PB-WORD INTELLIGIBILITY AND SPEAKER IDENTIFIABILITY OF 5 MEDIUM BAND CODERS : A PILOT STUDY.

G.W. Boxelaar

1. INTRODUCTION

In literature many methods are mentioned to establish the quality of speech coders. The most commonly used measure defines some kind of signal-to-noise ratio between the outgoing and incoming speech of the coder. Such an objective measure hardly has any correlation with methods in which subjects judge the speech quality of a coder. For the latter various tests have been designed, where coders are evaluated with help of subjects. Well known examples of such tests are the Mean Opinion Score (Daumer and Sullivan 1982, Goodman and Nash 1982), where the quality of the speech is expressed in terms such as "good", "fair" and "bad", and the Diagnostic Rhyme Test (Voiers, 1983), which gives diagnostic information about the confusion of phonemes. Each of these tests only estimates one aspect of speech quality and usually only one test is applied to establish the quality of a coder. Our aim is to evaluate coders on various aspects: The intelligibility, the speaker identifiability, and the acceptability. This paper only deals with the first and second aspect. Later on we will concentrate more on naturalness and acceptability.
For the first aspect a word intelligibility test with subjects is designed. This test measures intelligibility on a segmental or phonemic level. Supra-segmental features are not considered in this phase of the project. Results of this test are in terms of percentages of words correct and confusion matrices of phonemes, which can be used for diagnostic purposes.
For the second aspect a speaker identifiability test has been designed in which subjects have to identify persons on the base of read aloud passages of various length. The score is expressed in terms of average shortest length at which the person is correctly recognized. This test measures to what extent speaker information is preserved by the coder systems.
The two tests will be applied to 5 coders systems and a reference system, being a telephone band. The coders are being developed to be used in an office environment, where they can be applied in speech store-and-forward systems. The test will be designed while we reckon with the office application. The reference system is included in the tests to compare the coders with a system which has often been evaluated and which possibly will have the highest quality.

2. DESCRIPTION OF THE SYSTEMS

The coder systems, designed and implemented by various partners in the ESPRIT-SPIN project "Speech interface at the office workstation", are summarized in this chapter. Most of the coders are still under development, so only prototypes and software simulations were available for the present pilot tests. We have evaluated 5 systems named A through F.

System A is a subband coder working at a bitrate of 9.6 kBit/s. The coder divides the speech spectrum into different parts by means of digital bandpass filters. The signals from these filters are separately coded. After transmission, or store and forward, the speech signal is reconstructed by decoding the bit streams from each of these filters and by summing them. For system B the same coder as system A is used but operating at a higher bitrate: 16 kBit/s.
System C uses a time domain harmonic scaling algorithm. At the encoder side each time two pitch periods (or 20 ms for unvoiced speech) are taken and merged to one period (or 10 ms). So a data compression of 2 to 1 is obtained before using further bit reduction like vector quantization. At the decoder side the two periods are reconstructed from one period. The bitrate of this coder is 12 kBit/s.
System D is a multi pulse linear predictive coder. After estimation of the prediction coefficients the residue signal is represented by a limited set of Dirac pulses of different amplitude and with different time locations. These pulses are determined in such a way that the speech signal, which is derived from the filtered residue signal, is the optimal approximation of the original speech signal, according to a perceptual distance measure. At the decoder side the residue signal operates as an source signal for the LPC filter. The transmission rate of this coder is 9.6 kBit/s.
System E is also a multipulse linear predictive coder, but somewhat different from system D. The bitrate of this coder is 10.1 kBit/s.
The partners in the SPIN project were asked to use a bitrate of 9.6 kBit/s or additionally a bitrate of 16 kBit/s. Unfortunately different bitrates were used for almost all systems.
As reference system we added system F, which is a bandpass filter with the characteristics of a Codec filter (telephone bandwidth). This is done because all coders use a Codec for filtering. They also apply 8 bit m-law analog to digital conversion, but this is not simulated in system F.

3. PB-WORD INTELLIGIBILITY

3.1. Methods

A. Speech material

The speech material consists of monosyllabic meaningless words of the type consonant-vowel-consonant, also called logatoms. The logatoms are grouped in lists of 50 words and they are phonetically balanced for Dutch. Coders are supposed to process the speech signal in a language independent way. It was therefore considered appropriate to test all systems with Dutch speech material only. Each list is read aloud by a speaker: one logatom every three seconds. Recordings were made of 6 speakers: 3 male and 3 female. One male speaker was experienced the others hardly. We recorded at least one list from each speaker and for two of them, a male and a female speaker, an extra list. To these extra word lists noise was added, with a signal to noise ratio of +10 dB. With this extra condition the sensitivity of coders to low level noise can be evaluated. In an office environment, where the coders are meant to be used, a moderate level of noise might decrease the speech intelligibility already. Table 1 gives an overview.

Speaker	Sex	list number	
		no noise	noise
1	male	30	31
2	male	32	
3	male	33	
4	female	34	
5	female	35	36
6	female	37	

Table 1. Speakers, noise conditions and list numbers

B. Experimental conditions

The systems, A through F, used in this test are 5 coders and a reference system as described in the previous chapter. Before the speech is processed by the coders it is attenuated at to levels: a high level of 10 dB below saturation and a low level of 25 dB (+/- 1 dB) below saturation. The -10 dB level is a level at which almost no peak clipping occurs. The low level is chosen, because we expect level variations of at least 15 dB due to different speakers, a varying distance between mouth and microphone and different line attenuations when speech from analog lines is transmitted. Because none of the coders is provided with an automatic gain control we expect degradation of the speech from quantization errors, when applying the low level. For the reference condition only the highest level is chosen because it is an analog system without quantization. Table 2 shows an overview of the system and level conditions.

Code	System	rate kBits/s	Level dB High	Low
A	Subband	9.6	-10	-26
B	Subband	16.0	-10	-26
C	Time dom. harm. scal.	12.0	-10	-24
D	Multip. lin. pred.	9.6	-10	-24
E	Multip. lin. pred.	10.1	-10	-24
F	Bandpass limiting	---	-10	---

Table 2. System and level conditions

C. Subjects

The listening experiment is done with 4 female listeners. An audiometric test showed normal hearing for all subjects. Only one of them was experienced in so far as she had participated in an earlier similar evaluation test. The subjects were paid for their cooperation.

D. Equipment

All recording were played back via analog reel and cassette tape recorders.

The listening was done with Beyer DT48 headphones. The subjects were seated in a quiet room. A construction of boards prohibited contact between the subjects during the listening sessions. As experiments with word lists produce a lot of data, a special testing device was designed to collect and process all responses. This device consists of four portable microcomputers, Radio Shack TRS-80 Model 100, and a central personal computer, IBM-XT. The portable microcomputers and the personal computer are connected via RS232-C serial communication ports. The portable computers act as a sort of terminals. Each is equipped with a keyboard, which has the size of a normal typewriter, and a horizontally positioned liquid crystal display with dimensions 8 lines and 40 positions per line. A program written in Basic supports all input and output. It reacts on commands from the communication port. For example: display a pre-stored message, or start and stop throughput of characters from the keyboard to the communication port. The whole experiment is controlled by the central personal computer, provided with 10 Mbyte hard disk, a floppy disk, monochrome screen, keyboard, printer and four RS232-C serial ports. A menu driven program written in Pascal and some Assembler routines supervise all these parts. It has options to route input and output devices in various ways. For example data can be stored on disk, but also directed to the printer during an experiment. Before every listening session all kinds of control variables like filenames, list numbers, and number of listeners have to be specified. This can be done by hand or automatically with the help of a sort of batch file. During the listening session the raw responses are displayed as soon as they are typed by the listeners. After a session the computer gives global intelligibility scores to allow the operator to give further instructions to the listeners, especially during a training phase. The scores are calculated from a comparison of the responses with the original lists. Usually one is not interested in comparison of the characters, but in comparison of phonemes. For this reason a translate option which translates characters, or strings of characters, into phonemes is included in the program.

E. Experimental procedure

i. Listening experiment

The 4 subjects are instructed to respond on CVC-type speech sounds, by using orthographic symbols. For Dutch an almost unique set of orthographic symbols exists covering all possible phonemes in this test. This is usually not the case in other languages. The subjects are trained for one day to get acquainted with the test and to respond with a speed of one word every three seconds. On every subsequent measuring day only a few lists are necessary for training. The actual measurements took ten days. On every day about 40 lists were run, not only from this experiment but also from another large CVC word intelligibility experiment to evaluate various communication channels, speech enhancement methods, and noise conditions. To avoid acquaintance with specific lists, every list of this specific experiment appeared only once a day. The systems and conditions were randomly ordered to avoid artificial ranking effects.

ii. Calculations

Average intelligibility scores over any combination of subjects, list

numbers (this includes noise condition and speakers), systems and level conditions, can be calculated. The average word score is calculated from number of words correct divided by the number of stimulus words. A word is noted correct if all phonemes, that is the initial consonant C1, the vowel V, and the final consonant C2 are correct. Earlier experiments showed that the word score is approximately equal to the product of the phoneme scores for C1, V and C2. Usually the score of the initial consonant is lowest, so this score will be most dominant for the word score.
An analysis of variance is done on the word scores. Because the condition with noise is only available for two speakers and the condition without noise for six speakers, a separate design is used for each of the two conditions. Also a separate design is used to compare the reference system with the coder systems, because the signal of the coder systems is attenuated at two levels, whereas the signal of the reference system is attenuated at only one level.
After the analysis of variance a Newman-Keuls post hoc analysis is done for the factor "systems". At two levels of significance we have examined whether the systems differ from each other: alpha=0.01 and alpha=0.05.
Diagnostic information is derived from phoneme confusion matrices. Stimulus phonemes are arranged at the vertical positions and response phonemes at the horizontal positions. A cell of the matrix is filled with the number of times the corresponding stimulus-response combination was found. Two extra columns of the matrix contain total number of stimuli for each phoneme and the percentage correct for that phoneme, respectively. The results of the listeners are summarized and displayed in this matrix. Matrices of different lists can be merged, in order to examine the total amount of confusions caused by a particular system, condition, or speaker.

3.2. Results

A. Average word score

Table 3 shows the word intelligibility score for all systems. The score is an average of the scores for all speakers and listeners. The scores in the "no noise" column, where high and low level condition are given separately, are based on 24 talker-listeners pairs whereas the results in the "noise" column are based on 8 talker-listener pairs.

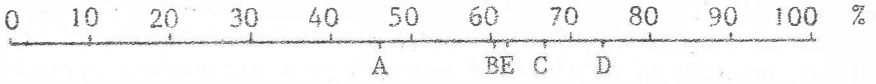
system	no noise			noise		
	high	low	average	high	low	average
A	55.9	38.3	47.1	44.3	40.3	42.3
B	69.1	51.4	60.3	69.0	56.5	62.8
C	66.9	62.3	64.6	58.8	54.0	56.4
D	75.3	72.4	73.8	62.8	68.5	65.6
E	59.7	61.8	60.7	54.8	53.0	53.9
F	81.6	----	81.6	78.0	----	78.0

Table 3. Percentage word correct score for all systems and all conditions.

From all coder systems system D (the multipulse coder working at 9.6 kbit/s) performs best for all conditions, but it performs not significantly better than system F (the reference system) which has a word correct score of 81.6 % for the no-noise condition. This is a somewhat low score compared with the results of earlier experiments (Steeneken, 1982), where the same system (telephone bandwidth) was tested and a word correct score of 94.6 % was found. This difference might be caused by the fact that we used both male and female subjects and only one of them was trained, whereas the earlier experiments were done with male and trained speakers.
The attenuation condition has not influenced the score much. Only system A shows a severe loss of word intelligibility of 17.6 percent. System D and E are effected by the noise most strongly, resulting in a loss of 8.2 percent word intelligibility for both systems.
To find out whether differences are significant, or not, various analyses of variance have been done. To compare the coder the following design is used:

4 listeners x 5 systems x 2 attenuation levels x 6 speakers

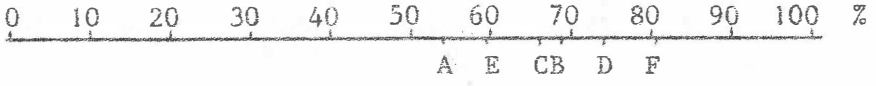
In this design the reference system is left out, because it is evaluated with only one attenuation level. Also the noise condition is omitted. We found significant differences (F-probability 0.003 %) between the systems. The average word correct scores of the 5 systems are shown in the plot below.



A Newman-Keuls post hoc analysis proved system D to be significantly better than all other systems at the significant 0.01 level and better than system A,B, and E at the 0.05 level. System A (the subband coder working at 9.6 kbit/s) is significantly worse than the others at the 0.01 level. Systems B,E, and C showed no significant differences for the average word score at both significance levels.
To examine whether the reference system is significantly better than the coders, we carried out a second analysis of variance, using the following design:

4 listeners x 6 systems x 6 speakers

In this design the noise condition and the low attenuation condition were left out, because the noise condition is not applied to all speakers and the attenuation has no effect on system F, as it is a analog system. The plot below shows the average word scores of each system on a bar.



Also for this design significant differences were found for the average word scores, so a Newman-Keuls post hoc analysis has been carried out. The reference system F proved to be better at significance 0,01 level than systems A,E,C, and B, but it did not significantly differ from system D. We compared coder system D with the other coders, and it became apparent that system D proved to be only better than system A an E, but was not

better than C and B at the significance 0,01 level.
To study the sensitivity for the coders to low level noise we also carried out a third and fourth analysis of variance. Only two lists of two speakers were treated with noise, so we used the following design to study the coders only:

4 listeners x 5 systems x 2 attenuation levels x 2 noise conditions x 2 speakers

To study both coders and reference system we used the following design:

4 listeners x 6 systems x 2 noise conditions x 2 speakers

Both analyses showed no significant differences, neither for the factor "systems" nor for the factor "noise conditions". This is probably due to the fact that we only used data from two speakers.

B. Phoneme correct score and confusions

In Table 4 the average correct scores for initial consonants are given, for system A through F. The average is taken for the no-noise and high level (-10 dB below saturation) conditions only. In the last column the frequency of occurence of the phonemes is given. Apart from these score we also calculated confusion matrices, which give us the relation between stimulus phonemes and response phonemes. These confusion matrices are not printed here.
The reference system F has a low score for /f/. According to the confusion matrix /f/ is often responded as /v/. If this confusion is due to voicing also /s/ would turn into /z/ and /p/ into /b/, but since this is not the case another distortion must be responsible for this confusion. System F, which is a telephone bandfilter, only passes spectral components between 300 and 3400 Hz, so the high frequency components, which are very characteristic for /f/, are not present in the stimuli, and /f/ can be easily confused with /v/. The coders (system A through E) show the same pattern, because the bandpass limiting is implemented in these systems too. The confusion between /f/ and /v/ could also be due to an improper pronunciation of the speakers.
Also the score for /p/ is somewhat low. From the confusion matrices it is found that /p/ is often perceived as /t/. The same counts for /b/ which is perceived as /d/. If we further notice that /f/ is not only perceived as /v/ but also in some cases as an /s/, it becomes apparent that graveness, or place of articulation, is not fully preserved by the systems used in this test. (The phonemes /t/, /d/, and /s/ are all alveolar whereas /p/ and /b/ are bilabial and /f/ is labiodental).
System E shows a low score for /b/, which is oftentimes perceived as /w/.

From Table 5 it can be seen that vowels are identified nearly correctly. This confirms the finding from earlier experiments with CVC-type words, where vowels proved to be resistent against all kinds of disturbances. However, the systems A and E show a remarkably low score of 50 percent for /oe/. According to the confusion matrix of vowels this is due to the confusion of /oe/ with /I/.

Looking at the intelligibility scores of the final consonants (Table 6) we notice a low score of /m/ and /ng/ through all systems. Both consonants

are oftentimes perceived as /n/. This confusion could be due to the pronunciation of the speakers, but it is also possible that they are easily confused by the listeners, because these phonemes look very much alike and a slight distortion could make them hard to distinguish. System A, the subband coder working at 9.6 kBit/s performs very badly. The consonant /f/ is often perceived as /s/ and /p/ as /k/ or /t/. System D, the multipulse coder working at 9.6 kBit/s performs almost equally well as system F, the reference system. Only the consonant /f/ is sometimes confused with /s/.

System	A	B	C	D	E	F	Freq.
Stimulus							
p	25.0	54.2	50.0	54.2	54.2	75.0	24
t	73.6	86.1	75.0	88.9	80.6	90.3	72
k	81.3	87.5	91.7	93.8	89.6	91.7	48
b	54.2	61.1	41.7	56.9	38.9	76.4	72
d	83.3	83.9	82.3	88.5	79.2	91.7	192
f	33.3	29.2	29.2	41.7	45.8	37.5	24
s	89.6	81.3	93.8	97.9	93.8	95.8	48
x	88.9	97.2	98.6	98.6	98.6	100.0	72
v	46.9	57.3	59.4	62.5	60.4	77.1	96
z	51.4	72.2	75.0	90.3	79.2	98.6	72
h	86.1	84.7	91.7	91.7	84.7	90.3	72
m	62.5	83.3	60.4	85.4	60.4	97.9	48
n	76.4	93.1	81.9	91.7	54.2	100.0	72
r	88.2	98.6	93.1	100.0	97.2	100.0	144
l	89.6	95.8	89.6	85.4	81.3	89.6	48
j	95.8	87.5	95.8	100.0	83.3	95.8	24
w	80.6	94.4	88.9	90.3	88.9	88.9	72

Table 4. Percentage of correctly identified initial consonants for the high level and no-noise condition.

System	A	B	C	D	E	F	Freq.
Stimulus							
AA	96.5	99.3	99.3	100.0	97.9	100.0	144
A	97.7	98.6	96.3	97.7	98.1	97.7	216
EE	86.7	96.7	95.0	95.8	90.0	96.7	120
E	76.7	97.5	95.8	98.3	78.3	99.2	120
IE	76.4	86.1	79.2	83.3	70.8	84.7	72
I	82.5	92.5	81.7	90.8	81.7	93.3	120
OE	85.4	72.9	77.1	83.3	81.3	83.3	48
OO	92.7	92.7	91.7	97.9	97.9	95.8	96
O	80.8	91.7	80.8	82.5	80.0	85.8	120
EI	91.7	99.0	97.9	96.9	87.5	97.9	96
UI	79.2	100.0	91.7	100.0	87.5	100.0	24
U	50.0	100.0	83.3	100.0	50.0	100.0	24

Table 5. Percentage of correctly identified medial vowels for the high level and no-noise condition.

System	A	B	C	D	E	F	Freq.
Stimulus							
p	50.0	79.2	87.5	79.2	70.8	79.2	24
t	82.1	95.0	94.2	98.3	96.3	99.2	240
k	98.6	100.0	98.6	98.6	98.6	100.0	72
f	37.5	75.0	75.0	75.0	62.5	91.7	24
s	94.2	87.5	97.5	100.0	99.2	100.0	120
g	86.5	95.8	94.8	95.8	95.8	95.8	96
m	31.3	31.3	22.9	41.7	35.4	64.6	48
n	80.8	80.4	88.8	85.6	70.8	91.0	312
NG	45.8	41.7	37.5	58.3	29.2	58.3	24
r	96.7	98.3	95.8	100.0	100.0	99.2	120
l	90.8	98.3	95.8	95.0	93.3	100.0	120

Table 6. Percentage of correctly identified final consonants for the high level no-noise condition.

3.3. Discussion

It is expected that using a higher bitrate for a coder would result in a better speech preservation and a higher intelligibility score. The listening test showed that the opposite is also possible. System D, one of the coders with the lowest bitrate (9.6 kBit/s), performs significantly better than the other coders, which vary in the range from 9.6 kBit/s through 16 kBit/s. The 75.3 % correct word score from system D does not significantly differ from the 81.6 score of the reference system.
From relations between the PB-word score and the scores of other tests we can predict the scores that would have been found if we performed those tests with system D. A 75 % score for PB-word test would result in a 90 % correct score of the Diagnostic Rhyme Test. Sentences would show nearly 100 % correct identification, but with the Mean Opinion Score system D would fall in the category "fair". That is in the middle of a 5 points scale.
If we added white noise to the unprocessed speech signal a signal-to-noise ratio of about +2 dB would be necessary to get a PB-word score of 75 %.
The phoneme confusion matrices did not further differentiate between the systems. The same patterns of phoneme confusions were found, more or less in quantity. In general we might say that the coders introduce errors in the identification of phonemes which are of the same kind and probably caused by the same mechanism. Only for the systems A and E a remarkable result was found for the phoneme /oe/. This phoneme, after being processed by the coders, was correctly identified for 50 percent, whereas the other systems showed nearly 100 percent. Not only for the "high" and "no-noise" condition, as described, did we found this result, but also for the "low" and "noise" conditions. Extensive spectral analyses of these phonemes, for both processed and unprocessed signals, might give an explanation of those low scores.
The PB-word experiment proved to be a convenient method to establish the intelligibility loss for medium band coders to be used in an office environment. For a final experiment, however, we propose a few modifications. For the "no-noise" condition we recorded 6 different lists. Listeners would get to much acquainted with these lists if they appeared more then once a day. To do the experiment with so few lists it is

necessary to mix them with lists of a larger PB-word experiment. We propose to use more lists in a final experiment to avoid the above mentioned problems.
The 6 lists of the "no-noise" condition resulted in significant differences for the systems. For the "noise" condition the 2 lists were not enough to gain significant results. For a final PB-word test we propose to record more lists.
We choose a +10 dB signal-to-noise ratio and noise with the same longterm average as speech to create the "noise" condition. The level and the spectrum of the noise will most probably differ from the situation in an office environment. For a final experiment we should adjust the level and choose a noise signal with the same spectral characteristics as office noise, or better use a recording made in an office.
In this pilot experiment we asked the partners of the SPIN group to create the "low" attenuation (-25 db) condition themselves. As a result from it the signals were not attenuated at the same level. In a final experiment we will redefine and record the signals at both levels and ask the partners to process the tapes at a constant level. It is more convenient then to store the signals on digital media instead of analog tapes. An analog recording at a level of 25 dB below saturation will already diminish the signal. The pulse code modulation system of Sony, using Betamax tapes would be a suitable choice to store and transport the signals in a digital form.
In an office application it is more likely to use loudspeakers instead of headphones for listening. So, it would be better to use loudspeakers for the listening tests too.

4. SPEAKER IDENTIFIABILITY

4.1. Methods

A. Speech material

The following two Dutch sentences have been used to perform a speaker identifiability test. "Binnen 10 jaar zullen kantoormachines door middel van automatische spraakherkenning en spraakproductie met de gebruiker communiceren. Deze tekst dient om de kwaliteit van dergelijke systemen te beoordelen." (Eng. : Within 10 years, office systems will communicate with the user by means of automatic speech recognition and speech synthesis. This text is meant to judge the quality of such systems.) We recorded these sentences of 6 speakers, 4 male and 2 female. They are all very well known in the Institute for Perception where the tests have been performed. We did not add noise to these recordings, like we did with the PB-word lists.

B. Experimental conditions

The speech material is processed by the 5 coders and the reference system, described in chapter 2. At the input of the systems the signal is attenuated at a level of -10 dB below saturation. At this level almost no peak clipping occurs. After the sentences have been processed, they are cut into fragments of various length. The first few fragments have been cut from the sentences at different places. These are separate words, the first one being a one-syllable word, followed by a two-syllable word , and

a three-syllable word. After that the whole passage is used in its original order, starting with the first part of the sentence, followed by longer and longer passages, until the whole passage is used. The following subdivisions were made:

1 tekst

2 zullen

3 kwaliteit

4 Binnen

5 Binnen 10 jaar

6 Binnen 10 jaar zullen kantoormachines

7 Binnen 10 jaar zullen kantoormachines door middel van automatische spraakherkenning

8 Binnen 10 jaar zullen kantoormachines door middel van automatische spraakherkenning en spraakproductie met de gebruiker communiceren.

9 Binnen 10 jaar zullen kantoormachines door middel van automatische spraakherkenning en spraakproductie met de gebruiker communiceren. Deze tekst dient om de kwaliteit van dergelijke systemen te beoordelen.

C. Subjects

All listeners, who joined this experiment, were appointed at the Institute for Perception for at least one year. Most of them were appointed for a longer period. A total of 12 listeners performed the test. According to their own statement they had no hearing losses.

D. Equipment

Analog reel and cassette tapes have been used for all recordings and playback.
The speech splicing was done with a Masscomp digital computer using a special speech editor program running on it. For this purpose the analog speech signal was digitized with a 12 bit linear analog-to-digital converter using a 15 kHz sample frequency.
Listening was done in a quiet room and via Beyer DT48 headphones.

E. Experimental procedure

1. Listening experiment

The speaker identifiability experiment was carried out with one listener at a time. Every listener had been given the same written instruction. During the listening session the listener was able to consult two pages, one page containing the 9 subdivisions of the passage, and the other one containing

the 130 names of the employees of the Institute for Perception which were appointed at the time the recordings were made. When we started the first tests one of the employees already had left the Institute.
It is only possible to present the speech of a particular speaker to a particular listeners once. After that, the listener can easily identify the speaker again. So, for every other system another speaker is necessary, if we work with the same listener. In order to test all 36 speaker-system combinations once, at least 6 listeners are necessary. The systems are arranged according to a latin square to avoid sequence effects. Sequence effect are important if the order in which the systems are presented could influence the results (Winer, 1962). The speakers were arranged in such way that all speaker-system combinations could be tested within each group of 6 listeners.
After playing one subdivision the tape is stopped and the listener is asked to write down the name of the speaker, and on a 5 point scale his feeling of being certain about the name. This action is repeated for all 9 subdivisions. If the listener has not identified the right speaker or if he is very uncertain about his answer at that time, the last fragment is played back again, but now using the unprocessed version of the same speaker. By this means we can verify the familiarity of the listener with the speaker.

ii. Calculations

The fragment at which the speaker is identified correctly is taken as the result from a session. The 5-point certainty scale has not been used till now. Every fragment is given a weight, starting with 1 for the first fragment and ending with 9 for the last and longest fragment. If the speaker was not identified after the 9 processed fragments but after the unprocessed fragment, a score of 10 was given, denoting a worse identification of a speaker via that specific system. If a speaker is not recognized at all, even after the unprocessed speech has been played back, it is assumed that the speaker is not known to the listener. In this case a score of 5.5, halfway the score range, is taken. The average score over all speakers and listeners is taken as an estimate for the preservation of speaker identifiability. An analysis of variance followed by a Newman-Keuls analysis is done on the scores.

4.2. Results

All 12 listeners were able to perform the task of this test easily. In Table 7 the scores for all system-speaker combinations are given. Each entry contains a score of 2 listeners. The symbol "-" marks those system-speaker combination where the speaker was not recognized at all. For these cells a score of 5.5 is used for further calculations.

From Table 7 it can be seen, that the speaker is correctly identified after the first fragment in 18 of the 72 cases. Especially speaker 4 is recognized oftentimes after the first fragment. Only in 8 cases the speaker is not recognized at all.
System D, with an average speaker identification between the third and the fourth fragment, performs best in this experiment, and the worst system is system B, with an average identification between the sixth and the seventh fragment. In order to explore whether differences are significant an analysis of variance is carried out with the following design:

6 speakers x 6 systems x 2 listener samples

The result of this analysis showed significant differences for the factor systems and for the factor speakers. A post hoc Newman-Keuls analysis proved system B to be significantly worse than systems E, A, F and D at the significant 0.05 level. Other significant differences were not found. To get an idea of the contribution of each fragment on the identification of the speakers the percentage of correctly identified speakers was calculated per fragment. By leaving out the cases in which a speaker was not recognized at all a 100 % identification level for the condition of unprocessed speech is defined. A curve of these scores is given in figure 1.

Coder	A	B	C	D	E	F	Aver.
Speaker							
1	5 5	10 10	7 5	- 1	5 1	4 3	5.12
2	4 1	2 2	1 4	2 5	5 1	8 6	3.42
3	5 7	7 10	5 -	1 -	8 1	2 1	4.83
4	1 1	5 5	3 1	1 2	1 1	1 1	1.92
5	8 6	6 -	10 10	10 -	10 -	8 -	7.50
6	4 1	10 5	3 -	2 1	10 3	3 5	4.38
Average	4.00	6.45	5.00	3.42	4.39	3.96	4.53

Table 7. Speaker identifiability score.

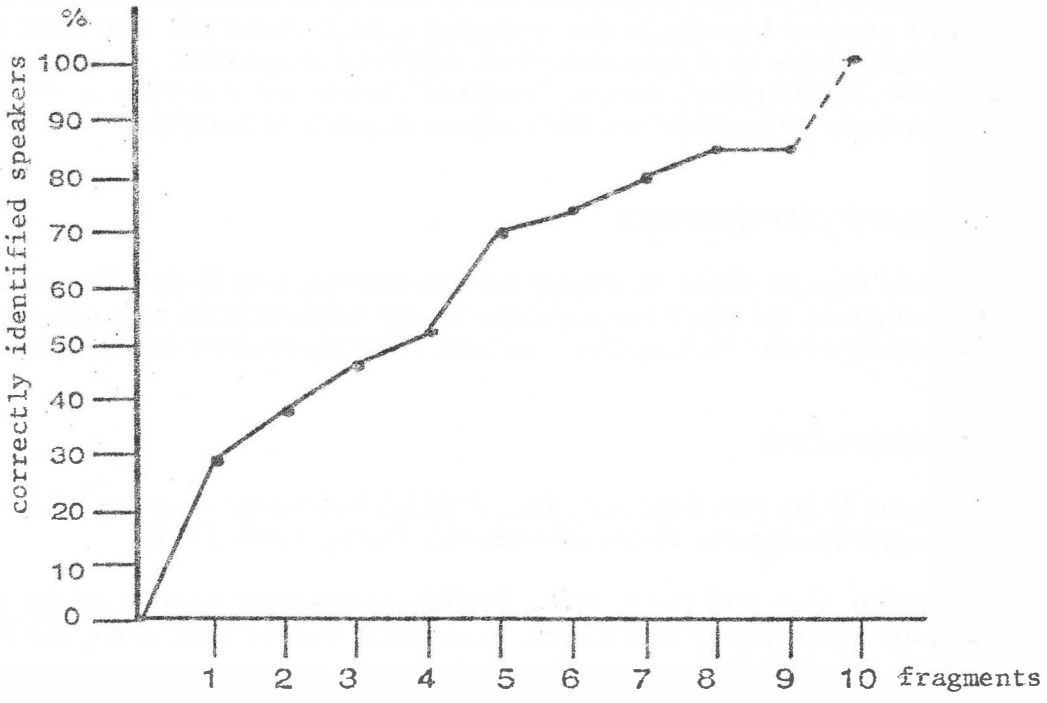


Fig. 1. Percentage of correctly identified speakers as a function of the fragments.

Already at the first fragment 28 % of the speakers is correctly identified. The second large jump in identifiability is to be seen at fragment 5. This fragment is the first short sentence in which prosodic information might be a cue for speaker identification. At fragment 9 the same score is found as at fragment 8. Fragment 8 contains one full sentence and fragment 9 contains two full sentences. The second sentence seems to give no further information for the identification of a speaker. Apparently all information is already available in the first full sentence.

4.3. Discussion

The experiment showed that system D, the multipulse coder working at 9.6 kBit/s, preserves speaker information best, but significant differences with the other systems were not found. The method seems to be convenient to measure speaker identifiability, but a larger group of listeners should be used to gain more significant data. If 36 listeners are used to test 6 systems with 6 speakers a balanced design for this experiment could be made.
The largest fragment, containing two full sentences, did not change the score of the preceding fragment, which was one sentence long. It is proposed to use only one full sentence as the longest fragment in a final experiment.
For the first fragment a stressed one-syllable word was used, resulting in 28 % correct identified speakers. To get a further refinement of the test a unstressed one-syllable word could be used for the first fragment.
The segmentation of the sentences into fragments was a very time consuming activity, because this had to be done for each system. To save time it is better to do the segmentation before the speech is processed by the systems. In that case one has to do the segmentation only once, but it might cause a change in the resulting speech material. The first method allows leakage in a fragment from adjacent fragments, as a result of system disturbances. As the fragments in the second method are isolated no leakage or interaction with adjacent parts is possible.

ACKNOWLEDGEMENTS

I would like to thank R. Plomp for his advice, and H.J.M. Steeneken and G. Langhout for their cooperation in the experiments. All are appointed at the Institute for Perception. L.C.W. Pols supervised this project.

REFERENCES

Daumer, W.R. and Sullivan, J.L., (1982), Subjective quality of several 9.6-32 kb/s speech coders, Proc. ICASSP82, Paris, 1709-1712.

Goodman, D.J. and Nash, R.D., (1982), Subjective quality of the same speech transmission conditions in seven different countries, IEEE Trans. Comm. 30, 642-654.

Steeneken, H.J.M., (1982), Development and evaluation of a Dutch Diagnostic Rhyme Test for assessing the intelligibility of speech communication channels, Institute for Perception TNO, report IZF 1982-13, 30 p. (in Dutch).

Voiers, W.D., (1983), Evaluating Processed Speech using the Diagnostic Rhyme Test, Speech Technology, Jan./Feb., 30-39.

Winer, B.J., (1962), Statistical principles in experimental design, New York, McGraw-Hill Book Company, 672 p.