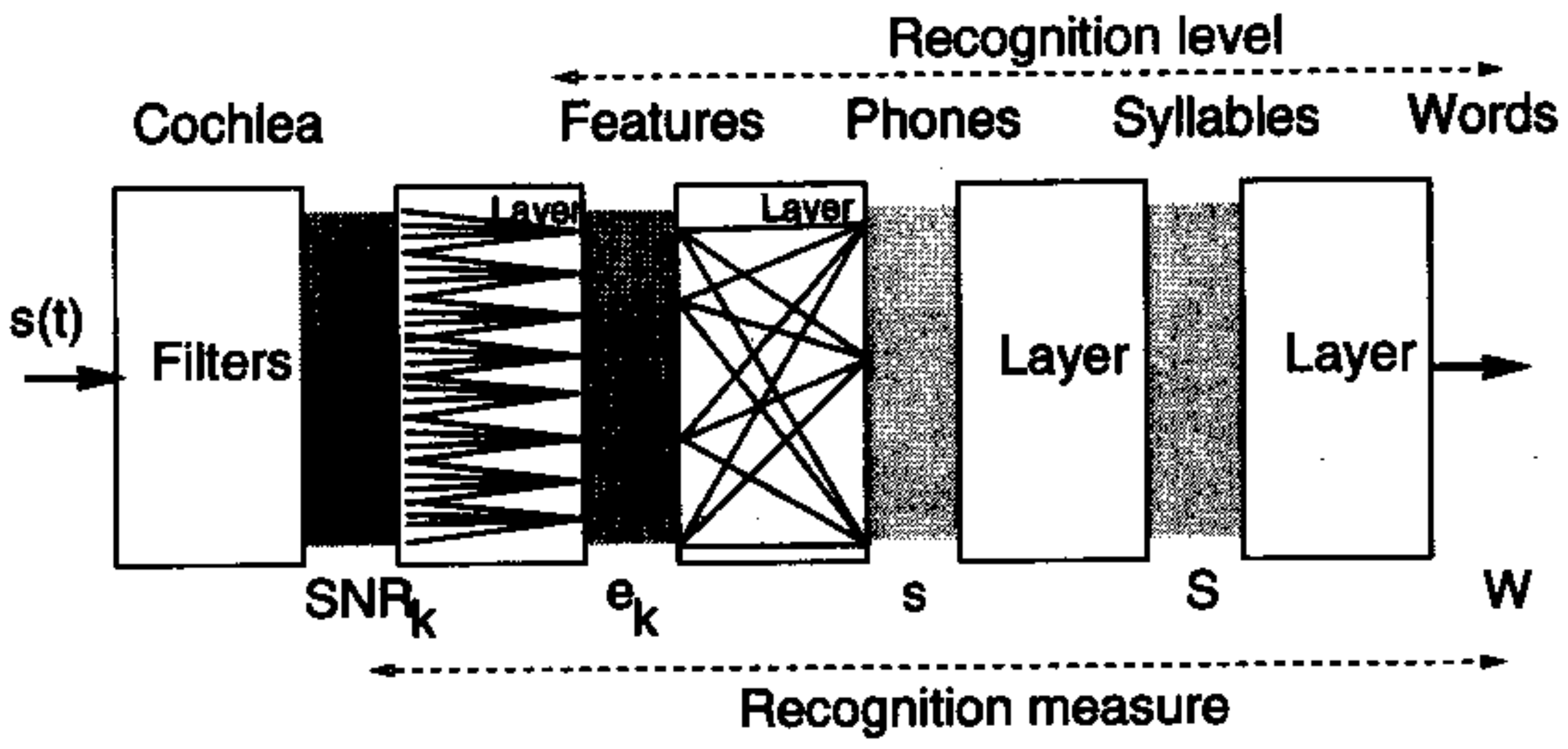


Fig. 6. Hypothetical cascade of recognition layers, starting with the cochlea. The articulation measures shown at the bottom are defined in Table II. The words along the top describe the physical correlate of the measure. No feedback is assumed between layers in this oversimplified model of HSR. The first layer, the cochlea, determines the signal-to-noise ratio in about 2800 overlapping critical band channels. The next layer extracts features (i.e., partial recognition) from the speech in a local manner, as indicated by the network wiring. The output of this layer is measured in terms of the $K = 20$ or so feature errors e_k. Next, the features are mapped onto the $M = 20$ or so phones. This process necessarily integrates across the entire tonotopic axis. Then syllables and words are formed.

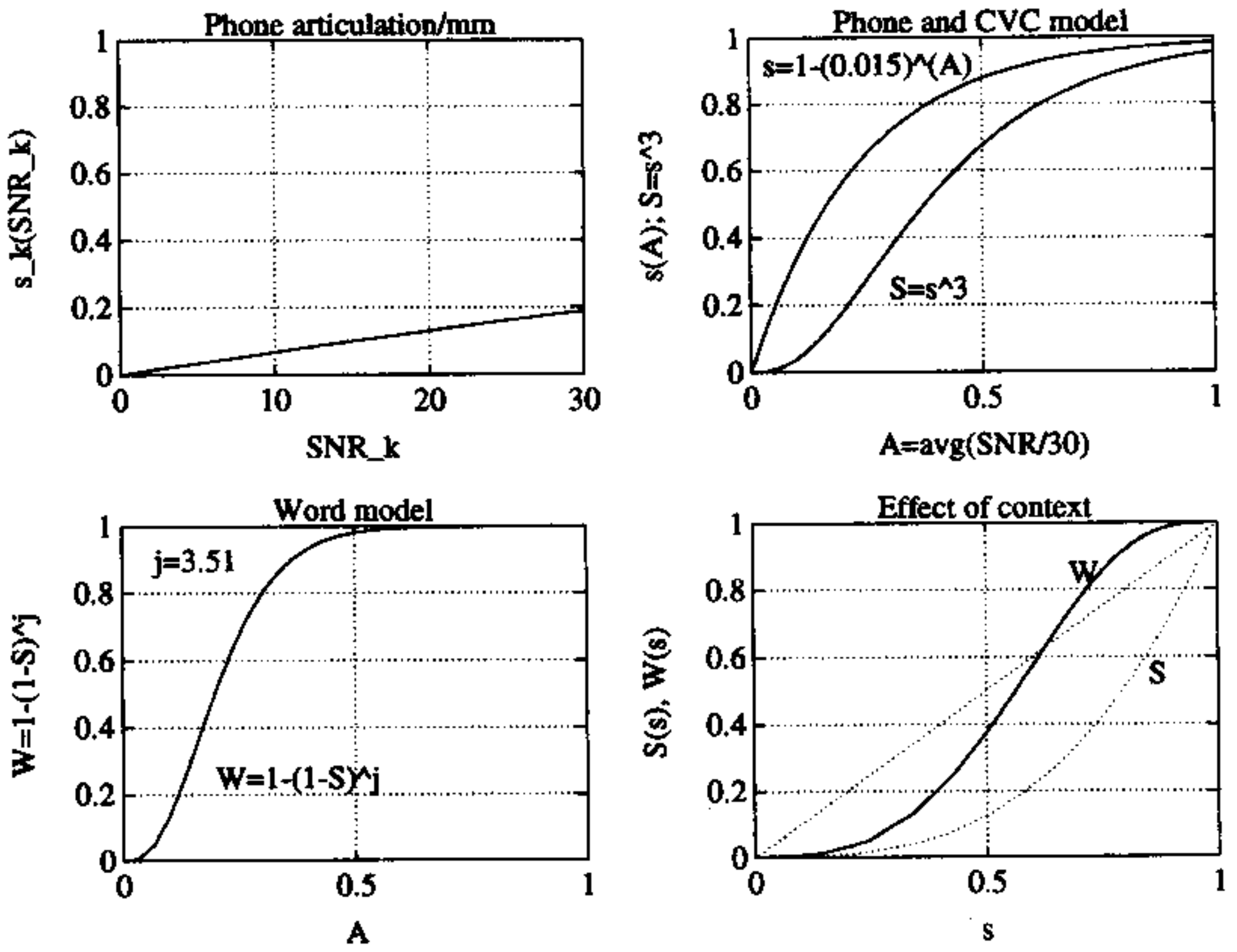


Fig. 7. The partial phone articulation density $s_k = 1 - e_k$ along the basilar membrane is shown in the upper-left plot, as a function of the signal-to-noise ratio SNR_k in dB. The partial articulation density is in units of probability of correct identification of nonsense phone units per millimeter along the basilar membrane. The phone articulation s is shown as a function of the average SNR/30 in the upper-right plot along with the nonsense CVC syllable articulation S. When meaningful sounds (words) are used, the effects of context must be taken into account. This effect is shown in the lower-left plot. One may also plot one variable against the other, as shown in the lower-right plot.

done on the automatic recognition of nonsense CVC's. From the work of Fletcher and Steinberg, it should be clear that the real challenge in machine recognition today is human-like performance for phones and nonsense CVC under conditions of typical channel distortions. Since the human performance is well-known under these conditions [15], [19], nonsense CVC's represent an excellent database. Decreasing the error rates for these elementary signals would have a major impact on overall system performance and robustness.

B. Average Phone Entropy

If we use the phone frequencies given by Fletcher (p. 95 of [13]) and calculate the phone entropy, we find 4.3 bits for the initial consonant C_i, 4.5 bits for the V, and 4.1 bits for the final consonant C_f. The average phone entropy is then 4.3 bits/phone, or $M = 2^{4.3} = 19.7$ possible phones, on the average. This entropy would be relevant for telephone speech when using a phone recognizer front-end that treats the phone string as context free. The entropy would drop as context effects are included.

C. What Is the Nature of the Phone Feature Space?

In this section we should like to speculate on the possible nature of the phone feature space, given what we know about the auditory system. From this point on I shall use the term *feature* to mean 1 bit of partial recognition. This definition seems consistent with the notion that a feature is a binary concept, namely it describes something that is, or is not, present. If M objects are to be described, then it takes $\log_2(M)$ features to describe them. If we interpret the bits computed from the phone entropy as features, then there are 4.3 features/phone, on the average.

It is common practice to use 20 articulation bands ($K = 20$), where each band corresponds to 1 mm along the basilar membrane. Thus, on the average, there are $4.3/20 = 0.215$ features/mm. Perhaps a more intuitive way of expressing this is that there are $20/4.3 = 4.65$ mm/feature. Since 1 octave corresponds to about 5 mm along the basilar membrane, the average phone feature density is about 1 feature/octave and the average feature length is 1 octave/feature.

We know that the cochlea (as well as the eye) breaks the external world down into a tonotopic array of critical bands (pixels). It is then the job of the CNS to "reconstruct" the scene from these pieces. There is evidence that this is done by making a huge cascade of local binary decisions. Much of this decision-making process is done at a subconscious level. For example, in a visual scene, the edge of a piece of paper is seen to be continuous, not because it is continuous, but because that is a rational decision to make given the local pixel inputs. Each time the elements are found to be correlated, they are fused together (e.g., the edge of the paper looks continuous; a musical chord sounds as one; a voice in noise is fused as a unit, independent of the noise).

From this point of view, the actual number of articulation bands ($K = 20$) is unimportant as long as they are not underrepresented. What *is* important is the feature length along the tonotopic axis as defined by the local feature extractors. Assigning probabilities to the detection of (binary) features provides a natural and unique reduction of the data from 4000 × 20/35=2300 narrow band neural channels to 4.3 (on the average) tonotopic feature regions of various lengths, that depend on the specific input phone. In this model, the number of correlated regions determines the dimension of the space and the length of the correlation determines the coordinates (e.g., the probability of the feature being present). This model is a natural generalization of Fletcher's (and Munson's) model of tonal loudness which was the first model to propose the idea of a neural excitation pattern [1].

VI. Discussion

Somehow the early CNS forms independent error estimates of features across frequency, identifies incomplete information and conflicts, resolves the conflicts, and then merges the resulting feature information. For example, if noise fills the upper band and clean speech fills the lower band, the articulation will be the identical to the case of sharply low-pass filtered speech having no energy in the upper band. The presence of noise in the upper band confounds a template-based approach which does not treat the frequency channels as independent. Since $\mathcal{A}$ depends on the SNR rather than the energy spectrum, filtering the speech does not change the articulation unless the filtering reduces the SNR below 30 dB. It would seem that the concept of a "perceptual norm" is not only an unobtainable dream, it is also a highly suboptimal processing strategy. It is surprising that there has been so little discussion about the significance of these articulation formulae in the speech recognition literature.

The partial recognition errors (articulation channels) were found to be independent (23). This implies that the CNS is processing the cochlear channels over time and extracting independent features across frequency. Estimates of the feature density give one feature per 5 mm, corresponding to 1 octave in frequency. One octave is about 5 to 10 critical bands since each critical band is between 0.5 and 1 mm. Only after the features are detected are they merged into estimates of the phones.

The extraction of frequency-local independent features might be done by finding correlations between auditory (critical band) channels. For example, when noise is added to the speech and the SNR across frequency changes, a correlation measure between bands will decrease in a manner that is consistent with the phone articulation reduction. In such a model, a group (114 or 1140, depending on whether you count inner hair cells or neurons) of critical band filter channels would be grouped by the CNS to form one of the $K \approx 20$ independent articulation channels.

This view is most strongly supported by "comodulation release from masking" (CMR) experiments which demonstrate that correlations of neural envelope signals are computed by the auditory system. These CMR experiments measure the detection threshold of a tone in noise when a flanking band of "comodulated" noise is present in the same ear. For example, suppose a band of noise from 1–1.4 kHz is amplitude modulated by a 20 Hz low-pass noise. A tone is presented in the center of the band at 1.2 kHz. The tone is increased in level until it is detected by the subject 75% of the time, and the level of the tone is noted. Then a second band of noise, say from 1.6–2.0 kHz is presented, again amplitude modulated by the same 20 Hz low-pass noise. When the second noise is added to the stimulus, the tone at 1.2 kHz is easily heard, well above threshold. The CMR effect shows that the CNS "correlates" the two bands of comodulated noise, and as a result "discovers" that a tone is present in the lower band. To do this it must identify that the bands are correlated, and recognize that the tone, present in the lower band, reduces this correlation!

A. Effect of the Cochlear Bandwidth at High Levels

At high sound levels the articulation s "rolls over" (it decreases). This effect was modeled by Fletcher as a change in the critical bandwidths of the ear, as measured by the ratio of the level of a tone to the spectral level of wide band noise at the tone's detection threshold (i.e., the critical ratio $\kappa(f)$) [16], [11], [12]). Above about 65 dB SPL the critical bandwidth increases by a small amount [12], [19], [15], [13]. An analysis of the magnitude of the cochlear bandwidth increase is complicated by the uncertainty of the signal-to-noise ratio at the output of the cochlear filters at the signal detection threshold [1]. The increased bandwidth of the cochlear filters leads to a reduction in the frequency resolution of the ear and therefore increased channel correlation under all signal conditions. Reduced speech recognition performance follows. *This implies that the actual cochlear filter bandwidths of the ear may be an important variable if we are to attain human-like performance in ASR.*

B. Across-Time Versus Across-Frequency Processing

The template-based approach used in ASR could be called an *across-frequency* processing scheme. It appears that HSR is solved using an *across-time* processing scheme, with only local coupling across frequency. There is some evidence for this. First, the articulation channels are independent. Second, the human listener is quite insensitive to dispersive (frequency dependent) delay, such as all-pass filters. This famous fact is frequently referred to as "Ohm's Law of Acoustics," which claims that the ear is *phase-deaf.* Room reverberation is an important form of degradation that is an example. The reverberation time in a room must reach at least 0.3 to 0.5 seconds before one is even aware of its presence, and must be in the range of seconds before it becomes a degradation to speech communication. Reverberation is typically very frequency dependent, with only short delays at high frequencies, and long delays at low frequencies. With the feature extraction done along time rather than across frequency, the system is much less insensitive to this common type of frequency-dependent degradation.

Coarticulation: Across-time processing may also resolve the paradox of coarticulation which results from trying to assign each phone a spectral template. When one tries to associate a spectral template to a particular sound, one is assuming (incorrectly) that the timing of the features must be synchronous. From psychophysical experiments, we know that under many conditions, our ability to perceive the magnitude (and even the relative order) of temporal events can be very poor. Phone recognition is most certainly not the synchronous timing of feature events, but some more abstract relation between the presence and absence of features, and their geometrical relations in a multidimensional feature space [6]. This transformation may be viewed as a form of "categorical perception" [20].

VII. Summary

How do humans process and recognize speech? (Remember the rhetorical title of this paper?) We are still looking for the answer, and Fletcher's experiments and analysis tell us where to look. The most elementary auditory speech processing model (Fig. 6) is a cascade of the cochlea followed by the following recognition layers: features, phones, syllables, words, sentences, meaning, etc. The basis of this model is the recognition data for the various context levels.

The most important problem in HSR for those interested in ASR is feature and phone perception because this is the part of the system that goes from an acoustic signal to the most basic speech recognition element. The speech sounds are divided into a time-frequency continuum of feature-bearing frequency bands by the cochlea. There are about 4.3 independent binary features represented along approximately 20 mm of basilar membrane. These feature channels form the basis for the articulation channel errors e_k. The bands are processed in such a way as to robustly extract and isolate the $\approx$ 20 possible elemental sounds for each phone time slot (each C or V) (i.e., 4.3 bits/phone).

Equations (10) and (13) indicate that the articulation error *information* $\mathcal{I}$ defined by

$$\mathcal{I}(e) = \log_2(e)$$

is additive and defines a tonotopic density because the partial recognition errors are independent (23).

The signal-to-noise ratio of each cochlear inner hair cell signal is important to the formation of the feature channels since e_k is known to depend directly on these SNR's rather than on the spectral energy. There are many more articulation bands than features, and we have estimated that each feature, on the average, occupies about 5 mm (1 octave) along the basilar membrane. The model is consistent with the idea of using correlations between neighboring cochlear channels to form the output of the feature layer. If two filter bands were not correlated due to a poor SNR, then the correlator output would be small. These correlations are undoubtedly generated in an early processing stage and form a very basic processing system. The auditory system then fuses these features into units (phones). This fusion is called an *auditory stream*, which is the subject of the book *Auditory Scene Analysis* by Bregman [7].

To understand how speech is recognized it is necessary to systematically control context factors since context is very important. This was done by working with a database of balanced nonsense CVC, CV, and VC syllables. Syllables having context decreases the speech entropy. We know the relation between the phone and nonsense syllable error rates (4). The phones may be represented as a multidimensional features space [6] leading to the idea of the categorical perception of these units [20].

At each layer, the entropy is decreased as meaning is utilized and extracted. By removing meaning or context from the speech, we may effectively disable the processing for that layer. This allows us to quantify the layer's utilization of the speech's entropy. Using this method, the recognition probability $W(S)$ between the words and nonsense syllables, due to the word intelligibility, has been empirically estimated. The same technique has been applied to quantify meaning in sentences.

Since Fletcher's theory only attempts to predict the average articulation, it does not address the important question of exactly how the articulation channel signals are processed to form the sound-unit recognition. An interesting clue may be provided by the "McGurk effect," where visual features *dominate* those of the auditory channel. This might be viewed as a visual side channel that has an input to certain terms in (23) by contributing information to the formation of that subset of features that are correlated to the lip or jaw movements. Alternatively the visual input might directly contribute to the feature space of phones. Braida is exploring a multidimensional perceptual space model in an attempt to model the articulation index in terms of what is happening in the CNS [5], [6].

Finally, these measures provide an important knowledge database against which we may benchmark the machine recognizers, to drive their performance toward that of the human listener.

Acknowledgment

The author would like to thank R. Bakis, E. Gilbert, A. Orlitsky, and J. Lagarias for helpful discussion, A. Rosenberg, D. Roe, and D. Berley for comments, and J. Flanagan for his very strong encouragement, without which this work could not have been started or finished.

References

[1] J. B. Allen, "Harvey Fletcher 1884–1981," in *The ASA Reprint of Speech and Hearing in Communication*, J. B. Allen, Ed. New York: Acoustical Society of America, 1994.

[2] A. Boothroyd, "Statistical theory of the speech discrimination score," *J. Acoust. Soc. Am.*, vol. 43, no. 2, 1968.

[3] ——, "Speech preception, sensorineural hearing loss, and hearing aids," in *Acoustical Factors Affecting Hearing Aid Performance*, G. A. Studebaker and I. Hochberg, Eds. Boston: Allyn and Bacon, 1993, pp. 277–299.

[4] A. Boothroyd and S. Nittrouer, "Mathematical treatment of context effects in phoneme and word recognition," *J. Acoust. Soc. Am.*, vol. 84, no. 1, 1988.

[5] L. D. Braida, "Crossmodal integration in the identification of consonant segments," *Quarterly J. Exper. Psychol.*, vol. 43A, no. 3, 1991.

[6] ——, "Integration models of speech intelligibility," in *Speech Communication Metrics, and Human Performance*, 1993, 1–20 Washington, DC, NAS-CHABA.

[7] A. S. Bregman, *Auditory Scene Analysis, The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1990.

[8] A. W. Bronkhorst, "A model for context effects in speech recognition," *J. Acoust. Soc. Amer.*, vol. 93, no. 1, pp 499–509, Jan. 1993.

[9] H. Fletcher, "The nature of speech and its interpretation," *J. Franklin Instit.*, vol. 193, no. 6, pp.729–747, June 1922.

[10] ——, *Speech and Hearing*. New York: Van Nostrand, 1929.

[11] ——, "Loudness, masking and their relation to the hearing process and the problem of noise measurement," *J. Acoustic. Soc. Amer.*, vol. 9, pp. 275–293, Apr. 1938.

[12] ——, "The mechanism of hearing as revealed through experiments on the masking effect of thermal noise," *Proc. Nat. Acad. Sci.*, vol. 24, pp. 265–274, 1938.

[13] ——, *Speech and Hearing in Communication*. New York: Krieger, 1953.

[14] ——, "Speech and hearing in communication," in *The ASA Edition of Speech and Hearing in Communication*, J. B. Allen, Ed. New York: Acoustic. Soc. Amer., 1994.

[15] H. Fletcher and R. H. Galt, "Perception of speech and its relation to telephony," *J. Acoustic. Soc. Amer.*, vol. 22, pp. 89–151, Mar. 1950.

[16] H. Fletcher and W. A. Munson, "Relation between loudness and masking," *J. Acoustic. Soc. Amer.*, vol. 9, pp. 1–10, 1937.

[17] H. Fletcher and J. C. Steinberg, "Articulation testing methods," *Bell Syst. Tech. J.*, vol. 8, pp. 806–854, Oct. 1929.

[18] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoustic. Soc. Amer.*, vol. 19, pp. 90–119, 1947.

[19] K. P. Green *et al.*, "Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect," *Perception, Psychophys.*, vol. 50, no. 6, pp. 524–536, 1991.

[20] J. C. Steinberg, R. H. Galt and A. B. Anderson, "Study of speech and hearing - case 20871-3," *Bell Syst. Tech. Memorandum*, vol. 328, no. 67, pp. 1–36, I:1–17, II:1–16, III:1–9, Figs 1–115, 1937.

[redacted]

Jont Allen (M'76–SM'79–F'85) received the B.S. in electrical engineering from the University of Illinois in 1966, and the Ph.D. from the University of Pennsylvania in 1970.

He joined Bell Laboratories in 1970, where he is in the Acoustics Research Department as a Distinguished Member of the Technical Staff.

Dr. Allen is a Fellow of the Acoustical Society of America (ASA) and is a past member of the Executive Council of the ASA. He is a past member of the Administrative Committee of the Acoustics, Speech, and Signal Processing Society of the IEEE, served as Editor of the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, Chairman of the Publication Board of the ASSP Society, and General Chairman of the International Conference of the ASSP (ICASSP-1988). In 1986 he was awarded the IEEE ASSP 1986 Meritorious Service Award. In 1986–88 he participated in the development of the AT&T multiband compression hearing aid, now sold under the Resound name. In 1990 he was an Osher Fellow at the Exploratorium Museum in San Francisco. In 1991–92 he served as an International Distinguished Lecturer for the Signal Processing Society. In 1994, he was a Visiting Scientist and Lecturer at the University of Calgary.