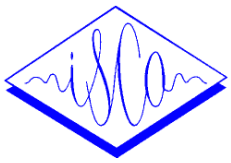

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

PROPOSAL OF A COMPOSITE MEASURE FOR THE EVALUATION OF NOISE CANCELLING METHODS IN SPEECH PROCESSING

R. LE BOUQUIN, G. FAUCON, A. AKBARI AZIRANI

Laboratoire de Traitement du Signal et de l'Image
Université de Rennes I - Campus de Beaulieu - 35042 Rennes cedex - FRANCE

ABSTRACT

This work addresses the problem of evaluating noise reduction techniques for hand-free telecommunications. It summarizes objective quality measures that are used in such a context. Finding a high correlation between objective and subjective tests remains a key point. In this paper, we propose a composite measure that appears more correlated with listening tests than classical ones. Its parameters are first determined by a training period and then this measure is validated using different noise cancelling approaches.

***Keywords** : speech enhancement, objective quality measures*

1. INTRODUCTION

Noise reduction is needed in speech enhancement for hand-free telecommunications to get a good-quality signal more pleasant to listen to. Noise cancelling approaches perform a processing on one or several channels to give a speech signal with a minimum residual noise and a slight distortion. However, such methods may bring some additional noise, called "musical noise", which must be negligible. If the evaluation of such methods may be done by listening tests, these tests must be rigorous and are quite expensive. Moreover, conditions such as the language choice, the test duration and listener fatigue can lead to biased results ; the listeners may be more or less trained to appreciate speech quality and so, it is difficult to obtain reproducible results. On the other hand, objective measures can provide near real-time estimation of voice quality in comparison with days or weeks required for subjective tests. Therefore, we must propose some objective measures which are easy to perform on the enhanced signals and are highly correlated with the subjective tests at the same time. Now to compute the objective measures, the clean speech signal must be available and used as the reference : to this end, speech signals and noises are recorded separately and then added together ; of course, this procedure does not take the Lombard effect into account. Nevertheless, we can judge whether the speech signal is well transmitted and the disturbing noise reduced.
In section 2, we give an overview of objective measures that are currently used ; then, we present a new composite measure which correlates better with subjective results than the conventional ones. In section 4, we indicate how we have conducted our listening tests. Finally, we present practical results and some concluding remarks.

2. CLASSICAL OBJECTIVE MEASURES

If most of objective measures have been proposed in the evaluation of coders, only a few have been tested in the noise reduction problem [1, 2, 3, 4]. These include, for example, the signal-to-noise ratios, the distance measures, the coherence function and the information index. Among these measures, we have tested those encountered frequently in the literature. Note that, if the objective tests have been used to evaluate noise reduction, the validity of these tests has not been proven yet nor the correlation with the subjective tests.
To estimate the enhancement brought by a method, we can compute objective measures at the input, where x represents the clean speech signal and x' the noisy speech signal, and at the output of the noise reduction processing, where x represents the clean speech signal and x' the signal estimated by this processing.
In the case of several microphones, the choice of the reference signal may be different. If the method estimates one of the signals received on the sensor, this signal becomes the reference. On the other hand, when the observations are combined to give an estimated signal, this one is compared to a signal derived by averaging all input clean signals. Let's recall briefly the expressions of the objective measures [5, 6, 7] we have retained :

• *The gain G*
The segmental SNR, proposed by Noll, does indeed correlate better with subjective results than the conventional SNR :

$$SNR_{seg} = \frac{1}{L}\sum_{k=0}^{L-1} 10\log_{10}\left(\frac{\sum_{i=0}^{M-1} x(i+Mk)^2}{\sum_{i=0}^{M-1}(x(i+Mk)-x'(i+Mk))^2}\right) \quad (1)$$

The speech signal is divided into L frames consisting of M samples each. The SNR_{seg} is obtained by averaging the signal-to-noise ratio in dB over all frames. A classical value of M is 256 (32ms for a sampling frequency f_s = 8kHz). The gain G is obtained by subtracting the input segmental SNR from the output segmental SNR.
Let's indicate that the primed variables refer to the enhanced speech signal afterwards.

• *The cepstral distance*
This distortion measure is a truncated version of the L_2 norm of the log spectral distortion measure between x and x' :

$$d_{cep} = \sum_{n=1}^{2p} \left(c_n - c_n'\right)^2 \qquad (2)$$

where c_n and c_n' are the cepstral coefficients corresponding to x and x' respectively ; the term c_0, corresponding to the gain of the model, has been dropped in the expression of d_{cep} and p is the order of the model (set to 8 in our application).

• *The Weighted Likelihood Ratio (WLR) distortion measure*
This measure is evaluated in the time domain using the auto-correlation and cepstral coefficients :

$$d_{WLR}^{N} = \sum_{n=1}^{N} \left(\frac{r_n}{r_0} - \frac{r_n'}{r_0'} \right) \left(c_n - c_n'\right) \qquad (3)$$

One thing that should be kept in mind is that, unlike the cepstral distortion measure which maintains its positive definiteness after truncation, the truncated d_{WLR} is no longer guaranteed to be positive definite. It seems that in practice, using N = 2p, we do not encounter any problems resulting from non-positive definiteness on the part of d_{WLR}^{N}.

• *The Itakura-Saito distortion measure*
It takes the following form :

$$d_{IS} = \frac{\sigma^2}{\sigma'^2}\frac{\delta}{\alpha} + \ln\frac{\sigma'^2}{\sigma^2} - 1 \qquad (4)$$

$$\text{where } \delta = a'^{T} R a' \,,\ \alpha = a^{T} R a$$

$$a^{T} = \left[1, a_1, \ldots, a_p\right] \,,\ a'^{T} = \left[1, a_1', \ldots, a_p'\right]$$

where a and a' are the AR model coefficients vectors, R is the (p+1)x(p+1) input sample autocorrelation symmetric Tœplitz matrix whose first row consists of (p+1) autocorrelation values of the signal from zero to p time lags, i.e. $[r_0, r_1, \ldots, r_p]$ and σ, σ' are the gain terms of the signals.

• *The Likelihood Ratio (LR) distortion measure*
An alternative choice is to set the gain terms, σ and σ', so that the test and reference patterns are compared with each other solely on the basis of their spectral shapes i.e. set $\sigma = \sigma'$. The resulting distortion measure is called the likelihood ratio distortion measure and is represented as :

$$d_{LR} = \frac{a'^{T} R a}{\alpha} - 1 \qquad (5)$$

• *The MOS derived from the information index*
The information index was developed by J. Lalou [8] and accounts for transmission loss, circuit noise, room noise, attenuation, frequency distortion and sidetone. The auditory system is modeled by dividing the spectrum into 16 critical bands, and applying empirical frequency weights and hearing thresholds for each band. The signal-to-distortion ratio (SDR), denoted QS(i), is computed first :

$$QS(i) = 10\log_{10} \frac{\sum_{j \in b_i} \left|X(f_j)\right|^2}{\left| \sum_{j \in b_i} \left|X(f_j)\right|^2 - \sum_{j \in b_i} \left|X'(f_j)\right|^2 \right|} \qquad (6)$$

where j ranges over all frequencies specified for the ith band, b_i. X(f) and X'(f) are Fourier transforms of the signals x and x'. The information index is given by :

$$RII = \sum_{i=1}^{16} W_2(i) \frac{3}{0.1 + 10^{-[QS_a(i) + W_1(i)]/10}} \qquad (7)$$

where $QS_a(i)$ is the average of QS(i) over all frames and $W_1(i)$ and $W_2(i)$ are tabulated weighting functions accounting for the hearing threshold and the perceptual importance of the ith frequency band respectively. MOS is estimated from RII using the following mapping :

$$RIT = \ln\left(\frac{RII}{27.6 - RII}\right)$$

$$YT = 1.00356 RIT - 1.4027$$

$$MOS = \frac{3.4 e^{YT}}{1 + e^{YT}} \qquad (8)$$

3. PROPOSAL OF A COMPOSITE MEASURE

The objective measures presented previously are based upon the computation of a difference between the reference signal and the signal to be compared (the noisy/enhanced signal). This difference includes the distortion of the speech signal and the residual noise. We don't attempt to get (in the noise reduction methods) a speech signal which is strictly identical to the clean speech signal. A weak distortion is not necessarily prejudicial, as far as the signal remains intelligible (good quality and noise reduced). The same importance cannot be awarded to the residual noise and to the distortion.
In the method we present now, we evaluate the distortion and the noise reduction separately and these two quantities are combined to give a new measure. Such a procedure is applied as long as the method to be tested is equivalent to a filtering applied to the observations. To evaluate the different measures, let's recall that noise and speech are recorded separately so that we can compute the signal distortion as well as the noise reduction. The procedure takes the following form :
- from the noisy observation, we deduce the filtering used in the signal processing procedure ;
- this filtering is applied to the clean speech signal (x) to obtain the filtered signal x_f, whose distorted part may be written $x_f - x = \varepsilon_x$, and to the disturbing noise (n) to obtain the residual noise ε_n.
We compute a segmental distortion of x, D :

$$D = \frac{1}{L} \sum_{k=0}^{L-1} \left(\frac{\sum_{i=0}^{M-1} \varepsilon_x(i + Mk)^2}{\sum_{i=0}^{M-1} x(i + Mk)^2} \right) \qquad (9)$$

In the same manner, we compute a segmental noise reduction factor R, defined as follows :

$$\frac{1}{R} = \frac{1}{L} \sum_{k=0}^{L-1} \left(\frac{\sum_{i=0}^{M-1} \varepsilon_n(i + Mk)^2}{\sum_{i=0}^{M-1} n(i + Mk)^2} \right) \qquad (10)$$

These quantities, evaluated on L blocks, are now combined to define the new measure M-:

$$M = \alpha H_{\beta}^{\gamma}(D) + 1/R \qquad (11)$$

$$\text{where } H_{\beta}(u) = \begin{cases} u-\beta \text{ if } u>\beta \\ 0 \quad \text{otherwise} \end{cases}$$

That means, by using (11), we don't consider the distortions lower than β. Of course, the lower the measure the more performant the method. The parameters α, β and γ will be determined using a learning phase, by looking for correspondence between the informal listening tests and this new measure.

4. INFORMAL LISTENING TESTS

It will be necessary to conduct listening tests to find some correspondence with the objective measures. These listening tests have been conducted in the following manner :
- speech signals (corresponding to phonetically balanced sentences) and noises have been recorded in a stopped car and in a moving car respectively and are added together to get the noisy speech files.
- different noise reduction methods are applied to the noisy speech signals.

The tape recording is presented like this :
- the clean signal is first recorded and repeated 3 times ;
- then, the noisy speech signal is also repeated 3 times ;
- the two previous sequences are repeated again ;
- the different results, which are repeated three times consecutively, are recorded three times at random.
- the listening tests are conducted with ten listeners ; each listener has to appreciate the distortion, the residual noise and the defaults brought by the processing and to give a global note. Notes have a range of 0 to 5 points : for each method, a final note is computed by averaging the 30 notes given by all listeners.

5. PRACTICAL RESULTS

a) Learning phase

The aim of this phase consists in determining the parameters α, β, γ of the measure M. Various noise cancelling algorithms have been applied to our noisy speech files ; in this training phase, we choose more or less performant noise reduction methods to extract the objective measures which appear strongly correlated with the listening tests. The methods, notated A, B, C, D, E, F, G, H and I, are not described there : we are only interested in the agreement between objective and subjective tests and not in the efficiency of each method. Then, we compute the objective measures defined in sections 2 and 3. The processed sentences are also recorded on a tape-recorder and we performed the listening tests described previously. An averaged score (between 0 and 5) is given for each method. The table I gives the results obtained with the different methods for one sentence (this sentence is : "il se garantira du froid avec ce bon capuchon") ; the disturbing noise is recorded in a Renault 25 at 90km/h and the input SNR is 5dB. The parameters used in the measure M are as follows : $\alpha = 1/0.3$, $\beta = 0.3$, $\gamma = 2$. These values lead to the best matching between objective and informal subjective tests. The right column corresponds to the subjective tests ordered from the best result to the worst one.

G	d_{cep}	d_{LR}	d_{WLR}	d_{IS}	MOS-RII	M	List. tests
2.18	0.205	0.237	0.161	11.49	1.12	0.094	C(3.81)
1.84	0.132	0.143	0.111	3.674	1.052	0.08*	H(3.66)
2.60	0.168	0.182	0.134	1.512	1.119	0.13*	I(3.46)
2.02	0.359	0.428	0.267	3.760	0.991	0.109	F(3.05)
4.13	0.127	0.162	0.101	1.043	1.469	0.109	B(2.96)
1.12	0.432	0.541	0.325	11.60	0.876	0.158	G(2.61)
3.47	0.282	0.310	0.216	0.628	1.242	0.202	E(2.52)
4.15	0.118	0.127	0.109	0.337	1.495	0.21	A(1.93)
3.46	0.237	0.245	0.194	0.341	1.305	0.307	D(1.85)

Table 1

It appears that no classical objective measure follows the evolution of the subjective tests. On the other hand, the new measure M seems more correlated with these listening tests except in the case of the methods marked with an asterisk. An explanation may be given in such cases : the processing is a multi-channel approach and the estimated signal is a combination of the two input speech signals ; so, even if the informal listening tests give satisfying results, the objective measures compare the output signal to one of the two clean speech signals.

It is obvious that the choice of parameters is not definitive ; we have to perform noise cancelling for a number of configurations (i.e. various sentences and noises, different input SNRs ...) and to conduct more rigorous listening tests. Moreover, the tested methods are not always sufficiently performant and a comparison between more efficient methods could yield other optimal parameters. Other functions giving a new composite measure can also be developed and we must keep in mind that the work to find a measure which is highly correlated with the listening tests is an all-important problem.

b) Results validation

Since the classical objective methods are not satisfactory, we think it is interesting to validate the measure M on other noise cancelling methods with other environmental conditions, while maintaining the parameters found during the training period. To limit the experimental study, we keep only two common objective measures, for instance the segmental gain and the cepstral distance, since they are simple and easy to implement, as well as the measure M. This second experiment has been conducted on eight mono-channel approaches (notated A' until G') with two noisy sentences ; the noise is recorded in a moving car (Renault 25) at 130km/h and the input SNR is 2dB. The results are given in table 2.

G	d_{cep}	M	List. tests
5.74	0.3756	0.1079	H' (4)
5.4	0.3298	0.1055	F' (3.5)
5.717	0.2771	0.1242	E' (3.4375)
2.987	0.3196	0.1481	C' (3)
5.918	0.2755	0.1331	G' (2.875)
3.073	0.4351	0.1588	B' (2.75)
5.817	0.2287	0.1553	D' (2.5625)
4.267	0.2457	0.1902	A' (0.5625)

Table 2

To have a better representation of our results, we represent the performance obtained with these three objective measures on a diagram (Figure 1) to compare it with that given by the informal listening tests. Each processing is pointed out according to its classification for each measure. The noise reduction methods are ordered on the x-axis from the best one to the worst one in accordance with the subjective tests.

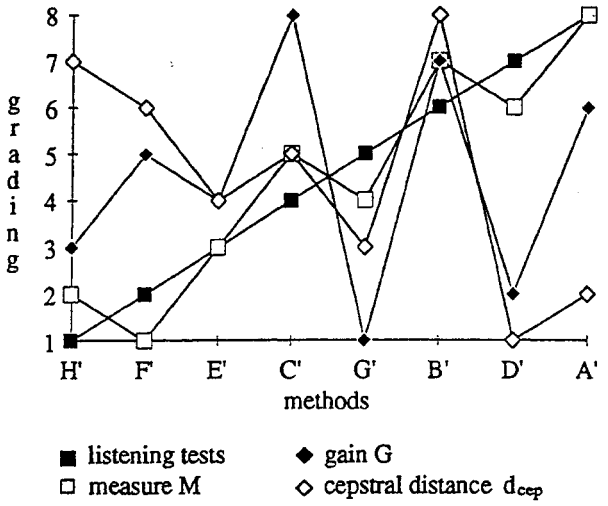

Figure 1

We observe that the curve depicting the measure M is the one which provides the best fit according to the listening tests. Another way to characterize the fitting between objective tests and informal listening tests lies in computing the correlation between these two kinds of measure [7]. This correlation is defined as :

$$\rho = \frac{\sum_p (s-\bar{s})(o-\bar{o})}{\left(\sum_p (s-\bar{s})^2 \sum_p (o-\bar{o})^2\right)^{1/2}} \qquad (12)$$

where p designs the processing, o and s are the objective and subjective results respectively. From the table 2, we find that $|\rho| = 0.319$ for the gain G and $|\rho| = 0.917$ for the measure M. These values tend to prove the correlation between our measure and the subjective tests but caution must be taken with these quantities because ρ is computed on a restricted number of results.

6. CONCLUSION

The results obtained by the composite measure are quite encouraging since they are better correlated with the listening tests than the common objective measures. The parameters of this measure have been determined during a training phase and then used in a validation step. Although the conditions are different, the measure M appears as a very interesting measure and its use may avoid long and expensive subjective tests. The latter can only be conducted when different noise reduction techniques yield very close results.
Our perspectives are the following :

- we can suggest other combinations of functions to get a new composite measure ;
- in the evaluation of the combined measure M, the total noise is considered in the computation of the noise reduction factor R. In fact, a part of the output noise can be masked by the signal. The problem is to find a masking curve (in the frequency domain) to determine if there is inaudible noise. Two models [9] to compute the masking curve are currently studied in psychoacoustics and coders, the ISO 1 model (german school) and the ISO 2 model (american school), near by the physiology of the human ear. Another method, called Perceval, is derived from the noise injection model described by B. Paillard [10], who develops a noise injection model and shows that a noise whose spectrum is at least 13dB below the signal spectrum remains inaudible. So, the inaudible noise won't be taken into account in the computation of the noise power ;
- the separate evaluation of the noise reduction factor and the distortion factor requires that noise reduction algorithms can be set in the form of a filtering applied to the noisy observation. Another possibility is to perform the evaluation directly on the enhanced signal, to find one measure that reveals the distortion and another measure mainly sensitive to the noise and then combine them.

REFERENCES

[1] M.S. Ahmed, "Comparison of Noisy Speech Enhancement Algorithms in Terms of LPC Perturbation", IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 37, n°1, pp. 121-125, Jan. 1989.

[2] J.H. Hansen and M.A. Clements, "Iterative Speech Enhancement with Spectral Constraints", ICASSP, Dallas, pp. 189-192, 1987.

[3] H.L. Nguyen Thi, C. Jutten, J. Caelen, "Speech Enhancement : Analysis and Comparison of Methods on Various Real Conditions", Signal Processing VI, EUSIPCO 92, pp. 303-306, Brussels, 1992.

[4] E. Masgrau, J.A. Rodriguez-Fonollosa, A. Ardanuy, "Enhancement of Speech by Using Higher-Order Spectral Modelling", Signal Processing VI, EUSIPCO 92, pp. 307-310, Brussels, 1992.

[5] R.F. Kubichek, "Standards and Technology Issues in Objective Voice Quality Assessment", Digital Signal Processing, 1, pp. 38-44, 1991.

[6] N. Nocerino et al., "Comparative Study of Several Distortion Measures for Speech Recognition", Speech Communication 4, pp. 317-331, 1985.

[7] S.R. Quackenbush et al., "Objective Measures of Speech Quality", Prentice-Hall, 1988.

[8] J. Lalou, "The Information Index : an Objective Measure of Speech Transmission Performance", Annales des Télécommunications, 45, n°1-2, pp. 47-65, 1990.

[9] ISO standard project : "Coding of Moving Pictures and Associated Audio for Digital Storage Media at about 1.5Mbits/s", ISO/IEC JTC1/SC2/WG11 MPEG 92/0, Committee Draft, Mar. 1992.

[10] B. Paillard, "Codage Perceptuel des Signaux Audio de Haute Qualité", Sherbrooke University, Canada, Feb. 1992.

This work was partly funded by the EEC under contract ESPRIT 6166 FREETEL "Enhancement of Handsfree Telecommunications".