institute for perception

soesterberg
kampweg 5, p.o. box 23
the netherlands
telephone 03463 - 1444

report nr. IZF 1986-20 copy nr. 22
of 50 copies

A COMPARISON OF SOME METHODS FOR MEASURING SPEECH LEVELS

H.J.M. Steeneken and T. Houtgast

Classifications:
Report : unclassified
Title : unclassified
Abstract : unclassified

Number of pages: 21

national defence research organization tno

CONTENTS

Rep. IZF 1986-20, TNO Institute for Perception, Soesterberg

A comparison of some methods for measuring speech levels

H.J.M. Steeneken and T. Houtgast

ABSTRACT

A comparison of several methods for measuring speech levels was made. These methods were divided into three groups: direct sampling, envelope sampling and meter readings. Speech fragments from connected discourse and from (embedded) word lists were used as the test materials. Based on the measuring results, a relative scale for the relations between the different measures was obtained. A selection of preferred methods could be made with respect to: the application of a threshold (in order to ignore silent periods between isolated utterances), band-pass limiting, background noise and the relation to intelligibility. In general, measures based on the envelope of A-filtered speech fragments, with a rejection of silent periods are preferred.

Vergelijking van meetmethoden voor spraakniveaus

H.J.M. Steeneken en T. Houtgast

SAMENVATTING

Een aantal meetmethoden voor het meten van spraakniveaus werd vergeleken. Deze methoden kunnen worden verdeeld in drie groepen: golfvormbemonstering, omhullendebemonstering en meteraflezing. Voor het vergelijken van deze methoden werd gebruik gemaakt van spraakmateriaal bestaande uit "lopende-spraakfragmenten" en woordenlijsten met testwoorden in een korte dragerzin. Op basis van de meetresultaten kan de relatie tussen de verschillende spraakniveaumaten worden vastgelegd op een relatieve schaal. Er werd een selectie van aanbevolen methoden gemaakt op basis van: het toepassen van een drempelwaarde (teneinde pauzen tussen afzonderlijke spraakuitingen te elimineren), bandbreedtebeperking (telefoonspraak), achtergrondlawaai en de relatie met de verstaanbaarheid. In het algemeen genieten methoden gebaseerd op de omhullende van het spraaksignaal, na frequentieweging (A-filter) en met onderdrukking van de pauzen de voorkeur.

1 INTRODUCTION

A comparison of the results of experiments in which the level of connected discourse or isolated test words is involved, is often complicated by the variety of methods used for defining the speech level. Because of the wide dynamic range of speech signals and the influence of silent periods between utterances, the quantitative relationship between, for example, the 1% peak level and the A-weighted equivalent sound-pressure level is not known a priori, but has to be determined experimentally.
There are a large number of methods for establishing a speech level. These methods can be divided into three groups:

1. methods where the original waveform is considered;
2. methods where the envelope (after rectification and integration) is considered; and
3. methods where the speech signal is fed into a meter (such as a VU meter or a sound-level meter) and the (peak) deflections on the instrument are considered.

In this paper a number of these different level measures are applied to connected discourse and to embedded test words. The main point of interest is to quantify the relation between the various measures. Also, on the basis of some further considerations, a selection will be made of preferred methods, i.e. those methods which result in a robust and relevant measure of the speech level.

2 DESCRIPTION OF SPEECH LEVEL MEASURING METHODS

Figure 1 shows a block diagram of the different methods, together with a listing of the various measures which can be obtained. The subscript thr (threshold) indicates that values below a given threshold are ignored in the calculation of that particular measure. In some cases, a frequency weighting (A-filter) is applied to the input signal. The measured values are then expressed in dB(A).

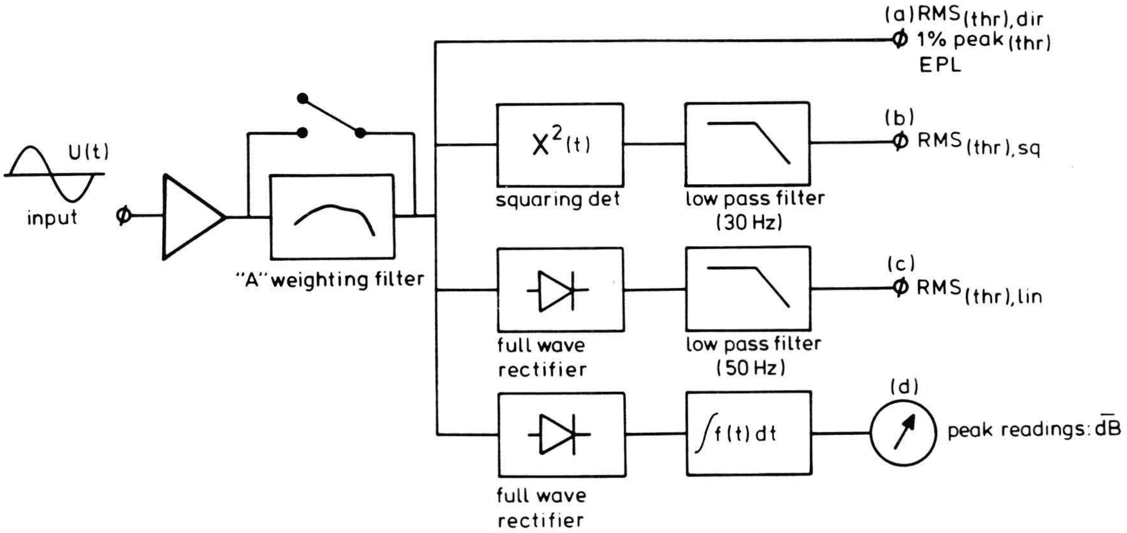

Fig. 1 Principle of a: waveform or direct sampling (dir), b: mean square detection (sq), c: envelope detection (lin) and d: mean of peaking readings (dB).

2.1 Waveform or direct sampling

The instantaneous signal value is converted to a number by analog-to-digital conversion and then processed by a computer. This processing usually is performed after the sampling sequence, then during sampling the sample values are stored in a histogram. Usually, the sampling rate is related to the highest significant frequency component in speech, and a sampling rate of typically 20 kHz (50 µs) is chosen. However, since information about the frequency content of the signal is not required, lower sample rates (e.g. 250 Hz) may also be applied (British Telecom Research Labs.). From the original histogram a conversion can be made to a log-amplitude scale using, for example, 1 dB intervals. Figure 2 shows an example of such a histogram (Van Heusden, Plomp and Pols, 1979). From the histogram the 1% peak level (level exceeded by 1% of the samples) and the RMS_{dir} value can be calculated.

A threshold can be applied mathematically by ignoring all samples below a certain value (Fig. 2). The application of such a threshold at direct sampling excludes not only silent periods between speech utterances, but also sample values from the speech signal near a zero crossing.

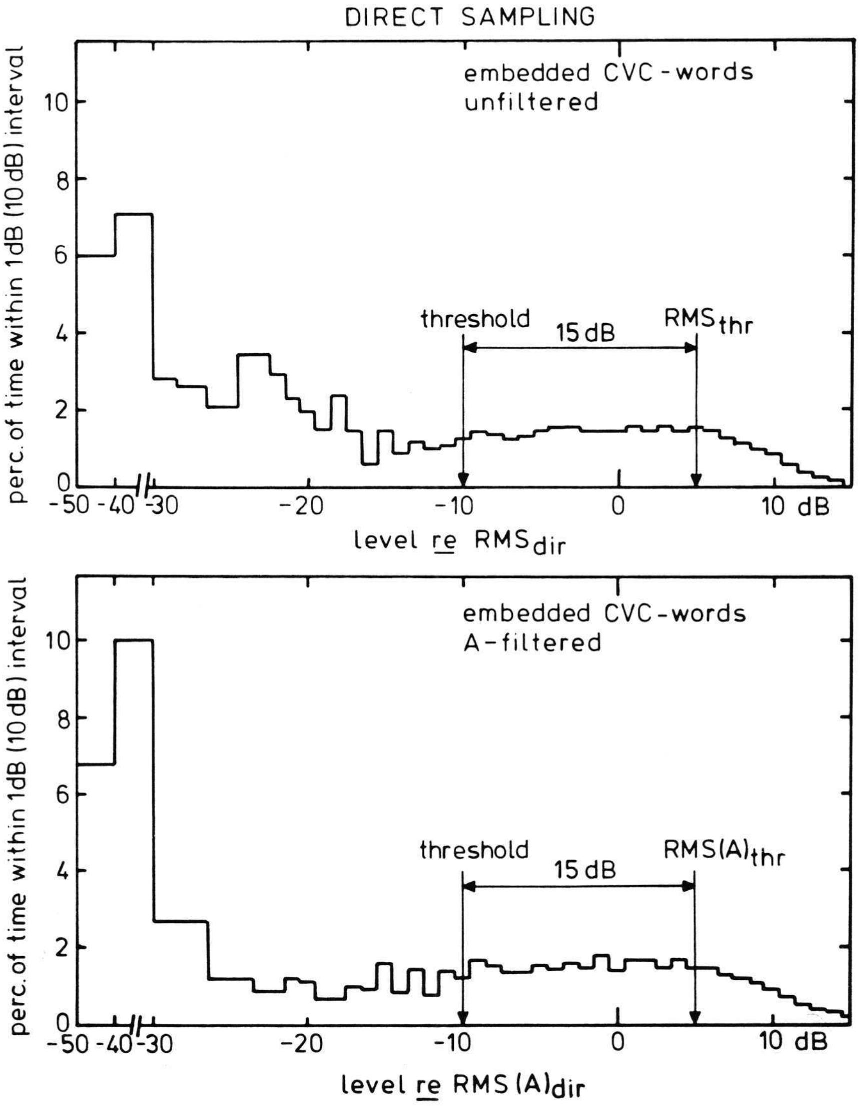

Fig. 2 Level distribution histograms of a 150-s speech fragment from embedded CVC words obtained with direct sampling. The left arrow line represents a threshold value, 15 dB below the $RMS_{thr,dir}$ value.

The shape of the upper part of the level distribution histogram obtained from unfiltered speech shows a uniform distribution. By using this feature a peak level can be defined which is, in a certain range, independent of the threshold value. The EPL method (Equivalent Peak Level: Brady, 1968) is based upon this.

2.2 Envelope Sampling

The envelope function of a speech signal can be derived from a full-wave rectification of the signal followed by integration or low-pass filtering. The rectification can be performed either with

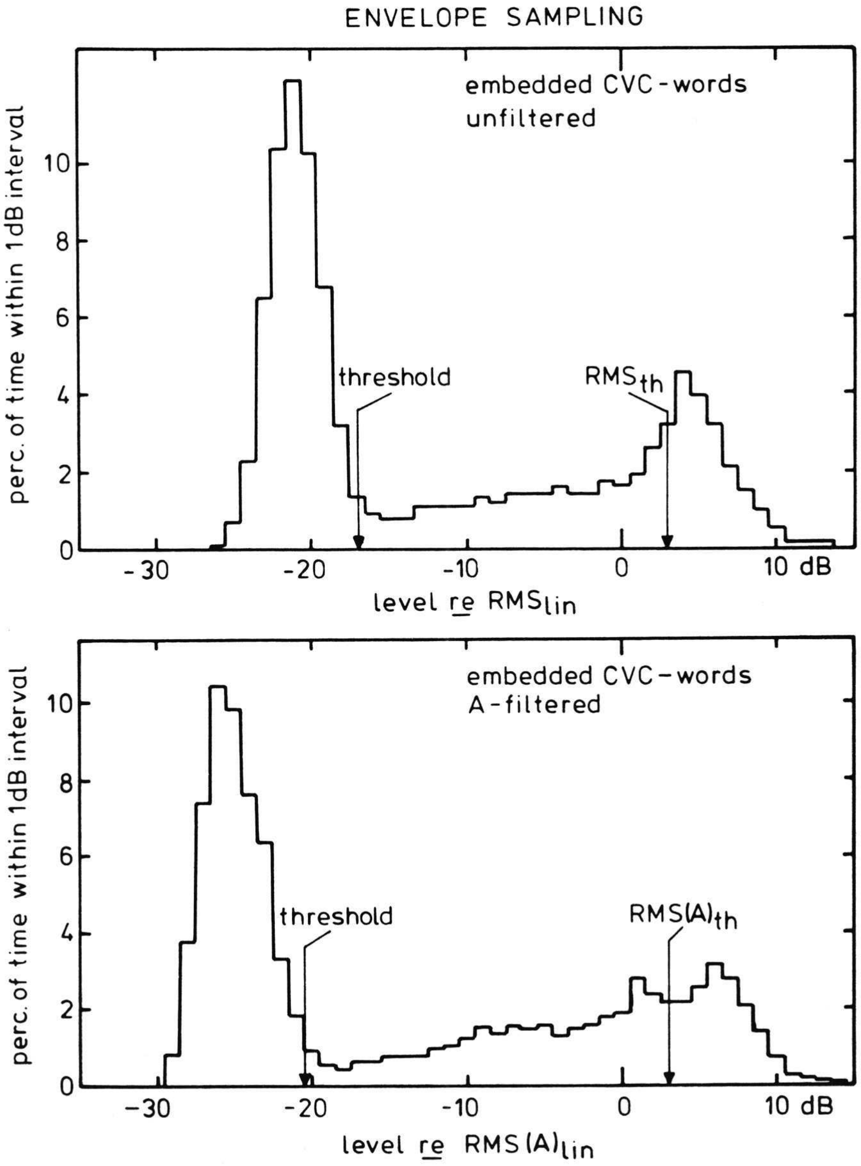

Fig. 3 Level distribution histogram of a 150-s speech fragment from embedded CVC words obtained with envelope sampling after a full-wave rectification and integration. The left arrow line represents a threshold value just above the distribution of the background noise during the silent periods.

analog circuits or by digital signal processing. For rectification a squaring detector (algorithm) (Berry, 1971) or a linear detector (algorithm) (Steeneken and Houtgast, 1978) can be applied. A squaring detector leads to an envelope reflecting the intensity of the signal (Fig. 1, output b), whereas a linear detector leads to an envelope related to the mean absolute amplitude of the signal (Fig. 1, output c).

For both methods the resulting envelope can be sampled and stored in a level distribution histogram (Fig. 3). The sample values obtained from the linear envelope detector must be squared for calculating the RMS value. However, because of the low-pass filtering, an estimated and not a true RMS value is obtained. The error depends on the cut-off frequency of the low-pass filter and on the crest factor of the signal (for speech we found an error of typically -1 dB). From a level distribution as the one in Fig. 3, a separation between speech samples and silent periods can be made very easily. In fact, two distributions are obtained, one from the speech signal and one from background noise during the silent periods between the speech utterances. The most likely threshold value (indicated by the left arrow in Fig. 3) lies just above the noise distribution.

2.3 Envelope Peak Reading

A simple method for obtaining a speech level is to average the peak deflections on a VU meter or a sound-level meter, with the sound-level meter set for fast response and 'A'-weighting (Fig. 1, output d) (Kryter, 1970).

The advantage of this method is its robustness for the influence of silent periods between the speech utterances, for the contribution of carrier phrases and for the influence of tape noise. The disadvantage is that the meter readings depend on the meter damping and on the accuracy of the readings. Therefore a level recorder is sometimes used to register the peak deflections. However, the dynamic behaviour of a (logarithmic) level recorder is level dependent and quite different from the sound-level meter standards.

3 EXPERIMENTAL CONDITIONS AND RESULTS

Two types of speech signals were used for the experiments:

- connected discourse;
- CVC test words embedded in a short carrier phrase of two or three words.

Speech signals from four male talkers, trained to speak at a constant level, were used. The speech signals were recorded on tape with the microphone placed at a distance of 50 cm in front of the speaker. The recordings were preceded by a calibration signal (sinusoid 1000 Hz) as a level reference and a pink noise signal to check the frequency response of the system.

The measurements with connected discourse were based on a speech sample of approximately one minute for each speaker. The measurements with the CVC word tests were based on word lists consisting of 50 different CVC words. The words were embedded in 5 different short carrier phrases. The duration of a word list was about 2.5 min, of which about half the time was occupied by the silent intervals between the successive utterances. For each speaker, three different lists were used.

3.1 Comparison of the levels

Table I specifies the relations between the different measures on the basis of the average of speech tokens from four speakers.

The relations between the same level measures either with or without the application of a threshold differ for connected discourse and embedded CVC-words because of the different relationship between the duration of the silent periods and the speech fragments.

The threshold value for the measures based on direct sampling was chosen 15 dB below the RMS_{thr} value according to the method described by Berry (1971; British Telecom Research Labs.). To obtain this value a table is calculated for the RMS_{thr} value as a function of the threshold value. From this table the final RMS_{thr} is selected being 15 dB above the corresponding threshold value. This threshold definition has the advantage of being related to the final RMS value.

Table I Relation between some speech level measures for connected discourse and embedded CVC words and the variation from four different speakers around their mean based on an equal intelligibility criterion.

	Connected discourse	Embedded CVC words	
METHOD	level in dB re RMS_{thr}	level in dB re RMS_{thr}	variation among speakers in dB at equal intell. (stand.dev.)
RMS_{dir}	-0.9	-2.7	1.0
$\mathrm{RMS}_{thr,dir}$	1.1	1.5	1.3
$\mathrm{RMS(A)}_{dir}$	-5.9	-7.5	0.9
$\mathrm{RMS(A)}_{thr,dir}$	-2.2	-1.7	1.5
1% peak	10.5	9.6	1.2
1% peak_{thr}	11.4	11.6	1.4
1% peak(A)	6.5	5.2	1.3
1% $\mathrm{peak(A)}_{thr}$	8.6	8.9	1.9
EPL	8.6	8.6	1.2
EPL(A)	4.7	5.3	1.4
RMS_{sq}	-0.9	-3.0	1.1
$\mathrm{RMS}_{thr,sq}$	0.0	0.0	1.1
$\mathrm{RMS(A)}_{sq}$	-5.4	-7.2	1.0
$\mathrm{RMS(A)}_{thr,sq}$	-4.9	-4.3	1.0
RMS_{lin}	-1.9	-4.1	1.2
$\mathrm{RMS}_{thr,lin}$	-1.2	-1.3	1.5
$\mathrm{RMS(A)}_{lin}$	-6.7	-8.8	1.0
$\mathrm{RMS(A)}_{thr,lin}$	-5.9	-5.7	1.2
dB(A,fast) BK2209	-4.5	-1.6	0.9
dB(A,100) BK2305	-3.9	-3.4	0.4

For the measures based on envelope sampling the separation between speech distribution and silent period is more obvious. The threshold value can be chosen just above the noise distribution as described by the level distribution histogram (see Figs. 2 and 3).

For connected discourse the effect of the threshold means an increase of the level with approx. 1 dB which implies that during 80% of the measuring time the considered speech token was above the threshold value. For the embedded CVC-words used in this study the level increase is about 3 dB, this value is related to an above-threshold time of 50%.

The levels of the two types of speech signals can be compared by considering those level measures which are independent of the duration of the silent periods. Some of these level values for connected discourse as well as for the embedded CVC-words from Table I are plotted in the two centre columns of Fig. 4. These level measures for the two types of speech material match within 1 dB (except peak

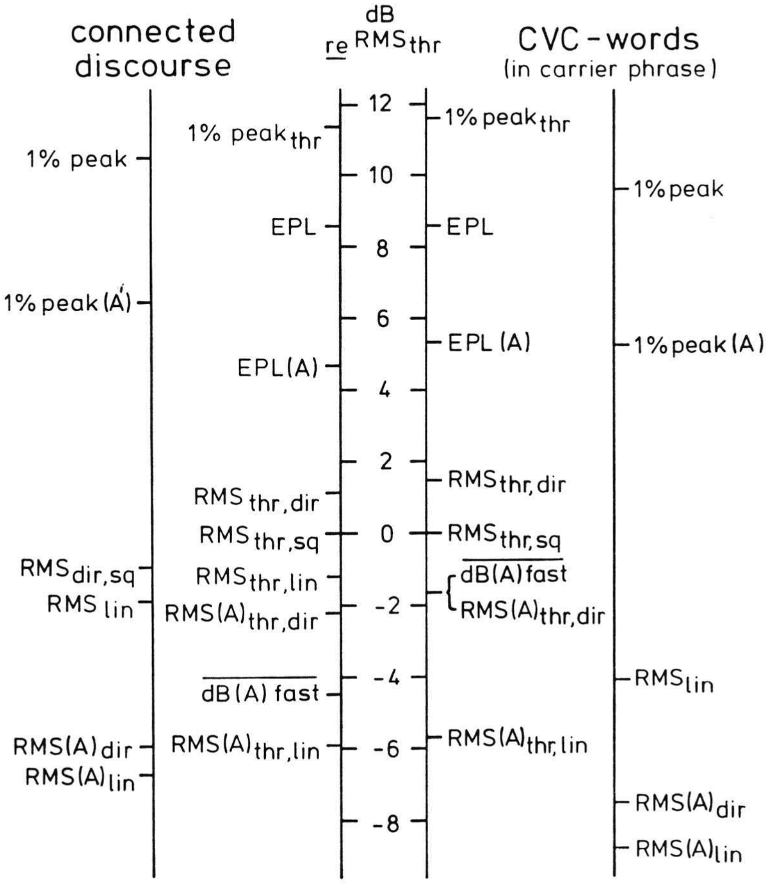

Fig. 4 Relative speech levels found with the different measures for connected discourse and embedded CVC words. The values are relative to the $\mathrm{RMS}_{\mathrm{thr,\,sq}}$ value.

readings on a level meter). This means that the threshold definitions for direct sampling and envelope sampling coincide. The effect of the threshold definition on the final level measure is given in Figs. 5 and 6.

As can be seen in these figures, the smallest level shift as a function of the threshold value is obtained with the EPL method. The EPL level shift in these graphs is plotted referring to the EPL obtained with a threshold value 30 dB below the RMS_{dir} as the algorithm cannot be used without a threshold definition. The $RMS_{dir,thr}$ and $RMS_{lin,thr}$ level shift are plotted relative to these values without a threshold. The fast increase for the RMS_{lin} and $RMS(A)_{lin}$ curves, with 3 dB for the embedded CVC-words, at a threshold value of -15 dB and -20 dB respectively, clearly indicates the optimal threshold value to be

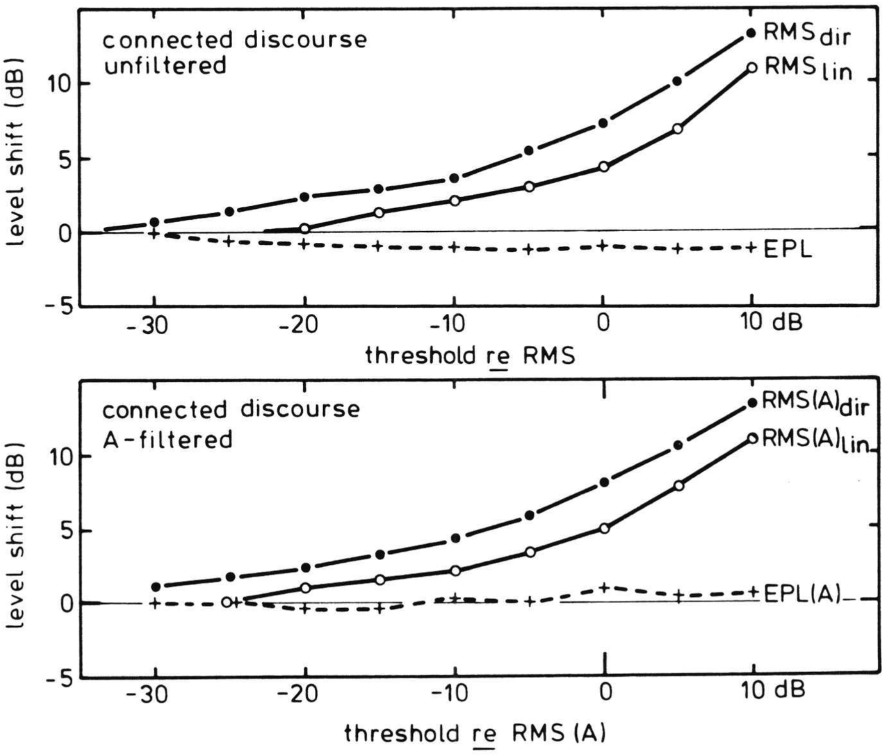

Fig. 5 RMS_{dir}, RMS_{lin} and EPL shift as a function of the threshold value of a 60-s speech fragment for connected discourse unfiltered and after A-filtering.

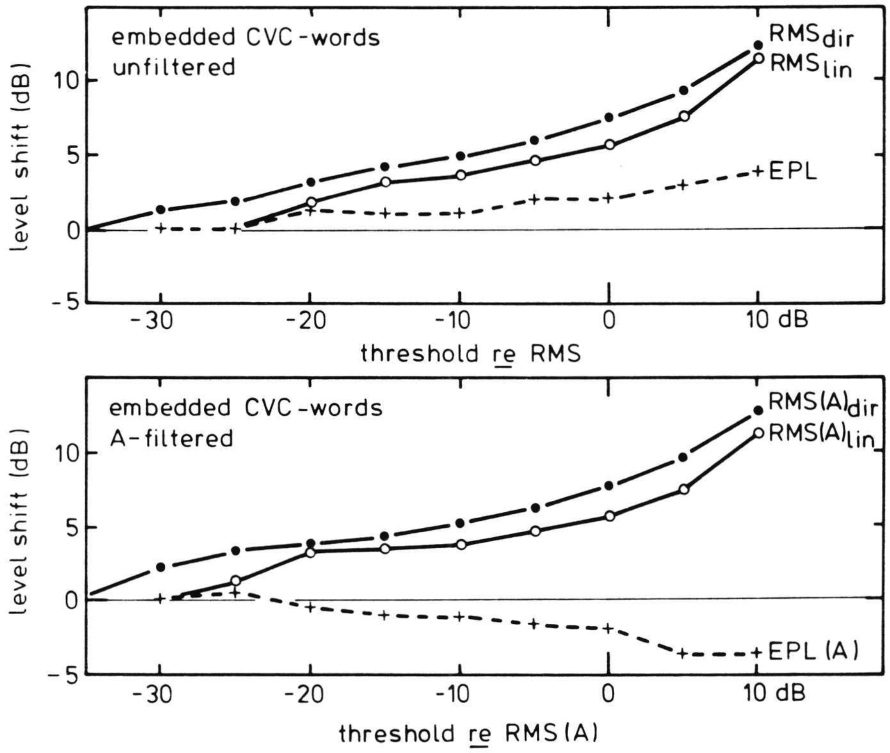

Fig. 6 $\mathrm{RMS}_{\mathrm{dir}}$, $\mathrm{RMS}_{\mathrm{lin}}$ and EPL shifts as a function of the threshold value of a 150-s speech fragment from embedded CVC-words unfiltered and after A-filtering.

just above the tape noise level. The level shifts for the direct sampled RMS values show a more continuous shape.

3.2 The effect of distortion on the speech level

In some situations the speech signal from which the level has to be determined is distorted by a band-pass filtering (as from telephone circuits) or by noise (background noise or system noise). Both distortions affect the speech level.

The level of an unfiltered speech signal mainly depends on the frequency components below 500 Hz. A high-pass filter as is used in telephone communications will, therefore, introduce a significant change in the level (typically -5 dB). Table I shows that the influence of A-filtered versus unfiltered speech, results in a level

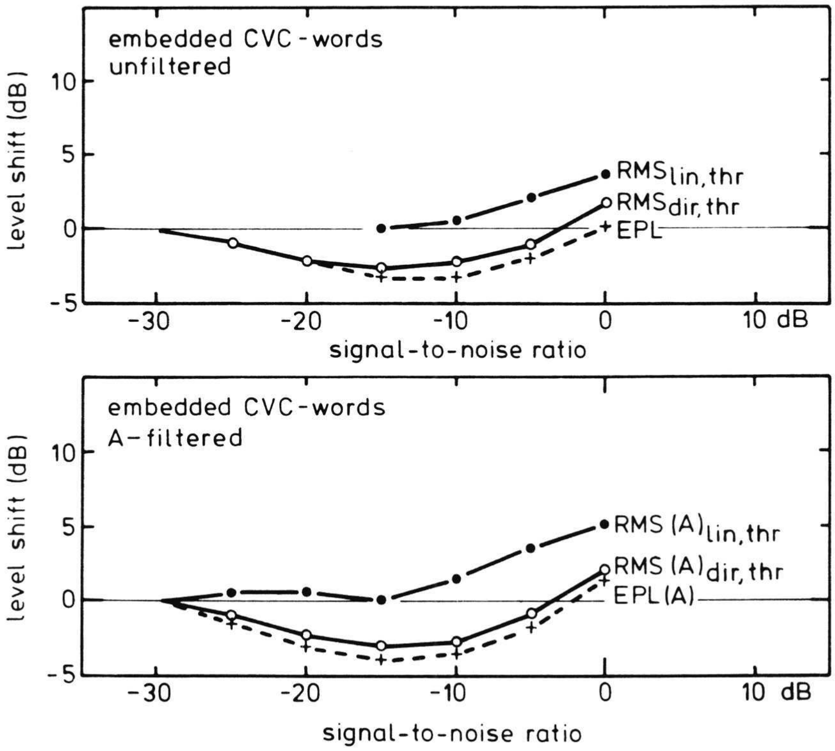

Fig. 7 RMS$_{dir}$, RMS$_{lin}$ and EPL level shifts as a function of the signal-to-noise ratio of speech and additive noise of a 150-s speech and noise fragment of embedded CVC-words of one speaker.

decrease of 4 to 5 dB. It was verified that the effect of band pass limiting as obtained for "telephone speech" (300-3400 Hz) will be less than 2 dB, if an A-weighted level measure is used.

The effect of background noise (pink noise spectrum) was studied by determining the level shift as a function of the signal-to-noise ratio. In Fig. 7 this level shift is given for the RMS$_{dir}$, RMS$_{lin}$ and EPL. The threshold value for the RMS$_{dir}$ and the related EPL was 15 dB below this RMS$_{dir,thr}$ value, for the RMS$_{lin}$ the threshold value was just above the amplitude distribution of the noise as described in paragraph 3.1.

Fig. 7 shows that the level shift for the RMS$_{dir}$ and EPL varies considerably for a signal-to-noise ratio between -30 dB and -5 dB. This negative level shift can be explained by the increase of the number of "above-threshold samples" while the sum of squares of the contributing samples (determined by the upper part of the histogram) does not change significantly. The RMS_{lin} based on the envelope of the signal does not change for signal-to-noise ratios up to -10 dB.

3.3 Relations with speech intelligibility

Speech level measures are often used in conjunction with speech intelligibility tests such as for the specification of signal-to-noise ratios or to adjust the speech level of individual speakers. For the CVC-words of the four speakers used in this study the intelligibility scores as a function of the signal-to-noise ratio for four different types of noise was known. Therefore we could calculate for each individual speaker the level correction for each measure to an equal CVC-score.

The variation among these corrections was calculated for all the level measures separately. In Table I this variation, expressed by the standard deviation, is given.

The envelope peak-reading measures show the closest relation with the intelligibility. These methods have the advantage that only the level of the test words can be measured and not the contribution of the carrier phrases. Generally measures based on A-filtered speech samples show a slightly better relation with intelligibility than the unfiltered speech samples.

4 DISCUSSION AND CONCLUSIONS

The robustness of a speech level measure depends mainly on the ability of the elimination of aspects of the speech signal which may contribute to the level but are in fact irrelevant. Such aspects are: silent periods between speech utterances (e.g. a comparison between connected discourse and word lists) and the effect of distortions such as background noise and band-pass limiting.

Another relevant feature is the relation between the speech level measure and the intelligibility. Of less importance but useful for a signal-to-noise ratio definition is the application of the level measure to non-speechlike signals such as noise or periodical signals.

- **The influence of silent periods and background noise**

(Pearsons and Horonjeff, 1982; Pearsons, 1983)
Some measures, such as reading peak deflections, are independent of the influence of silent periods. For other measures, the application of a threshold excludes the contribution of silent periods. A threshold above the level of the background noise can, to some extent, reduce the contribution of noise. This holds especially for envelope sampling methods ($RMS_{lin,sq}$) where, up to a signal-to-noise ratio of -10 dB, a good separation between speech and noise can be obtained (see Fig. 7). This separation is not obtained for direct sampling methods (RMS_{dir}; EPL).
The EPL, based on the shape of the upper part of the level distribution histogram, is to a large extent insensitive to the duration of silent periods but for noise conditions up to a signal-to-noise ratio of -10 dB, a level shift of -4 dB can be found; this value is similar to the level shift as obtained for the RMS_{dir} method.
Methods taking into account the peak levels of the signal ($\overline{dB(A)}_{fast}$) are, in principle, independent of the silent periods and, to a large extent, also of background noise. Even the contribution from the carrier phrases can be omitted.

- **The influence of band-pass limiting**

The level of an unfiltered speech signal mainly depends on the frequency components below 500 Hz. A high-pass filter as is used in telephone communications will, therefore, introduce a significant change in the level value. As indicated in Table I, the influence of A-filtered versus unfiltered speech results in a level decrease of 4 to 5 dB. The effect of an additional band-pass limiting as obtained on telephone circuits (300-3400 Hz), will be less than 1 dB, if an A-weighted level measure is used.

- **Relation to intelligibility**

The adjustment of the speech signal levels from different speakers to an equal intelligibility criterion can be done by considering A-filtered speech samples. Although Table I, column 4 indicates that the envelope peak reading methods have the closest relation with intelligibility, the measurement accuracy due to meter readings is less than with the sampling methods. A test with eight observers showed a standard deviation in meter readings of 2.1 dB.

From the considerations described above we may conclude that the level of a speech signal can best be described by an envelope measuring method such as RMS_{sq} or RMS_{lin} on with the EPL based on direct sampling. However the use of a direct sampling method has restrictions at low signal-to-noise ratios. A frequency weighting such as an A-filter makes the level measure more robust in conditions with band pass limited speech signals.
Only the RMS values are preferred to measure non-speechlike signals as noise and calibration signals.

REFERENCES

Berry, R.W. Speech Volume measurements on telephone circuits. Proc. IEE, Vol. 118, No. 2 (february 1971), 325-338.

Brady, P.T. Equivalent peak level: A threshold independent speech level measure. J. Acoust. Soc. Amer. 44, (1968), 695-699.

British Telecom Research Labs. Speech Volume Meter: type SV 6. Martlesham Heath, Ipswich, England.

Heusden, E.V. van, Plomp, R., and Pols, L.C.W. Effect of Ambient Noise on the vocal output and the preferred listening level of conversational speech. Applied Acoustics 12, 1979.

Kryter, K.D. The effects of noise on man. Academic Press (1970).

Pearsons, K.S., and Horonjeff, R.D. Speech levels in the presence of Time Varying Background Noise. NASA contract report 3547, (1982), Langley Research Center NAS1-16521.

Pearsons, K.S. Standardized methods for measuring speech levels. Proc. 14th Int. Congress "Noise as a public health problem". Vol. 1, Turin 1983, 465-476.

Steeneken, H.J.M., and Houtgast, T. Comparison of some methods for measuring speech levels. Report IZF 1978-22, Institute for Perception TNO, Soesterberg, The Netherlands.