IZF 1990 A-13

SPEECH DATA-BASE FOR INTELLIGIBILITY AND SPEECH QUALITY MEASUREMENTS

H.J.M. Steeneken
F.W.M. Geurtsen
E. Agterhuis

Number of pages: 22

CONTENTS

Rep.No. IZF 1990 A-13

TNO Institute for Perception,
Soesterberg, The Netherlands

Speech data-base for intelligibility and speech quality measurements

H.J.M. Steeneken, F.W.M. Geurtsen, and E. Agterhuis

SUMMARY

A data-base with speech recordings for intelligibility measurements has been made. This data-base contains, for each speaker, 50 CVC-word lists of 51 words per list, 65 sentences, connected discourse and whispered speech. This speech material is recorded for four male and four female speakers.
The recordings were made on digital audio tape (DAT). The recorded speech material was specified with respect to speech levels, spectra and level stability.

Rap.nr. IZF 1990 A-13

Instituut voor Zintuigfysiologie TNO,
Soesterberg

Spraak-databestand voor gebruik bij verstaanbaarheidsmetingen en spraakkwaliteitsbeoordeling

H.J.M. Steeneken, F.W.M. Geurtsen, en E. Agterhuis

SAMENVATTING

Een databestand met spraakmateriaal ten behoeve van verstaanbaarheidsmetingen is opgenomen. Dit databestand bevat per spreker 50 woordlijsten van 51 woorden per lijst, 65 zinnen, lopende spraak en fluisterspraak. Dit materiaal is opgenomen voor vier mannelijke en vier vrouwelijke sprekers.
De opnamen werden gemaakt op digitale audio tape (DAT). Het opgenomen spraakmateriaal werd gespecificeerd ten aanzien van de spraakniveaus, de spraakspectra en de stabiliteit van de niveaus.

1 INTRODUCTION

Since 1970, in our laboratory, a data-base with nonsense CVC-words (Consonant-Vowel-Consonant) is used for intelligibility tests. The words were embedded in a carrier phrase. There were recordings of four male speakers with 50 lists of 50 words for each speaker. This data-base was recorded on analog tape. Over the past years the tape (even the master) deteriorated, pre-echoes became noticeable and the noise level increased.
The data-base consisted of phonetically balanced word lists which is not optimal if confusion matrices have to be constructed. For these reasons we have recorded a new data-base for male and female speech. We also used a modern recording technique (DAT, digital audio tape). For each speaker not only CVC-lists were recorded, but also short sentences, connected discourse and whispered speech. This offers the possibility to perform different types of intelligibility tests, quality ratings, and analysis of different speech tokens for the same group of speakers.
With the extensive use of the former data-base, experience has been gained concerning the number and selection of speakers/listeners, recording conditions, and specification of the recorded speech. This knowledge is included in establishing this new data-base.

2 SPEECH MATERIAL

2.1 Word lists

The use of CVC nonsense words allows for experiments with an open response and a high discrimination between conditions with good and excellent quality. (Steeneken and Houtgast, 1987, 1990). For an effective use of the open response design a word list with all possible phonemes: initial consonants, vowels and final consonants, is required. For practical reasons however phonemes, which are used less than 1% in normal speech, should be omitted to keep the data-base manageable. This means for the Dutch language that 17 initial consonants (C_i), 15 vowels (V), and 11 final consonants (C_f) are required. Each word list consists of 51 words, hence each initial consonant is present three times in each list. For the vowels and final consonants a repetition of some of these phonemes is used in order to get a total number of 17. This selection is according to quasi phonetic balancing.

The phonemes according to the orthographic and the SAMPA notation (SAMPA, 1989) are given below. The additional phonemes are given between brackets.

C_i: p, t, k, b, d, g, f, s, h, v, z, m, n, l, r, j, w
p, t, k, b, d, x, f, s, h, v, z, m, n, l, R, j, w (SAMPA)

V: ie, ee, i, e, u, aa, a, uu, eu, o, oo, oe, ei, ui, au (a,e)
i , e:, i, E, Y, a:, A, y , 2:, O, o:, u , Ei, 9y, Au (SAMPA)

C_f: p, t, k, f, s, g, m, n, ng, l, r (l,t,r,s,n,m)
p, t, k, f, s, x, m, n, N , l, R (SAMPA)

All C_iV combinations are allowed for the Dutch language. However, some VC_f combinations do not exist. These restrictions are:

ijr, uir, aur,

ieng, oeng, uung, ijng, uing, aang, aung, oong, eung, eeng.

The test words are random combinations of CVC's, resulting in both nonsense or meaningful words. Some undesirable words ("dirty" words) were avoided.
We also avoided that successive test words have two or more phonemes in common. The lists were generated by a special program which makes a random combination of CVC sequences but with the restrictions mentioned above.
The CVC words are embedded in five different carrier phrases.
The carrier phrases were:

Attentie	...	einde
En nu	...	over
En zo	...	onder
Versta	...	uit
Volgende	...	aan.

The effect of such carrier phrases is that the pronunciation of the test word will be more natural and unstressed. For experiments with reverberation or other distortions in the time domain the premasking introduced by the first part preceding the test word, is essential while the second part is essential to mask echoes of the test word. The effect using a carrier phrase is shown by Houtgast and Steeneken (1984) where results of an international experiment are given from

eleven laboratories using different test material for the evaluation of the same conditions consisting of combinations of reverberation and noise.
Also, the vocal effort is more easy to control by the speaker when the test word is not isolated but part of a sentence. Appendix A gives an example of the application of these carrier phrases.
We recorded 50 different lists for each speaker giving a total of 400 lists.

2.2 Sentences

Sentences were recorded for experiments on quality rating, as used with the Mean Opinion Scoring method (Goodman and Nash, 1984), or for experiments based on the Speech Reception Threshold (SRT, Plomp and Mimpen, 1979). For each speaker 26 sentences were recorded in groups of 13 sentences as required for the SRT method.
An example of such a group is given in Appendix B.

2.3 Connected discourse

A recording of some text, ***read*** by the speaker serves as connected discourse. This text, given in Appendix C, contains all digits from 0 to 9. The text was composed for the evaluation of speech recognizers, trained for digits and numbers. The text size is according to a representative sample of the language.

2.4 Whispered speech

For some experiments the use of whispered speech is required. For instance to study effects of voiced/unvoiced decisions with vocoders. We recorded the same text items as used for the connected discourse, but asked the speakers to whisper. Some speakers had a difficulty to whisper properly without activating the vocal cords.

3 RECORDING CONDITIONS

3.1 Speakers

There were four male and four female speakers ranging in age from 22-52 years and in length from 1.60 m to 1.95 m. It is assumed that the age is related to the clearness of the voice and that the length is related to the size of the vocal tract.
The initials of each speaker, the sex, the age and size are listed below:

HVD	male	52 yrs	1.95 m
HB	male	32 yrs	1.85 m
OW	male	25 yrs	1.85 m
FG	male	33 yrs	1.80 m
EVB	female	38 yrs	1.75 m
MVR	female	33 yrs	1.60 m
IH	female	22 yrs	1.74 m
AL	female	29 yrs	1.83 m

We recorded 50 different lists for each speaker, where one list takes about 180 seconds.
Two groups of 13 sentences (75 seconds per group), one token of connected discourse (120-180 seconds), and one token of whispered speech (180 seconds). This results in two DAT-tapes for each speaker. At the beginning of each tape a calibration signal is recorded. This signal is used as a reference for the individual list levels and the levels of the other speech material.

3.2 Equipment, display and recording environment

In Fig. 1 a general block diagram of the recording equipment is given. The speaker is placed in an anechoic room in front of a recording microphone (1/2" condenser microphone, type B&K 4155), a speaking distance of 50 cm, and the registration of the vibration of the vocal cords with a laryngograph (Fourcin and Abberton, 1976). Both signals are recorded on two tracks of the DAT (16 bits resolution, 48 kHz sample rate).
The display of the words and the sentences was controlled and timed by the personal computer. The display, in front of the speaker, was a liquid crystal type display from a TANDY TRS-80 portable computer.

This computer was programmed as a terminal. The display was placed in front of the speaker on a table. Also in front of the speaker was a level display to control the (long term) vocal effort. This was a sound level meter set on fast response and A-weighting (B&K, type 2209).
A blanket was placed upon the table in order to suppress reflections from the table in front of the speaker. This table with the necessary equipment was placed in an anechoic room in order to avoid reflections from the walls, see Fig. 2. The experimenter and a part of the equipment was placed outside the anechoic room as illustrated in Fig. 1.
The recording rate for the words was 3.5 s per word. The display was programmed to display the word within the carrier phrase during 2.2 s and to display blanks during the remaining 1.3 s. The PC initiated also a trigger at the beginning of the display. This trigger was used to start a tone-burst which was mixed with the larynx signal. The tone-burst frequency (12.5 kHz) was above the frequency range of the larynx signal (< 500 Hz).

The following recording conditions were used for the speech data-base:

- anechoic room SPL < 45 dB or < 20 dB(A)
- omnidirectional condenser microphone, B&K 4155, 1/2"
- speaking distance 50 cm, approx. 15° from centre axis
- suppression of reflections via table to microphone (blanket on table).
- no tube light transformers in speaking room (no hum)
- airconditioning off (less low frequency noise)
- SNR (lin) on tape recording > 35 dB or 50 dB(A)
- no electro acoustical feedback to speaker
- no extra gradient microphone near speakers mouth
- automatic item display for words (silent LCD terminal)
- recording speed 3.5 s/word (words in carrier phrase)
- level control (feed back to speaker)
- laryngograph
- DAT-recorder 16 bit resolution
- level calibration after recording with threshold dependent, A-weighted, RMS measure.

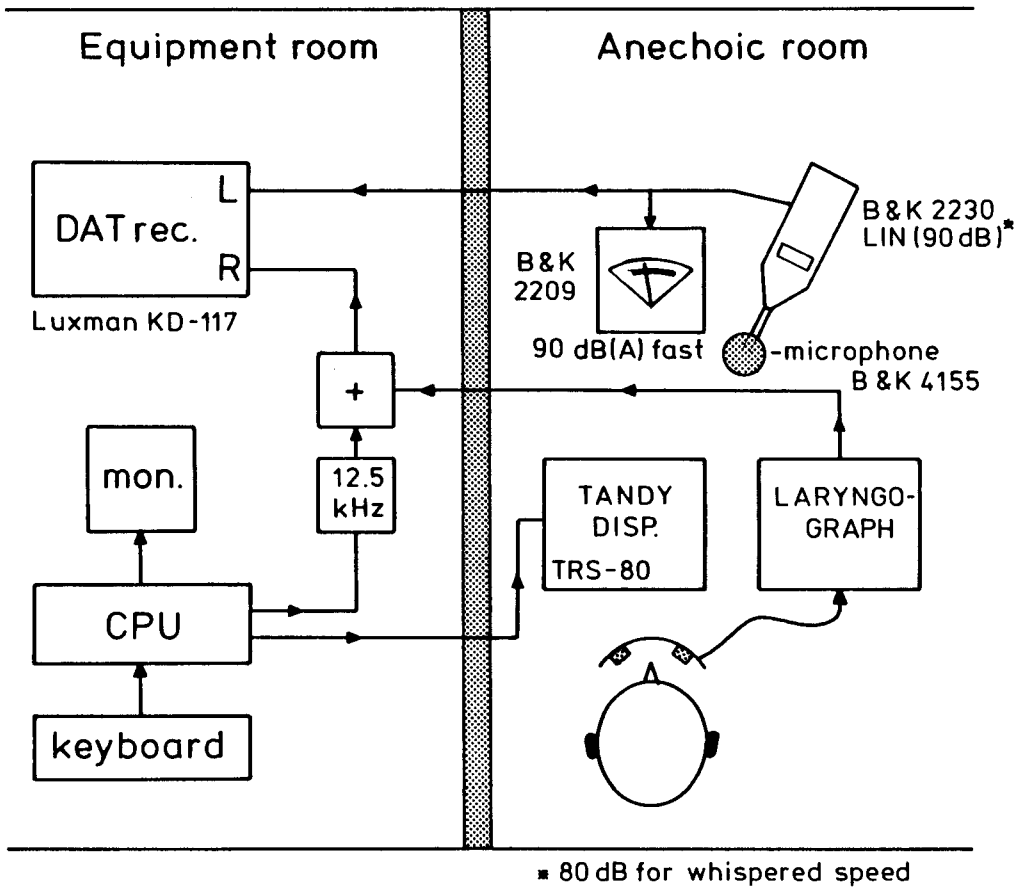

Fig. 1 Block diagram of the recording equipment.

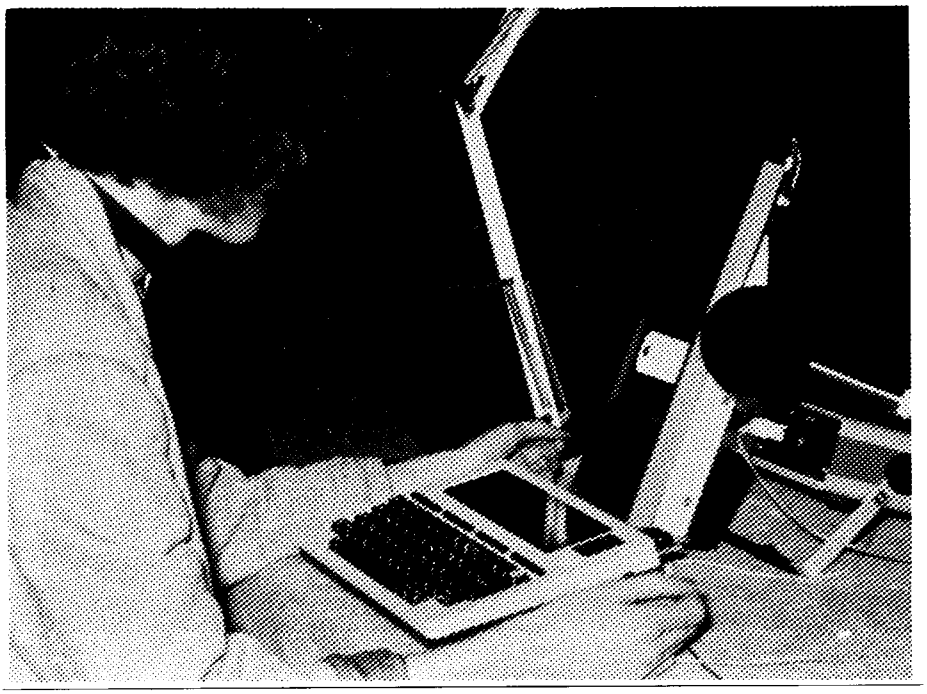

Fig. 2 Speaker position in the anechoic room.

4 SPECIFICATION OF RECORDED SPEECH MATERIAL

4.1 Speech level

A comparison between speech level measures and their effectiveness was made before (Steeneken and Houtgast, 1986). This study proposes a method which can be used for single words and for connected discourse, as the method uses a trigger level to exclude the contribution of silent periods to the final value. The speech signal is filtered through an A-weighting filter which results in a closer relation between wide band signals, male/female, and telephone speech. The method is used with the STI approach and programmed for the STI hardware. The method is comparable with the CCITT method but has some extra advantages such as more accuracy with noisy speech, no threshold during zero crossing of speech tokens, applicability for other signals than speech.
We applied this measure for all word lists, sentence lists, connected discourse, and whispered speech. These levels are given in Table I. As the values for the CVC-words are mean values obtained from 50 lists for each speaker, the standard deviation is also given. However this standard deviation reflects the variation within and between several speaking sessions on several days, and does not indicate the variation within a list.

Table I Mean speech levels in dB(A) (m) and standard deviation (s) for 50 CVC-word lists, and mean levels for the other speech material, for each of the 8 speakers.

Speaker	CVC-words (m)	(s)	sentences	connected discourse	whispered speech
HVD	-11.6	1.0	-12.2	-12.8	-10.4
HB	-17.5	1.8	-13.3	-14.5	-22.2
OW	-13.6	2.3	-11.6	-14.8	-18.7
FG	-16.3	0.7	-15.1	-15.2	-14.8
EVB	-13.6	1.3	-12.8	-12.1	-18.9
MVR	-18.9	0.9	-16.6	-17.9	-17.2
IH	-9.5	0.9	-19.9	-21.7	-24.9
AL	-13.8	0.6	-14.1	-13.7	-17.4

The vocal effort is found to be quite constant for each speaker and varies over 10 dB between speakers.

In order to obtain information concerning the level stability within a list, we measured for five lists per speaker the levels for the first and second half of a list. The difference between these two levels indicates the level stability within a list. The results of these measurements, for five lists for each speaker, are given in Table II. It was found that the level variation according to the given definition is maximal 1.2 dB, but typical (mean value) for males 0.5 dB and for females 0.3 dB.
The individual list levels (not given) can be used to adjust the level of each list to a reference level which is important for measurements at a predefined signal-to-noise ratio, modulation level etc.

Table II Difference in dB(A) level of the first and second half of a word list, for five lists for each speaker.

Speaker	List number of speaker specific set of lists				
	1	12	24	36	48
HVD	0.2	0.0	0.3	-0.2	-0.2
HB	-0.4	-0.6	1.0	0.3	-0.2
OW	0.4	-0.7	0.4	0.8	-1.0
FG	-1.5	-0.1	0.5	0.1	-1.2
EVB	-0.2	-0.2	0.2	-1.0	-0.9
MVR	0.5	0.2	0.2	-0.3	-0.2
IH	0.2	0.4	0.1	0.0	0.8
AL	-0.7	0.0	-0.2	0.5	0.1

4.2 Spectra and fundamental frequency

The 1/3 octave spectra for 5 CVC-word lists, two groups of sentences, connected discourse, and whispered speech were measured for each speaker. The measurements were performed with a Rhode and Schwarz 1/3 octave analyzer type FAR. The spectra were measured every 100 ms. The response of the system was set to "fast" identical to the response of a sound level meter. The measurement were made during 60 s for each type of speech. Based on the measured sample values, the equivalent level (energy summation) was calculated.
Because of the variation of the fundamental frequency between speakers, the spectra differ in the low frequency range (up to 400 Hz). We only present the mean spectra for the four males and the four females and separately for the four types of speech material. The

spectra are given in Figs. 3-6 for the males and Figs. 7-10 for the females.

The fundamental frequencies, based on the individual 1/3 octave spectra, are given in Table III. A resolution of 1/3 octave was considered to be sufficient because of the variation of the pitch during the speech tokens. As can be seen the fundamental frequency varies over an octave between the males and females.

Table III Fundamental frequency in Hz (1/3 octave resolution) for the 8 speakers.

Speaker	Fundamental Frequency
HVD	100
HB	100
OW	125
FG	100-125
EVB	200-250
MVR	200
IH	200-250
AL	200

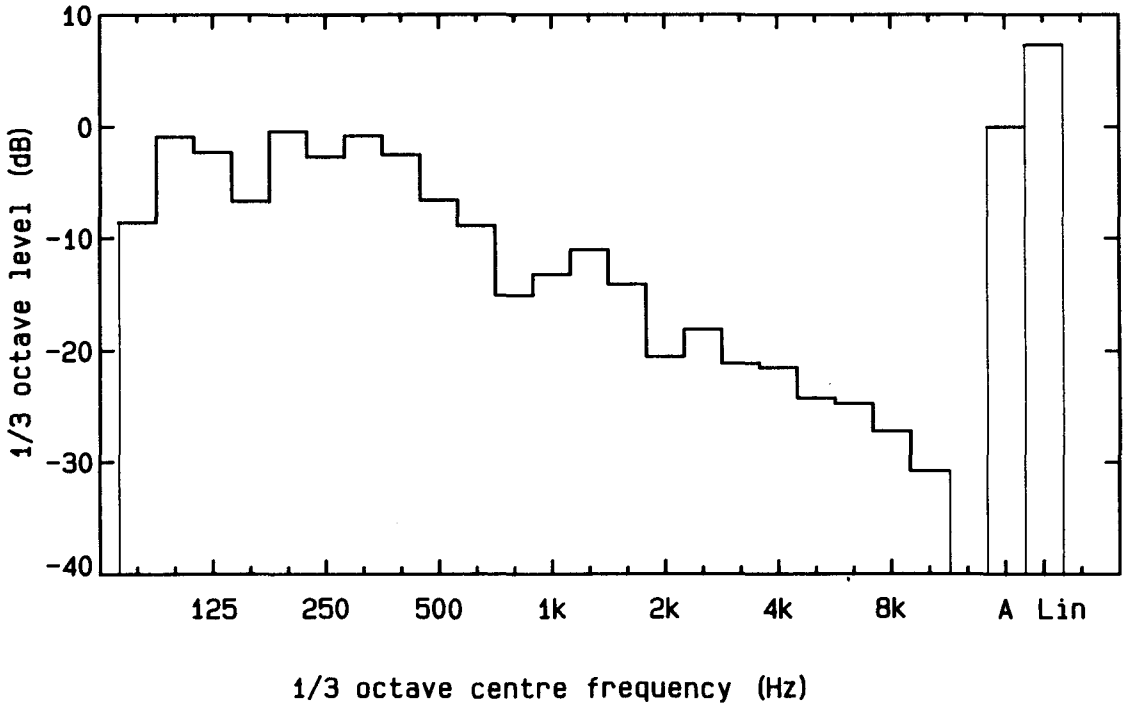

Fig. 3 1/3 Octave spectrum of the four male speakers and the CVC-word lists.

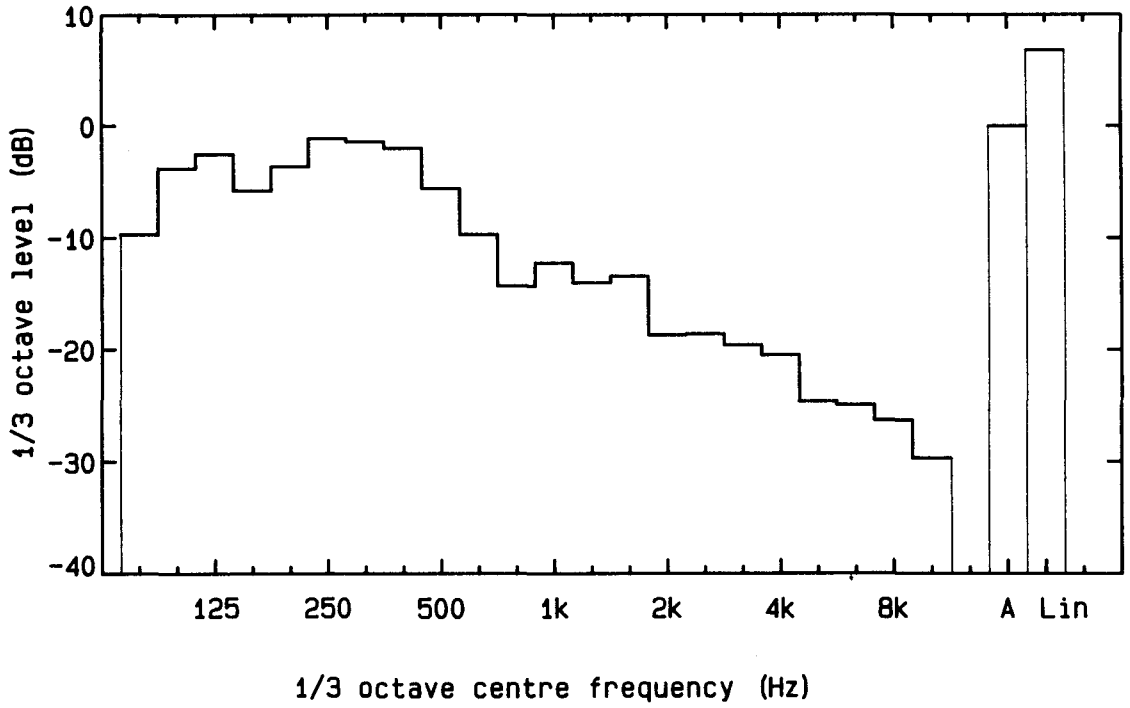

Fig. 4 1/3 Octave spectrum of the four male speakers and the sentence lists.

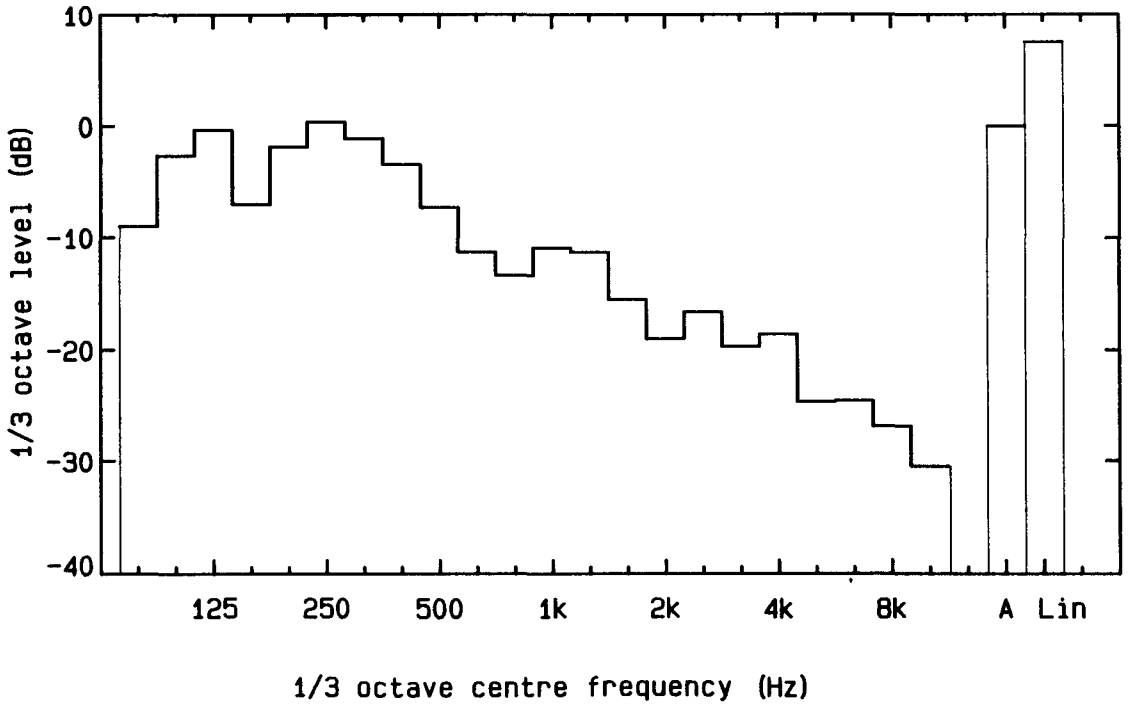

Fig. 5 1/3 Octave spectrum of the four male speakers and connected discourse.

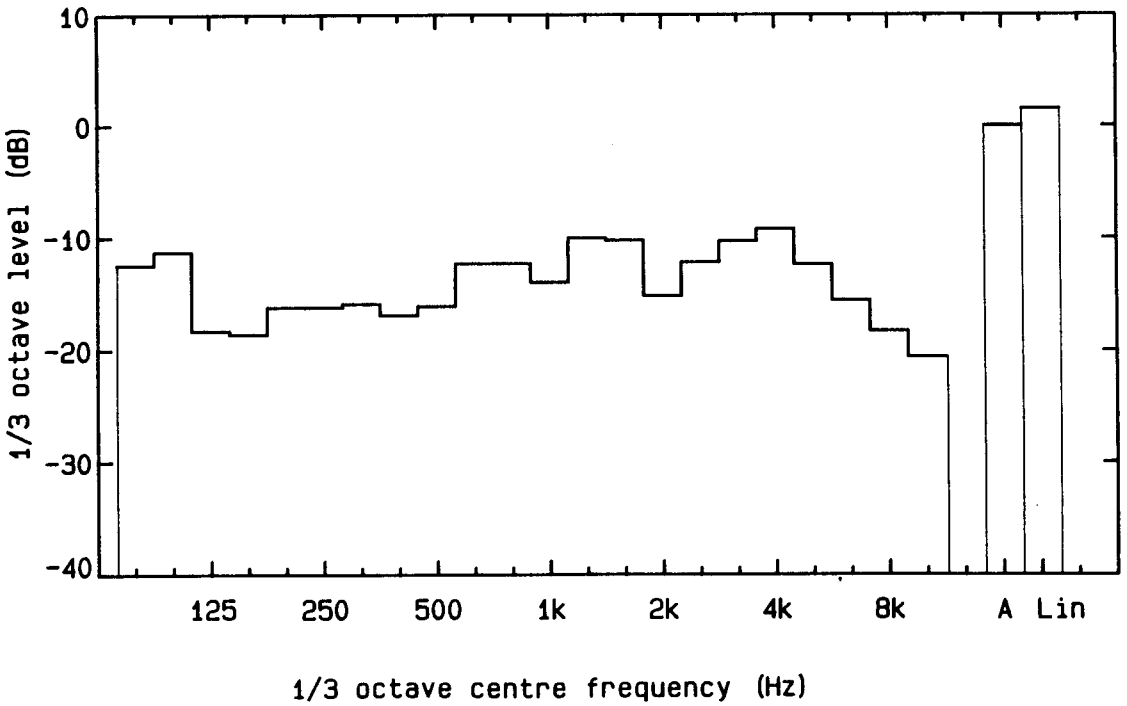

Fig. 6 1/3 Octave spectrum of the four male speakers and whispered speech.

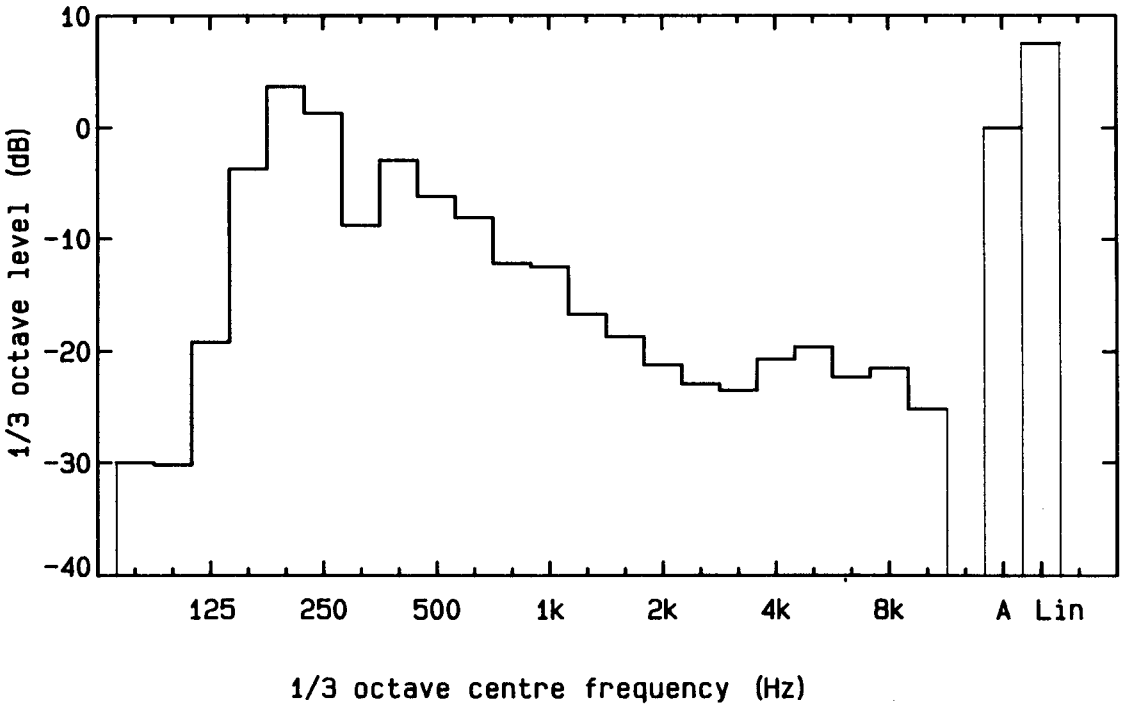

Fig. 7 1/3 Octave spectrum of the four female speakers and the CVC-word lists.

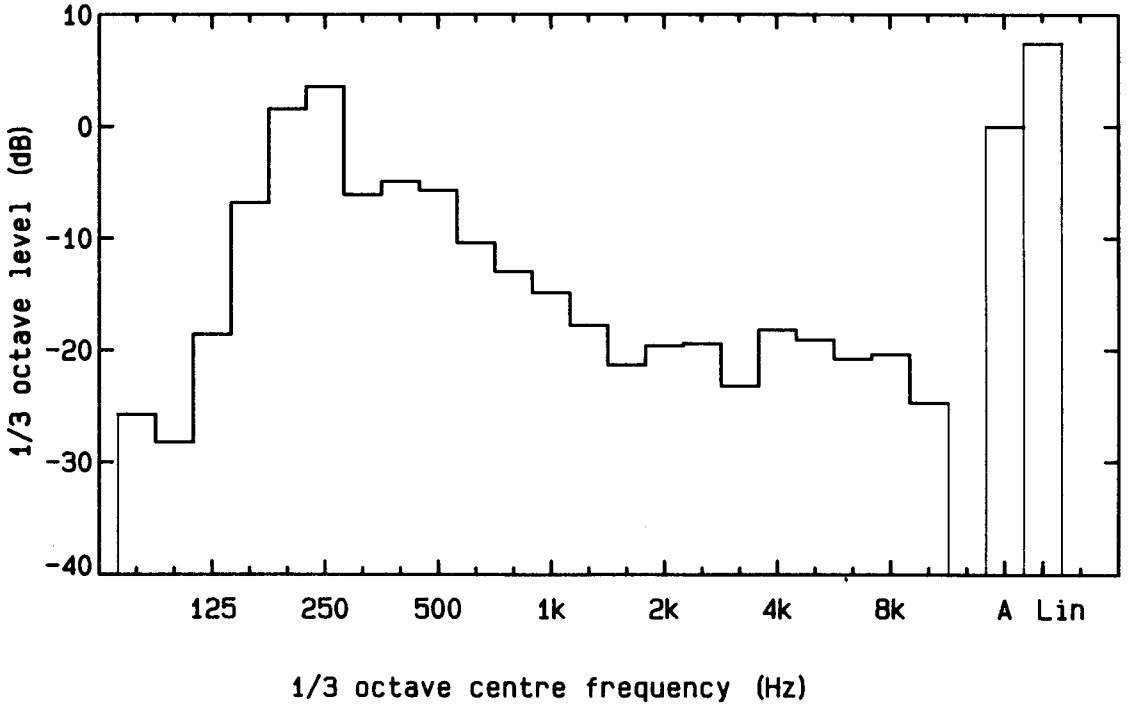

Fig. 8 1/3 Octave spectrum of the four female speakers and the sentence lists.

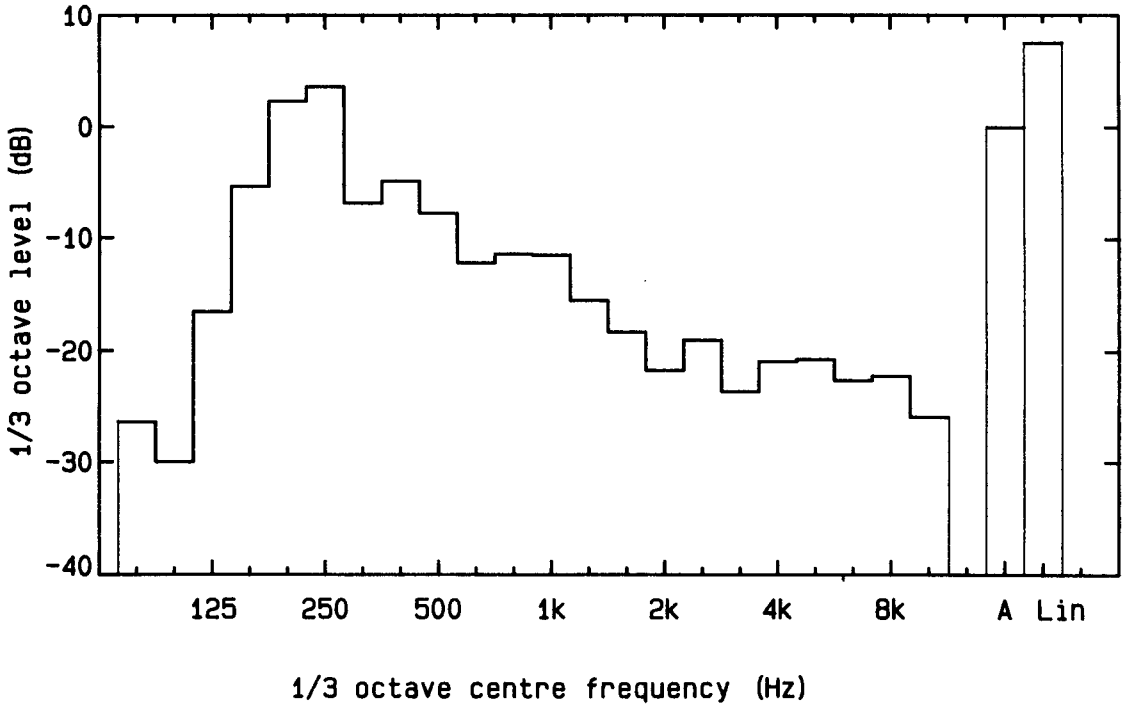

Fig. 9 1/3 Octave spectrum of the four female speakers and connected discourse.

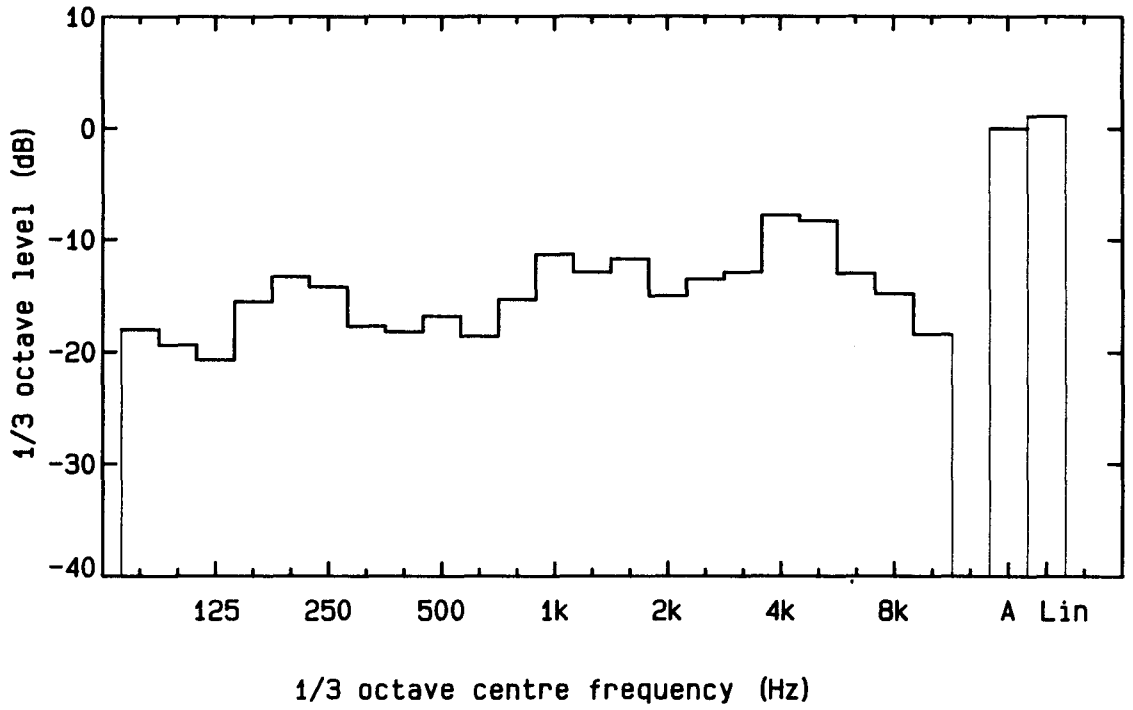

Fig. 10 1/3 Octave spectrum of the four female speakers and whispered speech.

5 CALIBRATION OF LISTENERS

Actual intelligibility measurements require extensive training of new listeners. Since the measuring sessions may extend over many days, any effect of learning on the task should be monitored. For this reason a number of reference lists is passed through reference transmission channels to measure learning effects. The results from these measurements, which are obtained during each daily session, can also be used to calibrate one group of listeners against another group in the same experiment or an other experiment. We have used five lists from each speaker for this purpose and recorded the speech through the reference channels in order to avoid hardware variations.

The description of the reference conditions will be given in the report concerning the digital speech transmission system simulator. With this simulator distortions as noise, band-pass limiting, automatic gain control, peak and centre clipping, digital modulation (PCM, delta mod etc), echoes and reverberation can be obtained. The five

reference conditions are selected to cover transmission qualities between excellent and poor and different types of distortions.

6 CONCLUSION

A data-base for CVC-words (in a carrier phrase), sentences, connected discourse and whispered speech has been recorded for four male and four female speakers. The data-base was carefully calibrated with respect to speech levels for each continuous speech token (lists, group of sentences,etc.) Also the level variability within a token was studied.

REFERENCES

Fourcin, A.J. and Abberton, E. (1976). The laryngograph and voiscope in speech therapy, E. Loebell (ed.), Proc. of the XVI Int. Congress of Logopeadics and Phoniatrics, Basel: S. Karger, pp 116-122.

Goodman, D.J. and Nash, R.D. (1984). Subjective quality of the same speech transmission conditions in seven different countries, IEEE Trans Comm. 30, 642-654.

Houtgast, T. and Steeneken, H.J.M. (1984). A multilanguage evaluation of the Rasti-method for estimating speech intelligibility in auditoria. Acustica 54, 185-199.

Plomp, R. and Mimpen, A.M. (1979). Improving the reliability of testing the speech reception threshold for sentences, Audiology 8, 43-52.

SAMPA, ESPRIT-SAM Project Nr. 1541 (1987). Phonetic alphabet (1989), See Wells, J., Computer coded phonetic transcription. J. of the International Phonetic Association, 17(2), 94-114.

Spiegel, M., Altom, M.J., Macchi, K. and Wallace, K. (1989). A monosyllabic test corpus to evaluate the intelligibility of synthesized and natural speech. Proceedings ESCA Workshop, Noordwijkerhout, The Netherlands.

Steeneken, H.J.M. (1986). Diagnostic information of subjective intelligibility tests. Internat. IEEE Proc., ICASSP, Dallas.

Steeneken, H.J.M. and Houtgast, T. (1986). Comparison of some methods for measuring speech levels. Report IZF 1986-20, TNO Institute for Perception, Soesterberg, The Netherlands.

Steeneken, H.J.M. (1987). Comparison among three subjective and one objective intelligibility test. Report IZF 1987-8, TNO Institute for Perception, Soesterberg, The Netherlands.

Soesterberg, March 2, 1990

Ing. H.J.M. Steeneken

Appendix A

Example of a CVC-word list

DIES	FIJS	ZEK
LAL	VAN	WAUM
KEL	JOOF	NENG
NIG	TUM	SUUG
KEUN	HUIK	JOER
BEUM	PUIT	BEUN
SOP	FAG	BUIM
MIES	NEER	RIS
HUNG	ROOL	LOK
GAN	REEM	ZAUL
SOON	MAAR	GIEP
TUUM	WET	PAR
FAUT	KIJT	HAAR
PUUN	GES	DIT
JOL	VIJP	DOER
VEF	ZAF	WUNG
TEEL	MOET	LAAS

Example of a test word in a carrier phrase

Attentie	dies	einde
En nu	fijs	over
En zo	zek	onder
Versta	lal	uit
Volgende	van	aan.

Appendix B

Example of a group of sentences according to Plomp and Mimpen (1979)

List 1

1. De bal vloog over de schutting
2. Morgen wil ik maar 1 liter melk
3. Deze kerk moet gesloopt worden
4. De spoortrein was al gauw kapot
5. De nieuwe fiets is gestolen
6. Zijn manier van werken ligt mij niet
7. Het slot van de voordeur is kapot
8. Dat hotel heeft een slechte naam
9. De jongen werd stevig aangepakt
10. Het natte hout sist in het vuur
11. Zijn fantasie kent geen grenzen
12. De aardappels liggen in de schuur
13. Alle prijzen waren verhoogd

Appendix C

Text for connected discourse and whispered speech

Onlangs verscheen in het blad "School en leerling" een artikel over het decimale stelsel. Daarin was een beschrijving opgenomen van een rekensom. Opvallend daarbij was dat in de hele tekst de getallen waren uitgeschreven.
Hieronder volgt een passage:
Het arabische of decimale getallenstelsel heeft slechts de getallen nul, één, twee, drie, vier, vijf, zes, zeven, acht, negen, en een komma nodig om oneindig veel getallen te kunnen samenstellen. De plaats van een getal ten opzichte van de komma is bepalend voor de waarde van dat getal.
Een voorbeeld:
Het getal acht-een-negen komma zes-vijf kan gelezen worden als acht maal tien tot de tweede, plus één maal tien tot de eerste, plus negen maal tien tot de nulde, plus zes maal tien tot de min eerste, plus vijf maal tien tot de min tweede.
Vooral het optellen binnen het tientallig stelsel is een eenvoudige zaak. De getallen op dezelfde positie ten opzichte van de komma, worden bij elkaar opgeteld. Wanneer het getal groter wordt dan negen, vindt er een overdracht plaats naar de volgende kolom. Een eenvoudig voorbeeld illustreert de bedoeling. Wanneer we de getallen vijf-zes-vijf en zeven-twee-zeven optellen, beginnen we achteraan. Tel vijf en zeven bij elkaar op en je krijgt twaalf maal tien tot de nulde. Dat is ook weer te geven als één maal tien tot de eerste plus twee maal tien tot de nulde. De één maal tien tot de eerste wordt nu overgedragen naar de volgende kolom en opgeteld bij zes en bij twee enzovoort...
Bovenstaand stukje illustreert met welk gemak we omgaan met getallen. Waarneer echter niet voor een cijfer-representatie wordt gekozen, komen er aanzienlijke problemen met de verwerking.