THE APPLICATION OF INFORMATIONTHEORY TO VOWEL RECOGNITION EXPERIMENTS.

This paper deals with the application of Information theory to the transmission of natural and artificial vowels. I want to start with a brief explanation of the main concepts of information theory for those not acquainted with them.

Information theory describes the phenomena of transmission as perceived by an outside observer who has full knowledge of both sides of the transmission channel.

The symbols to be coded by the transmitting part of the channel - in our case in sounds - will be referred to as input, the symbols decoded at the receiving end as output.

In this case we are only interested in the most simple situation in which the autocorrelation of the string of input symbols is zero, which means that the input symbols are in a random order. Incidentally this does not exclude the possibility that the decoding process is affected by the actual succession of two or more sounds. The number of different symbols will be finite.

Due to imperfections or instability of transmitter and receiver, distortion or interference. the string of output symbols will not be an exact replica of the string of input symbols. We speak therefore of a transmission channel with noise.

Let the number of different symbols be n .

The symbols can then be referred to as S_1 , S_2 ...... S_n .

The performance of the channel can be depicted by a table of confusion probabilities. (See fig. 1)

CONFUSION PROBABILITY MATRIX

INPUT		OUTPUT			
		S_1	S_2	S_k	S_n
S_1	p_{10}	p_{11}	p_{12}	p_{1k}	p_{1n}
S_2	p_{20}	p_{21}	p_{22}	p_{2k}	p_{2n}
S_j	p_{jo}	p_{j1}	p_{j2}	p_{jk}	p_{jn}
S_n	p_{n0}	p_{n1}	p_{n2}	p_{nk}	p_{nn}
Total	1	p_{01}	p_{02}	p_{0k}	p_{0n}

fig. 1

p_{j0} = probability that S_j is the input symbol,

p_{0k} = probability that S_k is the output symbol,

p_{jk} = probability of the combination of S_j as input symbol and S_k as output symbol.

In a forced-choice situation

$$\sum_j p_{j0} = \sum_k p_{0k} = \sum_{j,k} p_{jk} = 1$$

For a noise-free channel

$$p_{j0} = p_{jk} = p_{0k} \qquad \text{for } j = k$$

$$p_{jk} = 0 \qquad \text{for } j \neq k$$

For a channel with no correlation between input and output (that means no transmission at all, the receiver is only guessing)

$$p_{jk} = p_{j0} \cdot p_{0k} \qquad \text{for } j = 1, \ldots\ldots n; \; k = 1, \ldots\ldots n$$

A real channel will be somewhere between these extremes. Now we have to deal with different amounts of information. The information of the input H_x, that of the output H_y and that of the combination of input and output H_{xy}.

The unit of information is called a Bit.

One bit is the amount of information contained in the answer to a question to which there are two mutually exclusive answers with equal probability of occurrence. Take for example the information contained in the position of a coin. So the amount of information in Bits is the minimal number of questions of the type just mentioned necessary to obtain full knowledge. The amounts of information can be easily calculated using the formulae of fig. 2.

$$H_x = \sum_j - p_{j0} \, {}^2\log p_{j0}$$

$$H_y = \sum_k - p_{0k} \, {}^2\log p_{0k}$$

$$H_{xy} = \sum_{j,k} - p_{jk} \, {}^2\log p_{jk}$$

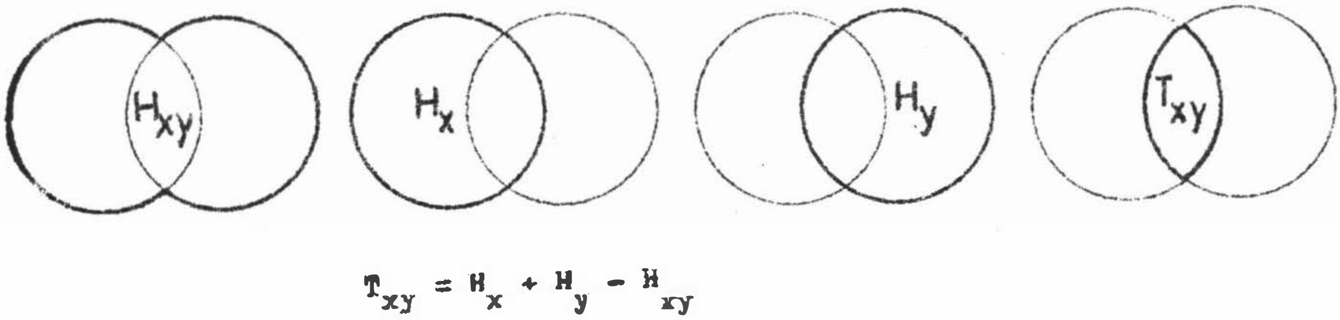
fig. 2

When information is transmitted by the channel we have the following unequality

$$H_x + H_y > H_{xy}$$

This means that given the output and our knowledge about the confusion matrix, we can make a good guess at the input.

The relations between H_x, H_y, and H_{xy} can be shown in simple Venn-diagrams. (See fig. 3)



$$T_{xy} = H_x + H_y - H_{xy}$$

fig. 3

The corss-section between H_x and H_y is called the transmission T_{xy}.

The physical meaning of the transmission is that part of the information of the input which we know when the output is known , in other words, the transmission is the information transmitted by the channel.

In order to calculate the transmission we have to make use of the confusion frequency matrix resulting from an experiment. (fig. 4)

CONFUSION FREQUENCY MATRIX

INPUT	TOTAL	OUTPUT			
		S_1	S_2	S_k	S_n
S_1	m_{10}	m_{11}	m_{12}	m_{1k}	m_{1n}
S_2	m_{20}	m_{21}	m_{22}	m_{2k}	m_{2n}
S_j	m_{j0}	m_{j1}	m_{j2}	m_{jk}	m_{jn}
S_n	m_{n0}	m_{n1}	m_{n2}	m_{nk}	m_{nn}
TOTAL	M	m_{01}	m_{02}	m_{0k}	m_{0n}

fig. 4

In this table the m's represent observed frequencies, the subscripts have the same meaning as in the probability matrix.

Taking the quotients $^m/M$ as best estimates for p's we can calculate the transmission.

The necessary calculations can easily be programmed for evaluation by an electronic computor.

All our calculations were carried out with the IBM 1130 system of the Institute of Phonetic Sciences of the University of Amsterdam.

To get some insight into the process of vowel perception we applied information theory to some data published in the literature.

We started with the well-known experiment by Peterson and Barney on formant measurements on vowels of different speakers. (JASA 1952) (fig. 5).

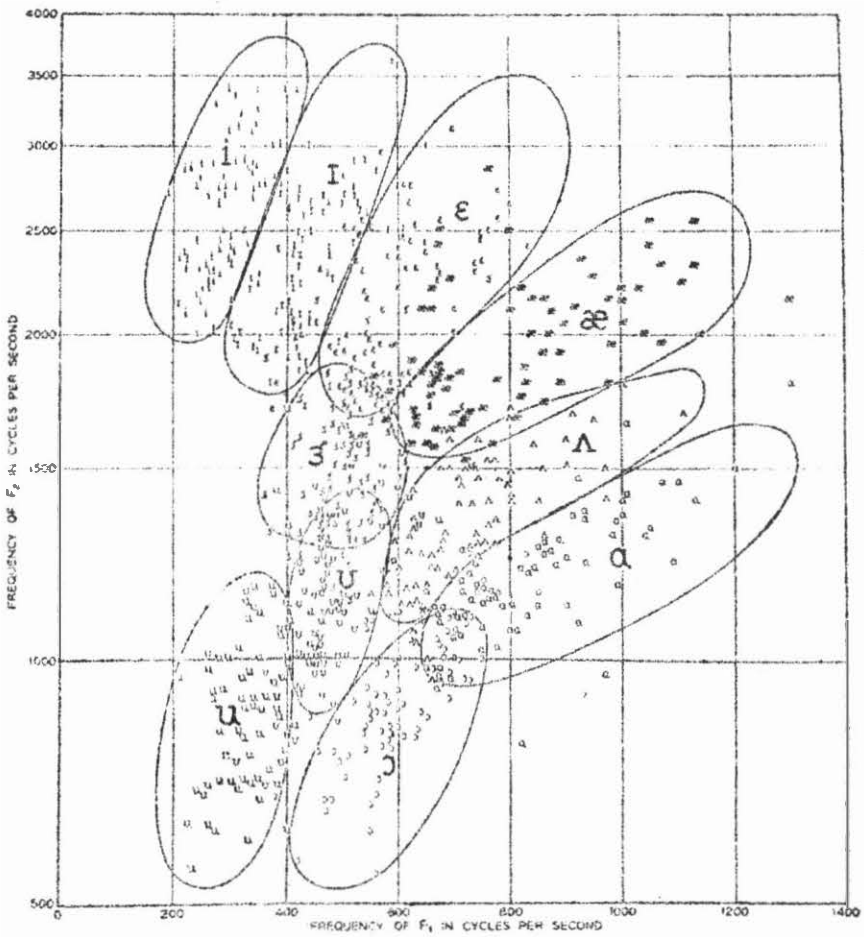


fig. 5

Suppose we have a vowel-recognition system that relates the sounds within a specific contour to one and only one vowelclass.

We determined the confusion frequency matrix for such a system shown in fig. 6 by a simple counting procedure, any sound falling in the cross-section of two areas being scored as 0.5 for each area. All frequencies are multiplied by 10 to avoid fractions.

As we see, the information of the input is 3.32 Bits, the transmission 2.19 Bits.

The same sounds were presented to a group of listeners. Peterson and Barney published the confusion matrix which is shown here as fig. 7.

When we apply our formulae to their matrix we find a transmission of 2.98 Bits.

It appears that human listeners do better than our hypothetical vowel recognition system. Our conclusion must be that man uses factors additional to the first two formants. These factors might be fundamental frequency, duration, the connection with surrounding consonants, and knowledge of the particular vowel system of an individual speaker. Although the speechsounds of different speakers were randomized, some knowledge of the position of the vowel system in the two-formant plane was available, due to the high correlation between fundamental freq. and the formant frequencies (MOL 1964)*.

As no confusion occurs when we listen to the sounds of a familiar voice we can list and add up our data as follows.

Information of input		3.32 Bits
Contributed by formant positions alone	maximal	2.19 Bits
Contributed by other factors than specific knowledge of a speakers vowel system	at least	0.79 Bits
	Sum of these factors	2.98 Bits
Contributed by specific knowledge of an individual speakers vowel system		0.34 Bits
		3.32 Bits

The next data to be examined are published by Cohen, Slis & 't Hart (Phonetica 1967) in an article entitled "On Tolerance and Intolerance in vowel perception".

They presented a highly interesting confusion matrix for a system of 12 synthetic vowels. They used 12 fixed two-formant positions and introduced duration as an extra parameter. The spacing of the vowels in the F_1, F_2 plane is somewhat exaggerated. The matrix is shown in fig. 8.

The information of the input is 3.62 Bits in formant positions and 1.55 Bits in duration which is redundant.

The transmission is 2.93 Bits. As the experimental conditions are comparable with the situation in which a person is listening to the sounds of one individual speaker, part of the information is lost. (Of course, some of the factors operating in experiments where monosyllabic words are used are absent in experiments with isolated sounds)

This low transmission is in agreement with our findings. It seems that a transmission channel operates less stably with artificial vowel-like sounds than with natural vowels.

From the results of the scaling experiment described by my collegue Meinsma an estimate can be made as to the confusion occurring between different areas of the perceptive vowel-triangle. We estimate the following data:

$$H_x = 3.6 \text{ Bits}$$
$$H_y = 3.6 \text{ Bits}$$
$$H_{xy} \approx 5.3 \text{ Bits}$$
$$T_{xy} \approx 1.9 \text{ Bits}$$

This means that the duration factor introduced by Cohen and collaborators must have contributed about 1 Bit of the 1.15 Bits of partly redundant transmitted information.

The present study is part of a larger programme which aims at the generation of vowel systems of optimal efficiency for the production of artificial speech.

UNIVERSITY OF AMSTERDAM
Institute of Phonetic Sciences
Herengracht 338

J.G. Blom

* Proceedings of the 5th Intern. Congress of Phonetic Sciences.

PETERSON&BARNEY FORMANT MEASUREMENTS

INPUT	TOTAL	OUTPUT 1	2	3	4	5	6	7	8	9	10
1	740	680	60	0	0	0	0	0	0	0	0
2	750	30	650	55	0	0	0	0	0	0	15
3	755	1	96	571	11	1	1	1	1	1	71
4	740	1	1	121	566	26	1	1	1	1	21
5	730	0	0	0	30	610	45	35	0	5	5
6	760	3	3	3	3	53	603	83	3	3	3
7	740	2	2	2	2	22	92	572	12	32	2
8	740	2	2	2	2	12	2	2	567	92	57
9	750	1	1	1	1	41	1	11	76	491	126
10	775	2	22	97	52	2	2	2	7	127	462
TOTAL	7480	722	837	852	667	767	747	707	667	752	762

$H(X) = 3.32$ $H(Y) = 3.31$ $H(XY) = 4.45$ $T(XY) = 2.18$

fig. 6

// XEQ

PETERSON&BARNEY LISTENING EXPERIMENT

INPUT	TOTAL	OUTPUT 1	2	3	4	5	6	7	8	9	10
1	10280	10267	4	6	0	0	3	0	0	0	0
2	10279	6	9549	694	2	1	1	0	0	0	26
3	10277	0	257	9014	949	1	3	0	0	2	51
4	10278	0	1	300	9919	2	2	0	0	15	39
5	10273	0	1	0	19	8936	1013	69	0	228	7
6	10279	0	0	1	2	590	9534	71	5	62	14
7	10279	0	0	1	1	16	51	9924	96	171	19
8	10279	0	0	1	0	2	0	78	10196	0	2
9	10277	0	1	1	8	540	127	103	0	9476	21
10	10279	0	0	23	6	2	3	0	0	2	10243
TOTAL	102780	10273	9813	10041	10906	10090	10737	10245	10297	9956	10422

H(X) = 3.32 H(Y) = 3.32 H(XY) = 3.66 T(XY) = 2.98

fig. 7

COHEN,SLIS,'T HART

INPUT	TOTAL	OUTPUT 1	2	3	4	5	6	7	8	9	10	11	12
1	1670	1628	42	0	0	0	0	0	0	0	0	0	0
2	1669	82	1584	0	0	1	0	0	0	2	0	0	0
3	1664	0	25	1570	0	0	12	0	0	57	0	0	0
4	1670	1	0	47	1546	68	1	4	1	0	0	2	0
5	1669	4	7	7	3	1628	0	3	10	6	0	0	1
6	1669	0	0	5	0	1	1475	128	6	53	0	0	1
7	1670	0	0	0	0	0	126	1536	2	4	0	2	0
8	1670	0	0	0	0	1	63	10	1592	0	2	0	2
9	1665	0	30	318	0	0	2	3	0	1309	1	2	0
10	1663	0	1	221	13	12	19	4	0	28	1358	7	0
11	1670	0	0	34	274	3	1	5	0	45	322	981	5
12	1670	0	0	0	0	269	1	0	3	0	109	4	1284
TOTAL	20019	1715	1689	2202	1836	1983	1700	1693	1614	1504	1792	998	1293

H(X) = 3.58 H(Y) = 3.56 H(XY)= 4.21 T(XY)= 2.93

COHEN,SLIS,'T HART VOWELS IN ONE DURATIONCLASS ADDED

INPUT	TOTAL	OUTPUT 1	2	3	4	5	6	7	8	9	10	11	12
1	6680	1629	1820	1545	985	42	71	3	49	5	81	128	322
2	8343	86	3	16	6	1621	1899	1605	1317	1287	325	66	112
3	4996	0	13	132	7	26	13	6	138	1	1796	1506	1358
TOTAL	20019	1715	1836	1693	998	1689	1983	1614	1504	1293	2202	1700	1792

H(X) = 1.55 H(Y) = 3.56 H(XY)= 3.97 T(XY)= 1.13

fig. 8