

ELIMINATION OF CLICKS FROM ARCHIVE SPEECH SIGNALS USING SPARSE AUTOREGRESSIVE MODELING

Maciej Niedźwiecki and Marcin Ciołek

Faculty of Electronics, Telecommunications and Computer Science, Department of Automatic Control
Gdańsk University of Technology, ul. Narutowicza 11/12, Gdańsk, Poland
maciekn@eti.pg.gda.pl, marcin.ciolek@pg.gda.pl

ABSTRACT

This paper presents a new approach to elimination of impulsive disturbances from archive speech signals. The proposed sparse autoregressive (SAR) signal representation is given in a factorized form – the model is a cascade of the so-called formant filter and pitch filter. Such a technique has been widely used in code-excited linear prediction (CELP) systems, as it guarantees model stability. After detection of noise pulses using linear prediction, the factorized model is converted into a generic sparse form in order to perform a projection-based signal interpolation. It is shown that the proposed algorithm is able to deal favorably with speech signals with strong glottal activity, which is a serious problem for algorithms based on the classical AR modeling.

Index Terms— Sparse autoregressive models, reconstruction of speech signals, elimination of clicks

1. INTRODUCTION

Archived audio recordings are often degraded by impulsive disturbances. Clicks, pops and record scratches are caused by aging and/or mishandling of the surface of gramophone records (shellac or vinyl). In case of magnetic tape recordings impulsive disturbances can be usually attributed to transmission or equipment artifacts (e.g. electric or magnetic pulses) – for more details see e.g. [1].

There are many methods that allow to eliminate noise pulses from audio signals [1], [2]. Most of these techniques are based on autoregressive (AR) signal modeling and model-based adaptive prediction: an on-line identification of the AR model of audio signal is carried out and its results are used to predict new samples from the old ones. If the magnitude of the prediction error is too large (e.g. if it exceeds three standard deviations of its nominal value), the sample is classified as an outlier and scheduled for interpolation. As shown in [3], the task of simultaneous identification/detection/interpolation can be stated as a nonlinear filtering problem and solved using the theory of extended Kalman filter (EKF).

Although the procedure described above yields very good results when applied to music recordings, it often fails on archived speech signals (e.g. historic speech recordings), especially those with strong voiced episodes. The reason is not difficult to guess. Since voiced speech sounds are formed by means of exciting the vocal tract (represented by the AR model) with a periodic train of glottal air pulses, the outlier detector is prone to confuse pitch excitation with noise pulses. One way of coping with this problem, worked out in some detail in [4], is based on blocking or desensitizing the click detector at the moments of expected pitch activity. However, the obvious disadvantage of this approach is that it works poorly in all cases where noise pulses happen to coincide with glottal pulses. In this paper we present a new procedure for declipping speech signals, which is free of the drawback mentioned above.

2. CLASSICAL AR APPROACH

Elimination of clicks can be handled by using classical AR modeling. In this approach the sampled audio signal $y(t)$ is represented by the following AR model of order p

$$y(t) = \sum_{i=1}^p a_i y(t-i) + n(t) \quad (1)$$

where $t = \dots, -1, 0, 1, \dots$ denotes normalized (dimensionless) discrete time, $a_1, \dots, a_p$ are the so-called autoregressive coefficients and $n(t)$ denotes white driving noise.

2.1. Prediction-based detection of noise pulses (outliers)

Model coefficients are continuously updated using a parameter tracking algorithm – such as exponentially weighted least squares (EWLS), least mean squares (LMS) or Kalman filter (KF) based [5] – which yields $\hat{a}_1(t), \dots, \hat{a}_p(t)$.

The first detection alarm is raised¹ when the magnitude of the AR model-based one-step-ahead prediction error exceeds μ times its estimated standard deviation (typically $\mu \in [3, 5]$)

$$|\varepsilon(t+1|t)| = |\hat{y}(t+1|t) - y(t)| > \mu \hat{\sigma}_{\varepsilon(t+1|t)} \quad (2)$$

¹ $d(t) = 1$ will further mean that the sample $y(t)$ is corrupted with a noise pulse; otherwise $d(t) = 0$

where

$$\hat{y}(t+1|t) = \sum_{i=1}^p \hat{a}_i(t)y(t-i), \quad \hat{\sigma}_{\varepsilon(t+1|t)} = \hat{\sigma}_n(t)$$

and $\hat{\sigma}_n^2(t)$ denotes the local estimate of the driving noise variance, obtained by means of averaging the recently observed squared one-step-ahead prediction errors (after excluding outliers).

$$\hat{\sigma}_n^2(t) = \begin{cases} \gamma \hat{\sigma}_n^2(t-1) + (1-\gamma)\varepsilon^2(t|t-1) & \text{if } d(t) = 0 \\ \hat{\sigma}_n^2(t-1) & \text{if } d(t) = 1 \end{cases}$$

The coefficient γ , $0 < \gamma < 1$, denotes the forgetting constant which determines the estimation memory of the averaging algorithm.

The detection process is continued for multi-step-ahead predictions until p consecutive prediction errors are sufficiently small, namely

$$|\varepsilon(t+k_0+i|t)| \leq \mu \hat{\sigma}_{\varepsilon(t+k_0+i|t)}, \quad i = 1, \dots, p \quad (3)$$

or until the length of the detection alarm k_0 reaches a prescribed value $k_{\max}$. The quantity $\hat{y}(t+k|t)$ can be obtained as a concatenation of k one-step-ahead predictions

$$\hat{y}(t+k|t) = \sum_{i=1}^p \hat{a}_i(t)\hat{y}(t+k-i|t) \quad (4)$$

where $\hat{y}(t+j|t) = y(t+j|t)$ for $j \leq 0$. The variance of the multi-step prediction errors can be evaluated recursively using the following algorithm proposed by Stoica [6]

$$\hat{\sigma}_{\varepsilon(t+k|t)}^2 = \hat{\sigma}_{\varepsilon(t+k-1|t)}^2 + \hat{\sigma}_n^2(t)f_{k-1}^2(t) \quad (5)$$

$$\begin{aligned} f_{k-1}(t) &= g_{k-1}^0(t) \\ g_k^i(t) &= g_{k-1}^{i+1}(t) + \hat{a}_{i+1}(t)f_{k-1}(t) \\ i &= 1, \dots, p-1 \\ k &= 2, \dots, k_{\max} \end{aligned}$$

with initial conditions: $\hat{\sigma}_{\varepsilon(t+1|t)}^2 = \hat{\sigma}_n^2(t)$, $f_0(t) = 1$ and $g_1^i(t) = a_{i+1}(t)$, $i = 1, \dots, p-1$.

When the detection process is finished, the sequence of irrevocably distorted samples $\{y(t+1), \dots, y(t+k_0)\}$ is interpolated using the available signal model (1). The projection-based interpolation is based on p samples preceding the missing block, and p samples succeeding the block [5]. In [3] all quantities needed to carry out the detection/ interpolation process are evaluated by the extended Kalman filter (EKF).

3. APPROACH BASED ON SPARSE AR MODELING

The sparse AR model of speech signal can be defined in the form

$$y(t) = \sum_{i=1}^r a_i y(t-i) + \sum_{j=\tau+1}^{\tau+s} a_j y(t-j) + n(t) \quad (6)$$

where the quantities τ ($\tau \gg r$) and s are chosen in such a way that $\tau+1 \leq T \leq \tau+s$, where T denotes the period of pitch excitation (if present). Even though formally of order $p = \tau+s$, such a model is sparse as it contains only $r+s \ll p$ nonzero coefficients.

Sparse AR models capture both short-term correlations [taken care of by the first component on the right-hand side of (6)] and long-term correlations [taken care of by the second component on the right hand side of (6)] of the analyzed time series. Hence, when appropriately fitted, such models can adequately represent both formant and pitch structure of speech signals. The main problem with the model (6) is that no identification algorithms seem to exist that can guarantee its stability. Since the order of the model $p = \tau+s$ is large (usually exceeding 100, even for moderate sampling rates), stability tests are also hardly practical.

3.1. Sparse AR model in a factorized form

Stability problems, mentioned above, can be easily solved if the sparse autoregressive (SAR) model is sought in the following factorized form, widely used for predictive coding of speech, e.g. in CELP coders [8]

$$y(t) = \sum_{i=1}^r \alpha_i y(t-i) + x(t) \quad (7)$$

$$x(t) = \beta x(t-T) + n(t) \quad (8)$$

Equation (7) describes the so-called formant filter, characterized by formant coefficients $\alpha_1, \dots, \alpha_r$, and equation (8) describes pitch filter, characterized by the pitch coefficient β . Note that the formant filter and the pitch filter form a cascade.

The factorized model (7) - (8) can be converted into the generic sparse form (6) by setting

$$\begin{aligned} a_i &= \alpha_i, \quad i = 1, \dots, r, \quad a_i = 0, \quad r < i < T \\ a_T &= \beta, \quad a_{T+i} = -\beta \alpha_i, \quad i = 1, \dots, r \end{aligned}$$

Stability of the factorized model is guaranteed if both filters (formant and pitch) are stable, which can be easily achieved using appropriate estimation tools and simple stability enforcement mechanisms [9].

Based on (7) - (8), the multi-step predictions can be evaluated using the following recursive algorithm

$$\begin{aligned} \hat{y}(t+k|t) &= \sum_{i=1}^r \hat{\alpha}_i(t)\hat{y}(t+k-i|t) + \hat{x}(t+k|t) \\ \hat{x}(t+k|t) &= \hat{\beta}(t)\hat{x}(t+k-T|t) + \hat{n}(t+k|t) \end{aligned}$$

where, for $j \geq 0$, $\hat{y}(t-j|t) = y(t-j)$ and $\hat{x}(t-j|t) = y(t-j) - \sum_{i=1}^r \hat{\alpha}_i(t)y(t-j-i)$. Estimation of the corresponding prediction error variances can be carried out using the algorithm (5), after converting the factorized model into the form (6).

When the detection process is finished, the sequence of irrevocably distorted samples $\{y(t+1), \dots, y(t+k_0)\}$ is interpolated using the sparse model (6). The projection-based interpolation is based on p samples preceding the missing block, and r samples succeeding the block.

3.2. Estimation of formant coefficients

To ensure the stability of the formant (short-term) model, the method of least squares with exponential data windowing (LSEW) is employed for the purpose of parameter tracking.

In this approach, the method of least squares is applied, at any time instant $t > r$, to the windowed data sequence $\{\lambda_0^{t-1}y(1), \dots, \lambda_0 y(t-1), y(t)\}$ where λ_0 , $0 < \lambda_0 < 1$, denotes the so-called forgetting constant. After extending this sequence with r zero samples at its beginning and at its end (the so-called autocorrelation technique), the LS estimate of the parameter vector $\theta(t) = [\alpha_1(t), \dots, \alpha_r(t)]^T$ can be expressed in the form

$$\hat{\theta}(t) = R^{-1}(t)s(t) \quad (9)$$

$$R(t) = \begin{bmatrix} \rho_0(t) & \cdots & \rho_{r-1}(t) \\ \vdots & \ddots & \vdots \\ \rho_{r-1}(t) & \cdots & \rho_0(t) \end{bmatrix}, \quad s(t) = \begin{bmatrix} \rho_1(t) \\ \vdots \\ \rho_r(t) \end{bmatrix}$$

where the recursively computable quantity

$$\rho_i(t) = \lambda_0^2 \rho_i(t-1) + \lambda_0^i y(t)y(t-i)$$

after scaling, can be interpreted as the local estimate of the i -th autocorrelation coefficient of $y(t)$.

Since the regression matrix $R(t)$ is, by construction, positive definite and Toeplitz, the estimates $\hat{\alpha}_1(t), \dots, \hat{\alpha}_r(t)$ can be evaluated using the well-known Levinson-Durbin algorithm [7].

The tracking properties of the LSEW algorithm (9) are similar to those of the classical EWLS algorithm equipped with the forgetting constant equal to $\sqrt{\lambda_0}$. However, unlike EWLS, the LSEW algorithm guarantees stability of the obtained model for all values of t . Note that while the quantities $\rho_0(t), \dots, \rho_r(t)$ should be continuously updated, the estimates $\hat{\theta}(t)$ can be evaluated on demand.

3.3. Estimation of pitch period

The usefulness of the SAR model (7) - (8) critically depends on precise knowledge of the pitch period T . In many cases detection/interpolation performance may drop significantly even if the estimated value of T differs from its true value by only

one or two samples. Since the pitch period may be subject to fast changes, its accurate estimation is a really challenging task.

Denote by $T_{\min}/T_{\max}$ the smallest/largest pitch lags that can be expected under a given sampling frequency, and let $\mathcal{T}_0 = [T_{\min}, T_{\max}]$. Our estimation scheme is based on multiple models. Four competitive estimates of the pitch period, given by

$$\hat{T}_j(t) = \arg \min_{T \in \mathcal{T}_0} S(t; T, k_j, f_j, g_j), \quad j = 1, \dots, 4, \quad (10)$$

are obtained by means of minimizing the sum of squared differences between the selected fragments of the analyzed speech signals

$$S(t; T, k, f, g) = \sum_{i=0}^{m-1} [f(t+k-i) - g(t+k-T-i)]^2$$

where

j	k_j	$f_j(t)$	$g_j(t)$
1	0	$\hat{y}(t)$	$\hat{y}(t)$
2	0	$\hat{y}(t)$	$y(t)$
3	10	$y_*(t)$	$\hat{y}(t)$
4	10	$y_*(t)$	$y(t)$

In the first two cases ($j = 1, 2$), one tries to find the best match between the m most recent samples (in our experiments m was set to 50) of the already declicked signal $\hat{y}(t)$ and the analogous sequence derived either from the past of the signal $\hat{y}(t)$ ($j = 1$) or from the past of the original (unprocessed) signal $y(t)$ ($j = 2$). The second variant helps one to avoid ‘serial’ detection errors which occur when a single incorrect decision – classification of a pitch pulse as an outlier – starts a chain of related ‘derivative’ decision errors.

In the last two cases ($j = 3, 4$), the reference frame, denoted by $y_*(t)$, is made up of the $m-10$ most recent samples and 10 ‘future’ samples (not yet processed). This often allows one to obtain more precise estimates of the pitch period – unless the ‘future’ samples are severely distorted.

3.4. Estimation of pitch coefficient

For each of the four estimates of the pitch period, the corresponding estimates of the β coefficient are obtained using the exponentially weighted least squares approach

$$\begin{aligned} \hat{\beta}_j(t) &= \arg \min_{\beta} \sum_{i=0}^{t-1} \lambda^i \left\{ \hat{x}(t-i) - \beta \hat{x}[t-i - \hat{T}_j(t)] \right\}^2 \\ &= \frac{p_j(t)}{r_j(t)}, \quad j = 1, \dots, 4 \end{aligned} \quad (11)$$

where λ , $0 < \lambda < 1$, denotes forgetting constant,

$$\hat{x}(t) = y(t) - \sum_{i=1}^r \hat{\alpha}_i(t)y(t-i)$$

and $p_j(t)$, $r_j(t)$ are recursively computable quantities:

$$\begin{aligned} p_j(t) &= \lambda p_j(t-1) + \hat{x}(t)\hat{x}[t - \hat{T}_j(t)] \\ r_j(t) &= \lambda r_j(t-1) + \hat{x}^2[t - \hat{T}_j(t)] \end{aligned}$$

According to [9], the pitch filter (8) is stable if $|\beta(t)| < 1$. If the condition $|\hat{\beta}_j(t)| < 1$, which is checked each time the pitch coefficient is updated, is not fulfilled, the last accepted coefficient is used instead.

3.5. Detection and interpolation

To increase robustness of the outlier detection scheme, and decrease the number of false alarms, the final decisions are worked out in a collaborative way. Five outlier detectors are operated simultaneously. Four detectors are based on SAR models corresponding to the same formant coefficients $\hat{\alpha}_1(t), \dots, \hat{\alpha}_r(t)$ and different pitch parameters $\hat{T}_j(t), \hat{\beta}_j(t)$, $j = 1, \dots, 4$. The fifth detector is based on the classical AR model which incorporates only the formant coefficients.

The detection alarm is switched on if *all* five detectors indicate the presence of the outlier. The detection alarm is switched off when at least one of the detectors ‘accepts’ r consecutive signal samples.

Detection is followed by the least squares signal interpolation based on parameters of the model responsible for termination of the detection alarm.

4. EXPERIMENTAL RESULTS

The quality of the outlier detection/elimination system was checked on real speech recordings using both objective and subjective performance measures.

First, a special data base was created, made up of 10 uncorrupted speech fragments (5 male voices and 5 female voices), each covering 2000 samples (the 22.05 kHz sampling rate was used), embedded in longer speech recordings. All recordings were appropriately scaled so as to equalize the mean square signal value in all 10 test areas.

As a test material we used fragments which, when processed by the detection/interpolation algorithm based on the AR model with a carefully selected order ($p = 10$), led to audible speech distortions. In each case the entire recording was processed, but the detector was active only in the test area. Even though, ideally, no detection alarm should be triggered (because the preprocessed signals were ‘clean’), the AR-model-based detector raised many false alarms, followed by poor signal interpolations. In each case this resulted in audible signal distortions. The same test fragments were next processed using the proposed algorithm based on the SAR model of order 5. False alarms were scarce and, even when they occurred, their effects were usually impossible to perceive during listening tests.

The experiment described above was then repeated after adding to each test fragment the artificially generated noise

pulses (each sample was corrupted, with probability 0.01, by adding a random number generated by a zero-mean Gaussian source with variance 0.5). 10 different realization of the noise sequence (containing a total number of 199 of noise pulses) were used, the same for all recordings. The results of both experiments are summarized in Table 1. As the performance indicators we have used: the total number and percentage of correctly detected noise pulses, the total number and the average length of false alarms (detection of nonexistent pulses), the mean squared signal interpolation error (MSE) and the mean opinion score (MOS). The mean opinion score was based on listening tests. Each of 20 test persons knew the localization of the test fragment, as it was signalled during audition. The listening order was chosen randomly for each test recording. The scores ranged from 1 (audible disturbances or distortions, highly irritable) to 5 (no audible disturbances or distortions). All results were obtained for the same experimental settings: $\mu = 3.5$, $\gamma = \lambda = 0.99$, $\lambda_0 = 0.996$, $k_{\max} = 50$, $T_{\min} = 50$, $T_{\max} = 500$ and $m = 50$.

Fig. 1 presents typical experimental results. Fig. 1D shows the detection signal $d(t)$ generated by the algorithm based on the 10-th order AR model. Even though this algorithm is capable of localizing correctly quite a number of noise pulses (it works very well on music signals), most of the time it fails to distinguish between pitch pulses and noise pulses. Since interpolations that follow are of poor quality, the output signal (Fig. 1E) suffers from audible distortions. In contrast with this, the algorithm based on a sparse AR model, with 5 coefficients in the formant filter and 1 coefficient in the pitch filter, copes favorably with all noise pulses (including those localized in regions of glottal activity), and does not rise false alarms – see Figs. 1F and 1G.

Both the objective and subjective performance measures confirm very good properties of the method based on SAR modeling: it generates much smaller number of relatively short false alarms, yields smaller interpolation errors, and – perhaps most importantly from the practical viewpoint – earns much better opinion scores during audition tests. The same conclusion was drawn from the results of a series of tests on real (gramophone) archive speech recordings (because of the lack of space not reported here).

5. CONCLUSION

The problem of eliminating impulsive disturbances from archive speech signals using sparse AR modeling was considered. Even though such models incorporate a small number of coefficients, when carefully designed they have very good predictive capabilities. To meet stability requirements, sparse AR models are constructed in a factorized form, as a cascade of a formant filter and a pitch filter. Experimental results confirm good detection and interpolation properties of the proposed approach, impossible to achieve when the classical autoregressive approach is used.

Table 1: Detection statistics for the two compared models: AR and SAR. The remaining symbols represent: FA - the total number of false detection alarms, FL - the average length of false detection alarms, MSE - Mean Square Error, MOS - Mean Opinion Score, CD - the total number of correctly detected noise pulses, CD[%] - percentage of correctly detected noise pulses.

No.	Model	Fragments without noise pulses				Fragments with noise pulses					
		FA	FL	MSE	MOS	CD	CD[%]	FA	FL	MSE	MOS
1	AR	15	16,1	1,4E-03	4,65	181	91,0%	104	11	1,4E-02	3,75
	SAR	4	6,5	8,7E-05	5,00	176	88,4%	52	2,1	1,5E-03	4,92
2	AR	13	31,5	2,0E-03	4,70	179	89,9%	99	14,6	2,0E-02	3,42
	SAR	10	3,7	2,9E-05	4,95	178	89,4%	86	1,9	4,2E-04	4,67
3	AR	19	12,5	1,1E-03	3,15	177	88,9%	144	8,4	1,1E-02	2,64
	SAR	0	-	0,0E+00	4,85	176	88,4%	6	1,1	2,1E-05	4,71
4	AR	20	16,8	1,1E-03	3,35	183	92,0%	138	12,7	1,2E-02	2,93
	SAR	12	3,4	1,9E-05	4,90	182	91,5%	119	2,2	3,1E-04	4,77
5	AR	15	11,2	4,4E-04	3,60	172	86,4%	136	7,5	6,1E-03	2,73
	SAR	1	8,0	1,6E-06	4,90	180	90,5%	21	1,6	2,2E-05	4,86
6	AR	19	19,0	1,8E-03	3,10	169	84,9%	148	12,1	1,5E-02	2,51
	SAR	7	2,9	4,2E-06	4,95	176	88,4%	70	1,9	8,1E-05	4,67
7	AR	13	20,2	9,1E-04	3,85	177	88,9%	99	9,9	9,4E-03	2,62
	SAR	1	14,0	1,0E-06	4,95	185	93,0%	17	2,1	1,2E-05	4,84
8	AR	12	22,2	9,5E-04	3,20	183	92,0%	88	10,6	8,3E-03	2,92
	SAR	1	1,0	7,3E-09	4,90	186	93,5%	13	1,4	4,9E-06	4,44
9	AR	19	13,8	2,4E-04	4,15	185	93,0%	143	11,1	4,7E-03	3,09
	SAR	1	7,0	3,6E-07	5,00	188	94,5%	14	1,3	3,6E-06	4,93
10	AR	10	23,2	9,4E-04	3,60	183	92,0%	83	11,7	1,0E-02	2,93
	SAR	0	-	0,0E+00	5,00	185	93,0%	6	1,1	2,2E-06	4,58

6. REFERENCES

- [1] S.V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*, Wiley, 2008.
- [2] J.S. Godsill, and J.P.W. Rayner, *Digital Audio Restoration*, Springer-Verlag, 1998.
- [3] M. Niedźwiecki, and K. Cisowski, "Adaptive scheme for elimination of broadband noise and impulsive disturbances from AR and ARMA signals," *IEEE Transactions on Signal Processing*, vol. 44, pp. 528–537, 1996.
- [4] M. Niedźwiecki, "Elimination of clicks from archive speech signals," 6th IEEE International Workshop on Intelligent Signal Processing and Communication Systems, Melbourne, Australia, pp. 1116–1120, 1998.
- [5] M. Niedźwiecki, *Identification of Time-varying Processes*, Wiley, 2001.
- [6] P. Stoica, "Multistep prediction of autoregressive signals," *Electronics Letters*, vol. 29, pp. 554–555, 1993.
- [7] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, 1979.
- [8] P.R. Ramachandran, and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 467–478, 1989.
- [9] P.R. Ramachandran, and P. Kabal, "Stability and performance analysis of pitch filters in speech coders," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, pp. 937–946, 1987.

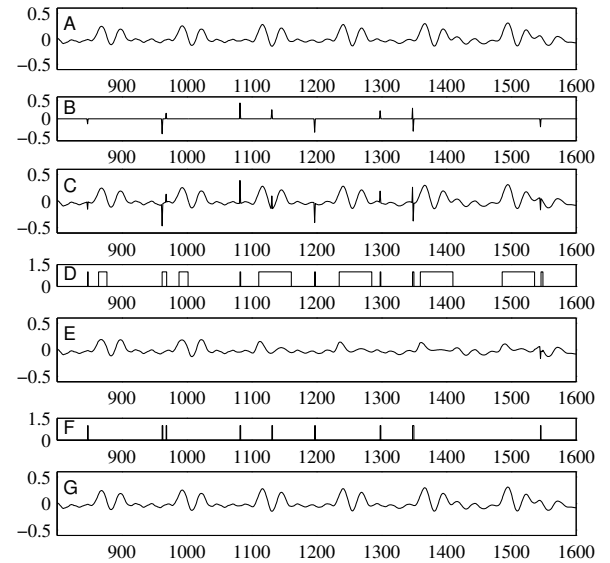


Fig. 1: Fragment of the original (noiseless) speech signal (A) and a sequence of randomly generated noise pulses (B). The remaining plots show: the speech signal corrupted with noise pulses (C), detection (D) and interpolation (E) results yielded by the AR model of order 10, detection (F) and interpolation (G) results yielded by the SAR model of order 5.