Speech Decomposition Based on a Hybrid Speech Model and Optimal Segmenation

August 31, 2021

Alfredo Esquivel Jaramillo, Jesper K.Nielsen, Mads G.Christensen
aeja@create.aau.dk, alfredoaudioaau@gmail.com

Ph.D. candidate
Audio Analysis Lab, CREATE
Aalborg University
Denmark

AALBORG UNIVERSITY
DENMARK

Introduction

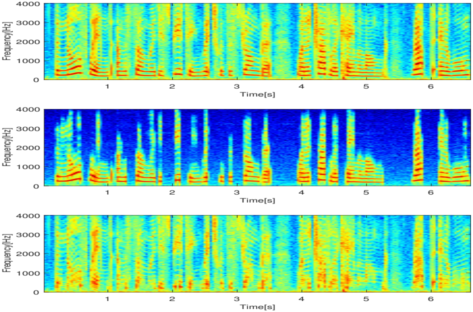
- Hybrid speech model: deterministic component+ stochastic component.
- Deterministic part: voiced speech $\leftarrow$ Harmonic model (sinusoids with frequency $kf_0, k = 1, ...L$), f_0 is the pitch or fundamental frequency and L the number of harmonics (model order).
- Stochastic part: unvoiced speech $\leftarrow$ AR process (turbulences, friction).
- Extracting both parts useful $\leftarrow$ coding, synthesis, diagnosis of illnesses.
- State-of-the-art methods do not distinguish between unvoiced speech and additive noise. This may be relevant for telemedicine applications.
- Although [[1]] estimates the pitch by whitening the periodogram, it does not use adaptive windows and does not estimate the number L of harmonics. [[2]] is based on the cepstrum, does not estimate L, and converges to the wrong solution.

---

[1] Elie, B., Chardon, G. (2016) Robust tonal and noise separation in presence of colored noise, and application to voiced fricatives, International Congress on Acoustics

[2] Yegnanarayana, B., D'Alessandro, C., Darsinos, V. (1998) An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. IEEE Transactions on Audio and Speech Processing.

Example and Motivation



- Time-varying segment lengths N have a better fit to specified models (vs. fixed lengths)
- If f_0 remains constant, N should be longer than when f_0 varies too fast.
- Parameters of deterministic and stochastic parts jointly estimated [3] for all candidate N and candidate models $\leftarrow$ to find N_{opt}
- A-priori information of AR parameters of unvoiced speech and noise.

[3] Jaramillo, A.E., Nielsen, J.K., Christensen, M.G. (2020) Robust Fundamental Frequency Estimation in Coloured Noise. ICASSP.

Signal Model and Filtering for Decomposition

- In the hybrid speech model $s(n) = v(n) + u(n)$, $v(n)$ described by the harmonic model and $u(n)$ is an autoregressive (AR) process.
- In additive noise $y(n) = v(n) + u(n) + c(n) = v(n) + x(n)$, the goal is to extract $v(n)$ and $u(n)$.
- For M ($< N$) samples, $\mathbf{y} = \mathbf{v} + \mathbf{u} + \mathbf{c}$, $\mathbf{R_y} = E\left[\mathbf{y}\mathbf{y}^T\right] = \mathbf{R_v} + \mathbf{R_u} + \mathbf{R_c}$, and since $\mathbf{R_x} = \mathbf{R_u} + \mathbf{R_c}$, $\mathbf{R_y} = \mathbf{R_v} + \mathbf{R_x}$, where $\Phi_x(\omega) = \Phi_u(\omega) + \Phi_c(\omega)$.
- First, extract $v(n)$, $\hat{\mathbf{v}} = \mathbf{H}_v\mathbf{y} = \mathbf{H}_v\mathbf{v} + \mathbf{H}_v\mathbf{x}$. From the joint diagonalization of $\mathbf{R_v}$ (parametrized by f_0) and $\mathbf{R_x}$ (in terms of AR parameters), use M eigenvectors and eigenvalues to form a Wiener filter matrix

$$\mathbf{H}_v = \mathbf{R}_v \sum_{q=1}^{M} \frac{\mathbf{b}_q \mathbf{b}_q^H}{1 + \lambda_q}. \tag{1}$$

- To extract $u(n)$, $U(\omega) = H_u(\omega)\hat{X}(\omega)$, where $\hat{X}(\omega)$ is the spectrum of the modelled residual and $H_u(\omega) = \frac{\hat{\Phi}_u(\omega)}{\hat{\Phi}_u(\omega)+\hat{\Phi}_c(\omega)}$ (using prior information).

Statistics and Parameter Estimation

- $v(n) = \sum_{l=1}^{L} \left[\alpha_l e^{j2\pi f_0 ln} + \alpha_l^* e^{-j2\pi f_0 ln}\right]$, more accurate for a length N_{opt}.
- L of harmonics, $\alpha_l = \frac{A_l}{2} e^{j\psi l}$ is the complex amplitude of the l'th harmonic with $A_l > 0$ the real amplitude, ψ_l the initial phase.

Statistics and Parameter Estimation

- $v(n) = \sum_{l=1}^{L} \left[\alpha_l e^{j2\pi f_0 ln} + \alpha_l^* e^{-j2\pi f_0 ln}\right]$, more accurate for a length N_{opt}.
- L of harmonics, $\alpha_l = \frac{A_l}{2} e^{j\psi l}$ is the complex amplitude of the l'th harmonic with $A_l > 0$ the real amplitude, ψ_l the initial phase.
- M samples $\mathbf{v} = \mathbf{Z}(f_0)\boldsymbol{\alpha}$, $\mathbf{Z}(f_0) = [\mathbf{z}(f_0)\ \mathbf{z}^*(f_0) \cdots \mathbf{z}^*(Lf_0)],$ where $\mathbf{z}(lf_0) = [1\ e^{jl2\pi f_0} \ \cdots \ e^{jl2\pi f_0(M-1)}]^T$, and $\boldsymbol{\alpha} = \frac{1}{2}[A_1 e^{j\psi_1}\ A_1 e^{-j\psi_1} \ \cdots \ A_L e^{-j\psi_L}]^T$ are the harmonics amplitudes.

Statistics and Parameter Estimation

- $v(n) = \sum_{l=1}^{L} \left[\alpha_l e^{j2\pi f_0 ln} + \alpha_l^* e^{-j2\pi f_0 ln}\right]$, more accurate for a length N_{opt}.
- L of harmonics, $\alpha_l = \frac{A_l}{2} e^{j\psi l}$ is the complex amplitude of the l'th harmonic with $A_l > 0$ the real amplitude, ψ_l the initial phase.
- M samples $\mathbf{v} = \mathbf{Z}(f_0)\boldsymbol{\alpha}$, $\mathbf{Z}(f_0) = [\mathbf{z}(f_0)\ \mathbf{z}^*(f_0) \cdots \mathbf{z}^*(Lf_0)]$, where $\mathbf{z}(lf_0) = [1\ e^{jl2\pi f_0} \cdots e^{jl2\pi f_0(M-1)}]^T$, and $\boldsymbol{\alpha} = \frac{1}{2}[A_1 e^{j\psi_1}\ A_1 e^{-j\psi_1} \cdots A_L e^{-j\psi_L}]^T$ are the harmonics amplitudes.
- $u(n) = -\sum_{i=1}^{P} \beta_{u_i} u(n-i) + e(n)$, where $\{\beta_{u_i}\}_{i=1}^{P}$ are the P AR coefficients and $e(n)$ is the excitation WGN process with variance σ_e^2. Also $c(n)$ modelled as an AR process with $\{\gamma_{c_i}\}_{i=1}^{P}$.

Statistics and Parameter Estimation

- $v(n) = \sum_{l=1}^{L} \left[\alpha_l e^{j2\pi f_0 ln} + \alpha_l^* e^{-j2\pi f_0 ln}\right]$, more accurate for a length N_{opt}.
- L of harmonics, $\alpha_l = \frac{A_l}{2} e^{j\psi l}$ is the complex amplitude of the l'th harmonic with $A_l > 0$ the real amplitude, ψ_l the initial phase.
- M samples $\mathbf{v} = \mathbf{Z}(f_0)\boldsymbol{\alpha}$, $\mathbf{Z}(f_0) = [\mathbf{z}(f_0)\ \mathbf{z}^*(f_0)\ \cdots\ \mathbf{z}^*(Lf_0)],$ where $\mathbf{z}(lf_0) = [1\ e^{jl2\pi f_0}\ \cdots\ e^{jl2\pi f_0(M-1)}]^T$, and $\boldsymbol{\alpha} = \frac{1}{2}[A_1 e^{j\psi_1}\ \ A_1 e^{-j\psi_1}\ \cdots\ A_L e^{-j\psi_L}]^T$ are the harmonics amplitudes.
- $u(n) = -\sum_{i=1}^{P} \beta_{u_i} u(n-i) + e(n)$, where $\{\beta_{u_i}\}_{i=1}^{P}$ are the P AR coefficients and $e(n)$ is the excitation WGN process with variance σ_e^2. Also $c(n)$ modelled as an AR process with $\{\gamma_{c_i}\}_{i=1}^{P}$.
- $\hat{\mathbf{R}}_{\mathbf{v}} = \mathbf{Z}(\hat{f}_0)\hat{\mathbf{P}}\mathbf{Z}(\hat{f}_0)^H$, where $\hat{\mathbf{P}} = E\{\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}^H\} = \frac{1}{4}\,\mathrm{diag}([\hat{A}_1^2\ \hat{A}_1^2 \cdots \hat{A}_L^2\ \hat{A}_L^2])$.
- Estimates of f_0 and the amplitudes are required from the segment of length N_{opt}, before $\rightarrow$ estimate parameters for all N candidates $\rightarrow$ N_{opt} is the one which maximises the posterior probability of the data.

Joint parameter estimation (f_0, L and AR parameters)

$y(n)$ → Pre-whitener (Pre-processor) → WGN-based method (NLS f_0 estimator) → Post-processing (Prewh.+WGN estim.)

- Which pre-whitener gives more reliability to the NLS f_0 estimator? Fit noise PSD to AR spectrum (MMSE-SPP), OK for $c(n)$ stationary → Par-NMF (pre-trained spectral shapes), better for $c(n)$ non-stationary[4]

---

[4] Jaramillo, A.E., et.al (2021) An Adaptive Autoregressive Pre-whitener for Speech and Acoustic Signals Based on Parametric NMF. Submitted to Applied Acoustics

[5] Srinivasan et al. (2005) Codebook driven short-term predictor parameter estimation for speech enhancement. IEEE Transactions on Audio, Speech and Language Processing.

Joint parameter estimation (f_0, L and AR parameters)

$y(n)$ → Pre-whitener (Pre-processor) → WGN-based method (NLS f_0 estimator) → Post-processing (Prewh.+WGN estim.)

- Which pre-whitener gives more reliability to the NLS f_0 estimator? Fit noise PSD to AR spectrum (MMSE-SPP), OK for $c(n)$ stationary → Par-NMF (pre-trained spectral shapes), better for $c(n)$ non-stationary[4]
- Post-processing consists in iterating between (until convergence):
  1. $\hat{f}_0 = \arg\max_{f_0} \underline{\mathbf{y}}_{\mathrm{W}}^T \mathbf{Z}(f_0) \left[\mathbf{Z}^H(f_0)\mathbf{Z}(f_0)\right]^{-1} \mathbf{Z}^H(f_0)\underline{\mathbf{y}}_{\mathrm{W}}$, a final L selected using model comparison (e.g., BIC-Bayesian Information Criterion).
  2. After $\hat{\boldsymbol{\alpha}} = [\mathbf{Z}^H(\hat{f}_0)\mathbf{Z}(\hat{f}_0)]^{-1}\mathbf{Z}(\hat{f}_0)^H\mathbf{y}$, estimate directly the residual $\underline{\mathbf{x}} = \underline{\mathbf{y}} - \mathbf{Z}(\hat{f}_0)\hat{\boldsymbol{\alpha}}$ (and $\hat{\mathbf{R}}_x$), and its AR parameters. Pre-whiten again $y(n)$.

[4] Jaramillo, A.E., et.al (2021) An Adaptive Autoregressive Pre-whitener for Speech and Acoustic Signals Based on Parametric NMF. Submitted to Applied Acoustics

[5] Srinivasan et al. (2005) Codebook driven short-term predictor parameter estimation for speech enhancement. IEEE Transactions on Audio, Speech and Language Processing.

Joint parameter estimation (f_0, L and AR parameters)

$y(n)$ → Pre-whitener (Pre-processor) → WGN-based method (NLS f_0 estimator) → Post-processing (Prewh.+WGN estim.)

- Which pre-whitener gives more reliability to the NLS f_0 estimator? Fit noise PSD to AR spectrum (MMSE-SPP), OK for $c(n)$ stationary $\rightarrow$ Par-NMF (pre-trained spectral shapes), better for $c(n)$ non-stationary[4]
- Post-processing consists in iterating between (until convergence):
  1. $\hat{f}_0 = \arg\max_{f_0} \underline{\mathbf{y}}_{\mathrm{W}}^T \mathbf{Z}(f_0)\left[\mathbf{Z}^H(f_0)\mathbf{Z}(f_0)\right]^{-1}\mathbf{Z}^H(f_0)\underline{\mathbf{y}}_{\mathrm{W}}$, a final L selected using model comparison (e.g., BIC-Bayesian Information Criterion).
  2. After $\hat{\boldsymbol{\alpha}} = [\mathbf{Z}^H(\hat{f}_0)\mathbf{Z}(\hat{f}_0)]^{-1}\mathbf{Z}(\hat{f}_0)^H\mathbf{y}$, estimate directly the residual $\underline{\mathbf{x}} = \underline{\mathbf{y}} - \mathbf{Z}(\hat{f}_0)\hat{\boldsymbol{\alpha}}$ (and $\hat{\mathbf{R}}_x$), and its AR parameters. Pre-whiten again $y(n)$.
- The modelled $\underline{\mathbf{x}} = \underline{\mathbf{y}} - \mathbf{Z}(\hat{f}_0)\hat{\boldsymbol{\alpha}}$ has a parametric spectrum $\hat{\Phi}_x(\omega) = \frac{\sigma_u^2}{\left|1+\sum_{i=1}^{P}\beta_{u_i}e^{-j\omega i}\right|^2} + \frac{\sigma_c^2}{\left|1+\sum_{i=1}^{P}\gamma_{c_i}e^{-j\omega i}\right|^2}$. Parameters estimated from single codebook entries which minimize d_{IS}. [5] Wiener filter is applied.

[4] Jaramillo, A.E., et.al (2021) An Adaptive Autoregressive Pre-whitener for Speech and Acoustic Signals Based on Parametric NMF. Submitted to Applied Acoustics

[5] Srinivasan et al. (2005) Codebook driven short-term predictor parameter estimation for speech enhancement. IEEE Transactions on Audio, Speech and Language Processing.

Criteria for optimal segmentation

- Each way in which a signal can be segmented (segment of kN_{Min} samples) is a model. The set of candidate models is M.
- For $v(n)$, use MAP to select the model which maximises the a posteriori probability: either $J_1(N) = \frac{N}{2}\ln\frac{1}{N}||\underline{\mathbf{y}}_{\text{W}} - \mathbf{Z}\boldsymbol{\alpha}_{\text{W}}||_2^2 + \frac{3}{2}\ln N + \hat{L}(N)\ln N$ or $J(N) = \frac{N}{2}\ln||\underline{\mathbf{y}}_{\text{W}}||_2^2$. (Obtain segmentation markers of $v(n)$ and N_{opt}).

Criteria for optimal segmentation

- Each way in which a signal can be segmented (segment of kN_{Min} samples) is a model. The set of candidate models is M.
- For $v(n)$, use MAP to select the model which maximises the a posteriori probability: either $J_1(N) = \frac{N}{2}\ln\frac{1}{N}||\underline{\mathbf{y}}_{\text{W}} - \mathbf{Z}\boldsymbol{\alpha}_{\text{W}}||_2^2 + \frac{3}{2}\ln N + \hat{L}(N)\ln N$ or $J(N) = \frac{N}{2}\ln||\underline{\mathbf{y}}_{\text{W}}||_2^2$. (Obtain segmentation markers of $v(n)$ and N_{opt}).
- After extracting $v(n)$, the residual $x(n)$ is segmented according to the log-likelihood $J_2(N) = \frac{N}{2}d_{\text{IS}}(\hat{\Phi}_x, \frac{\sigma_u^2}{|B_u^i(\omega)|^2} + \frac{\sigma_c^2}{|\Gamma_c^j(\omega)|^2}) + \frac{1}{2}\sum_{k=1}^{N}\ln\hat{\Phi}_x$. (Obtain segmentation markers of $u(n)$ and N'_{opt}).
- The cost of all possible segment lengths is compared, $\widehat{M} = \arg\min_M J_i$, $i \in \{1,2\}$. A minimal segment length, N_{min}, is defined, generating a subsegment of N_{min} samples and dividing the signal into S subsegments.

Segmentation algorithm [a]

[a]Prandoni et al. (2000) R/D Optimal Linear Prediction. IEEE Transactions Audio,Speech,Language Processing

while $s \times N_{\text{min}} \leq$ **length(signal)**

  Initialize $B = \min([s, B_{\text{max}}])$

  for $b = 1 : B$ **do**

    subsegment of signal to use is $s - b + 1, ..., s$

    For $v(n)$: estimate f_0, L and $\boldsymbol{\alpha}$, find $\mathbf{Z}$ if $\hat{L} \neq 0$

    calculate $J_{(s-b+1)m}$

$$J(b) = \begin{cases} J_{(s-b+1)s} + J_{(s-b),\text{opt}} & \text{if } s - b > 0, \\ J_{(s-b+1)s} & \text{otherwise.} \end{cases}$$

  end for

  $b_{\text{opt}}(s) = \arg\min_b J(b)$

  $J_{s,\text{opt}} = \min_b J(b)$

  $s = s + 1$

end while

$s = S$

while $s > 0$

  number of subsegments in segment is $b_{\text{opt}}(s)$

  $s = s - b_{\text{opt}}(s)$

end while

Summary
Steps of the decomposition

1. The noisy signal is pre-processed with an adaptive AR pre-whitener, yielding $y_{\mathrm{W}}(n)$.
2. Parameter estimates of $v(n)$ and $x(n)$ are jointly obtained for all candidate segment lengths. Followingly, based on $J_1(n)$, the markers of the optimal segmentation for voiced speech and N_{opt} are obtained.
3. Parameter estimates of $v(n)$ and $x(n)$ and statistics $\mathbf{R_v}$, $\mathbf{R_x}$ are obtained from the segments of length N_{opt}. If $\hat{L}(N_{\text{opt}}) \neq 0$, estimate $\mathbf{v}$ using $\mathbf{H}_v$ after joint diagonalization of $\mathbf{R_v}$ and $\mathbf{R_x}$.
4. Obtain the modelled residual $\underline{\mathbf{x}} = \underline{\mathbf{y}} - \mathbf{Z}(\hat{f}_0)\hat{\boldsymbol{\alpha}}$ in all the different obtained optimal lengths $\{N_{\text{opt}}\}$. Once the whole modelled $x(n)$ is obtained, estimate $u(n)$ parameters $\{\sigma_u^2, \{\beta_{u_i}\}_{i=1}^{P}\}$ for all candidate segment lengths.
5. Based on $J_2(n)$, obtain the markers of the optimal segmentation for $u(n)$ and N'_{opt}.
6. The $u(n)$ parameters $\{\sigma_u^2, \{\beta_{u_i}\}_{i=1}^{P}\}$ are obtained from the segments of length N'_{opt}. Extract $\underline{\underline{\mathbf{u}}}$ using Wiener filter in the frequency domain.

Experimental Evaluation
Optimal vs. fixed segmentation

- Segments of length $N = 160$ to $N = 400$ (i.e., 20-50ms) in steps of 40.
- $B_{\text{max}} = 10$. The cost for $b = 1, 2, 3$ (i.e., 5,10 and 15 ms) is set to ∞ (NLS f_0 estimator does not work well for low f_0 at too low N).
- $M = 40$, and filtering $\mathbf{H}_v$ updated every 20 samples (i.e., 50 % overlap).
- Ground truth from clean $s(n)$: $u(n) = s(n) - v(n)$, from which we obtain an AR codebook of 64 entries ($N = 160$ for the training)

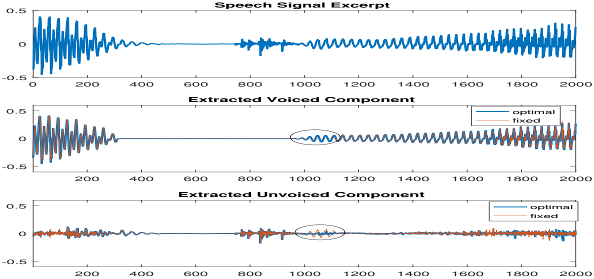


Figure: Extraction of voiced and unvoiced components from optimal and fixed segmentation on a clean signal excerpt.

Experimental Results

Performance under 4 noise types (babble, restaurant, factory, street)

- To find the segmentation markers, the signal is pre-whitened with the setup of [6], using 32 speech and 256 noise pre-trained spectral shapes.
- The segmentation of $u(n)$ is obtained from segments of 15 to 40 ms. A codebook of 16 noise entries was used for the Wiener filter.
- Compared to state-of-the-art methods when their input signal is enhanced using OM-LSA (since they do not take into account $c(n)$).

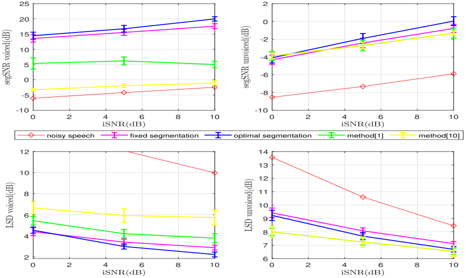


Figure: Averaged LSD and segmental SNR (segSNR) in different iSNRs.

[6] Jaramillo, A.E., et.al (2021) An Adaptive Autoregressive Pre-whitener for Speech and Acoustic Signals Based on Parametric NMF. Submitted to Applied Acoustics

Conclusion

- The use of an optimal segmentation combined with parameter estimates of a hybrid speech model allow to have a more accurate recovery of $v(n)$ and $u(n)$, compared to the use of fixed segments.
- An adaptive segmentation results in a better modelling of the periodic parts in $v(n)$ with a higher probability of improved segSNR and also of a lower LSD of both extracted $v(n)$ and $u(n)$.

[7] Rindom, J., et al., (2015) Noise reduction with optimal variable span linear filters. IEEE Transactions Audio, Speech and Language Processing

[8] Quinn, B.,et al. (2021) Fast algorithms for fundamental frequency estimation in autoregressive noise. Signal Processing.

Conclusion

- The use of an optimal segmentation combined with parameter estimates of a hybrid speech model allow to have a more accurate recovery of $v(n)$ and $u(n)$, compared to the use of fixed segments.
- An adaptive segmentation results in a better modelling of the periodic parts in $v(n)$ with a higher probability of improved segSNR and also of a lower LSD of both extracted $v(n)$ and $u(n)$.
- We considered prior spectral information stored in codebooks in order to differentiate between $u(n)$ and $c(n)$.
- A higher segSNR and lower LSD for $v(n)$ is possible when compared to reference methods, with a potential to reduce the LSD for the extracted unvoiced part (e.g., other variable span linear filters [[7]] or including masking curves/perceptual criteria).

---

[7] Rindom, J., et al., (2015) Noise reduction with optimal variable span linear filters. IEEE Transactions Audio, Speech and Language Processing

[8] Quinn, B.,et al. (2021) Fast algorithms for fundamental frequency estimation in autoregressive noise. Signal Processing.

Conclusion

- The use of an optimal segmentation combined with parameter estimates of a hybrid speech model allow to have a more accurate recovery of $v(n)$ and $u(n)$, compared to the use of fixed segments.
- An adaptive segmentation results in a better modelling of the periodic parts in $v(n)$ with a higher probability of improved segSNR and also of a lower LSD of both extracted $v(n)$ and $u(n)$.
- We considered prior spectral information stored in codebooks in order to differentiate between $u(n)$ and $c(n)$.
- A higher segSNR and lower LSD for $v(n)$ is possible when compared to reference methods, with a potential to reduce the LSD for the extracted unvoiced part (e.g., other variable span linear filters [[7]] or including masking curves/perceptual criteria).
- Future work → derive the segmentation based on the recently introduced joint f_0-AR estimator. [[8]], and evaluating the methodology for applications such as diagnosis of Parkinson's disease.

[7] Rindom, J., et al., (2015) Noise reduction with optimal variable span linear filters. IEEE Transactions Audio, Speech and Language Processing

[8] Quinn, B.,et al. (2021) Fast algorithms for fundamental frequency estimation in autoregressive noise. Signal Processing.

Thanks for your attention!

AALBORG UNIVERSITY
DENMARK