EXTRACTION OF SPEECH IN ACOUSTICAL NOISE BY MARKOV FILTERING

Y.N.PROKHOROV

A.V.MININ

Moscow Telecommunication Institute, USSR, 111024

ABSTRACT

This paper presents a general approach to the improvement of speech intelligibility in broad band acoustical noise. By using the methods of Markov filtering the digital processing algorithms of noise-added speech are being synthesized and their experimental study is being carried out.

INTRODUCTION

The telephone communication systems and the systems of automatic man-machine communication by voice often operate in a severe broad band acoustical noise situations. The organising protective measures and the compensation techniques do not always provide the effective noise suppression. In such cases the signal-noise ratio (SNR) of the microphone output may be 0-3 db, and the intelligibility S may be 40-50% /1,2/. The special digital processing for noise reduction is applied but it doesn't allow to increase intelligibility sufficiently so far /1,2/. The aim of this paper is to develop the effective processing methods by using Markov filtering.

FORMULATION OF THE PROBLEM

In interval the duration of which is about 20 - 50 ms the mixture of signal and noise is

$$z_t = x(\vec{\lambda},t)+n_t\ ,\ t=0,\pm 1,\pm 2,\dots, \qquad (1)$$

where $\vec{\lambda} = (\lambda_1,\lambda_2,\dots,\lambda_m)^T$ – is a vector of parameters describing the articulation apparatus state (ArA); $x(\vec{\lambda},t)$, n_t – the sample sequences of speech signals (SS) and noise. Because of the low accuracy of articulatory organs the parameters take a continuous set of values.

For the automatic recognition (reception) of corrupted speech the values of parameter $\vec{\lambda}$ should be classified. However, in the process of extraction it is quite enough while using z_t to formulate such signal $u(\vec{\lambda}^*,t)$ on hearing of which the maximum intelligibility is achieved: $S^{(z)}_{max} = S^{(u)}$, where $S^{(z)}$, $S^{(u)}$ – is the intelligibility of signals z_t and $u(\vec{\lambda}^*,t)$ respectively. Since a vector $\vec{\lambda}^*$ or an unknown function $g(\vec{\lambda}^*)$ is classified in the process of human perception, the value of $\vec{\lambda}^*$ should be chosen in such a way that $\varepsilon^2_\lambda = E[\vec{\lambda}^* - \vec{\lambda}]^T Q_\lambda [\vec{\lambda}^* - \vec{\lambda}] = min$ where E – mathematical expectation operation, Q_λ – a weighted coefficient matrix. The minimum attainable value ε^2_λ is defined by Kramer-Rao's inequality.

So the problem of speech extraction is interpreted as the construction of the best estimation of $\vec{\lambda}$ and the creation of the signal $u(\vec{\lambda}^*,t)$ with $S^{(u)}=max$. A general diagram of the extraction device is given in fig.1, where CD is a controlling human ear perception system device.

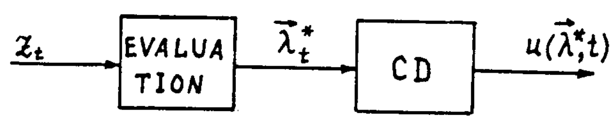

Fig.1. A General Diagram of Speech Extraction

SIGNAL AND NOISE MODELS

For the best evaluation of $\vec{\lambda}$ the adequate models of signal and noise are required. The simplest model of the broad band noise is a Gaussian sequence n_t with $En_t=0, En_t^2=G_n^2, En_{t_1}n_{t_2}=0, \forall (t_1 \neq t_2)$. The more precise model is the process of autoregression

$$v_t=\sum_{i=1}^{L}\alpha_i v_{t-i}+\mu_t, \quad L=2\div 10, \qquad (2)$$

where α_i is evaluated a'priory by the noise realization by means of a least-square technique with limitations. One can use the orthogonal projections as the forms of limitations/3/.
The signal is modelled by a nonlinear autoregressive process

$$y_t=\beta\varphi(y_{t-1})+b_y\eta_{t-1}, \qquad (3)$$

$$x_t=\sum_{i=1}^{m}a_i x_{t-i}+cy_{t-i}+b\xi_{t-1}, \qquad (4)$$

$$\vec{\lambda}=(\beta,a_1,a_2,\dots,a_m)^T, \quad m=2\div 10.$$

The function φ is found on the synthesis stage from the condition:

$$\min E|\varepsilon_x(t)\varepsilon_x(t-\tau)|,$$

$$\varepsilon_x(t)=x_t-\beta\varphi(x_{t-1}), \quad \tau=const.$$

The result is shown in fig.2.

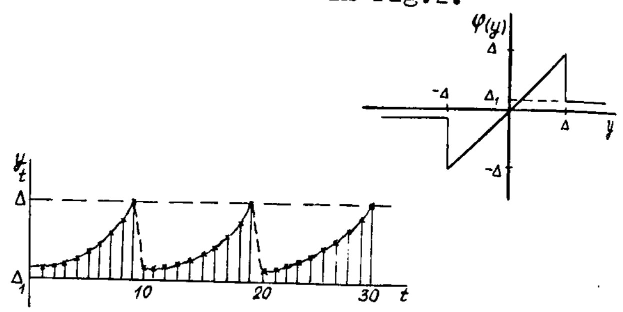

Fig.2. The function of Excitation

To reduce the number of the parameters evaluated model (3), (4) may be written in the following way:

$$x_t=\sum_{j=1}^{N}\lambda_j\psi_j(x_{t-1},x_{t-2},\dots,x_{t-m})+b\xi_{t-1},$$

where $N=2$-6 and the orthogonal functions ψ_j are found experimentally. To do this on the speech signal of a concrete speaker the set of the parameters values is defined in models (3),(4), then a set of autoregression functions is given

$$\psi(x_{t-1},x_{t-2},\dots,x_{t-m}),$$

and the Karunen-Loev basis is built for it.

THE EVALUATION OF THE POSSIBLE INTELLIGIBILITY

The intelligibility $S^{(z)}_{max}$ may be evaluated in the presence of noise with an average flat spectrum. Consider

$$z_t^{(1)}=x(\vec{\lambda},t)+n_t, \quad z_t^{(2)}=f[x(\vec{\lambda},t)]+n_t.$$

Chose the function f so that $S^{(z_2)}\approx S^{(z_1)}_{max}$. For example f may be the central clipping function. According to the articulatory tests $S^{(z_1)}, S^{(z_2)}$ may be found for different SNR_1, SNR_2 and the threshold of the clipping x_{thr}. Putting down the Kramer-Rao's inequalities the formulas for $\varepsilon^2_{\lambda 1}(SNR_1)$ and $\varepsilon^2_{\lambda 2}(SNR_2)$ can be obtained. In the situation where $\varepsilon^2_{\lambda 1}(SNR_1)=$

$=\varepsilon^2_{\lambda 2}(SNR)$ we can find a family of curves with the equal reception accuracy: $SNR_1=\varphi_\varepsilon(SNR_2)$ with parameter x_{thr}. Now the estimation $\hat{S}^{(z_1)}_{max}$ can be obtained in the following way: by using SNR_1, x_{thr} and function φ_ε we can find SNR_2, where $\varepsilon^2_{\lambda 1}(SNR_1)=\varepsilon^2_{\lambda 2}(SNR_2)$. Then $\hat{S}^{(z_1)}_{max}=S^{(z_2)}(SNR_2)\le S^{(z_1)}_{max}$ and $\Delta S^{(z_1)}=\hat{S}^{(z_1)}_{max}-S^{(z_1)}$ is a possible benefit in intelligibility in digital processing of the corrupted speech $z_t^{(1)}$.

THE FILTERING ALGORITHMS

For the simplification we can take $g(\vec{\lambda})$ as a mutually unique continuous (unknown) function. It can be shown that in this case $u(\vec{\lambda}^*,t)=x(\vec{\lambda}^*,t)$ and instead of CS we may use a speech synthesizor operating according to (3), (4). Thus, the extraction of speech is performed according to the algorithm No.1: "estimation of $\vec{\lambda}$ - synthesis $x(\vec{\lambda}^*,t)$" (analysis-synthesis). If $bE\xi_t^2 << Ex_t^2$, then the algorithm No.1 is very close in effectiveness to the mutual evaluation of $x, \vec{\lambda}, y$ or to the "adaptive filtering" - the algorithm No.2. If there are pauses in conversation and the consonants in speech, then the algorithm No.3 "a mutual evaluation of parameters, filtering of speech and classification of tone-consonant-pause" is quite optimal.
The above mentioned algorithms are synthesized by the maximum of a'posteriory probability criterion in /4, 6/ by using Markov filtering technique /5/.

EXPERIMENTAL RESULTS

Testing of algorithms No.1-3 are performed on the speech signal with the sampling frequence 15 kHz and with the number of quantizing levels 2^{12}. In fig.3 the power spectral densities of the initial (G_x), the processed (G_x^*) signals and the noise-added speech (G_z) are shown for the word "действие" (algorithms No.3, 1). In fig.4 the curves of likelihood function Λ and the current signal power Ex_t^2 recieved on the articulatory tables of syllables without any pauses are shown. The probability of a classification error of tone-consonant-pause is about $3\cdot 10^{-2}$ with a zero threshold. In Table No. 1 the results of tests are shown, where ΔSNR is a benefit

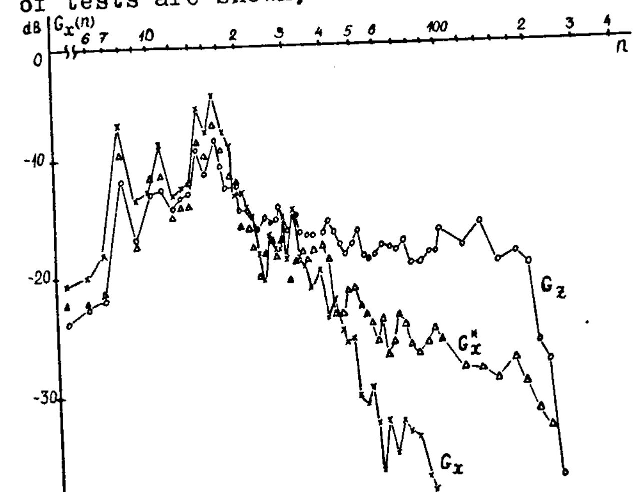

Fig. 3. Signal spectrum G_x, G_x^*, G_z

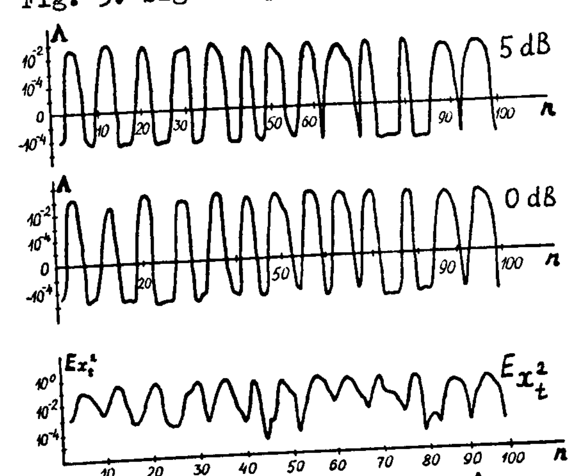

Fig.4. Likelihood Functions Λ

Table 1

No.	3-1	2	3
ΔSNR dB	7	6-7	10-12

in SNR , the number of algorithm 3-1 is a sequential application of the algorithms N3 and N1. These results are achieved for the noise with close to an average flat spectrum and model (3), (4). In pauses the mixture of z_t is multiplied to a coefficient $q < 1$. The coefficient of noise power in filters is chosen experimentally.

In Table No.2 the signal-error prediction ratio (SER) for models (4), (5) which is achieved on the initial speech signal is given. There the results of algorithm No. 2 with (5) in noisy environment because of the engines operation.

In Table П is percentage of the real favours given by the listeners to the processed signal. The number of listeners is 20-25.

Table 2

Model	SER, dB, for N			П	Δ SNR
	2	4	6	%	dB
(4)	21,3	24,7	26,5	-	2-3
(5)	26	27	27,3	85-88	3-5

CONCLUSION

The method of intelligibility improvement in noisy environment is worked out. The theoretical benefit of the digital processing for noise with long-term average flat spectrum is evaluated. By using the Markov filtering techniques the algorithms of mutual speech filtering, the parameters evaluation and the classification of tone-consonant-pause are developed. The algorithms provide the improvement of the corrupted speech intelligibility in broad band noise and can be technically done on the mikroprocessor devices Am 2900.

BIBLIOGRAPHY

1. H.Suzuki, J.Igurashi, Y.Ishii."Extraction of Speech in Noise by Digital Filtering", J.A.S. of Japan, v.33, No.8, 1977, pp.405-411.

2. D.Graupe, J.Grosspietsh, S.Basseas. "Self Adaptive filtering of Environmental Noises from Speech", Proc. AIAA/IEEE 6-th Dig. Avionics Syst. Conf., Bultimor, MD, 1984, N.Y., 1984, pp.263-269.

3. А.В.Минин, Ю.Н.Прохоров. "Оценка параметров речевых сигналов методом наименьших квадратов с ограничениями", Электросвязь, №3, 1986, с.26-29.

4.Ю.Н.Прохоров. "Статистические модели и рекуррентное предсказание речевых сигналов", Радио и связь, 1984. - СТС, вып.20.

5. A.P.Sage, J.L.Melse. Estimation Theory with Application to Communication and Control, N.Y., McGraw-Hill, 1972.

6.М.В.Назаров, Ю.Н.Прохоров. "Методы цифровой обработки и передачи речевых сигналов", Радио и связь, 1985.