# Audio intercept analysis: Can we identify and locate the speaker?

K V Vijay Girish, PhD Student, *kv@ee.iisc.ernet.in*

Research Advisor: Prof. A G Ramakrishnan, Research Collaborator: Dr. T V Ananthapadmanabha

Medical Intelligence & Language Engineering (MILE) Lab, Department of Electrical Engineering,
Indian Institute of Science, Bangalore

## Motivation

- Any audio intercept is a mixture of sounds including environmental sound in the background and speech at specific intervals.
- Analyzing the intercepts of conversations is of importance to forensics and other investigations
- Identifying the probable geographical location and the speaker is useful for tracking the suspect

## Dictionary and Sparse representation

- A dictionary is a matrix $D \in \mathbb{R}^{p\times K}$ (with $p$ as the dimension of the acoustic feature vector) containing $K$ column vectors called atoms, denoted as $\mathbf{d_n}, 1 \leq n \leq K$
- Given a feature vector $y$, it can be approximated by $\mathbf{y} \approx \hat{\mathbf{y}} = D\mathbf{x}$ where $D$ is known and
- The weight vector $\mathbf{x} \in \mathbb{R}^K$ is estimated by $\mathbf{x} = \underset{\mathbf{x}}{arg\,min}\ distance(\mathbf{y}, D\mathbf{x})$ such that $||\mathbf{x}||_0 \leq l$ and $l$ is the sparsity constraint
- A weight vector $\mathbf{x}$ viewed as a concatenation of block vectors $\mathbf{x_i}$ is k-block sparse if $||\mathbf{x_m}||_2$ is non-zero for $m$ taking at most $k$ number of values and $||\mathbf{x_i}||_2 = 0,\ \forall i \neq m$
- Dictionary atoms corresponding to each $\mathbf{x_i}$ belong to a dictionary block, $D_i$

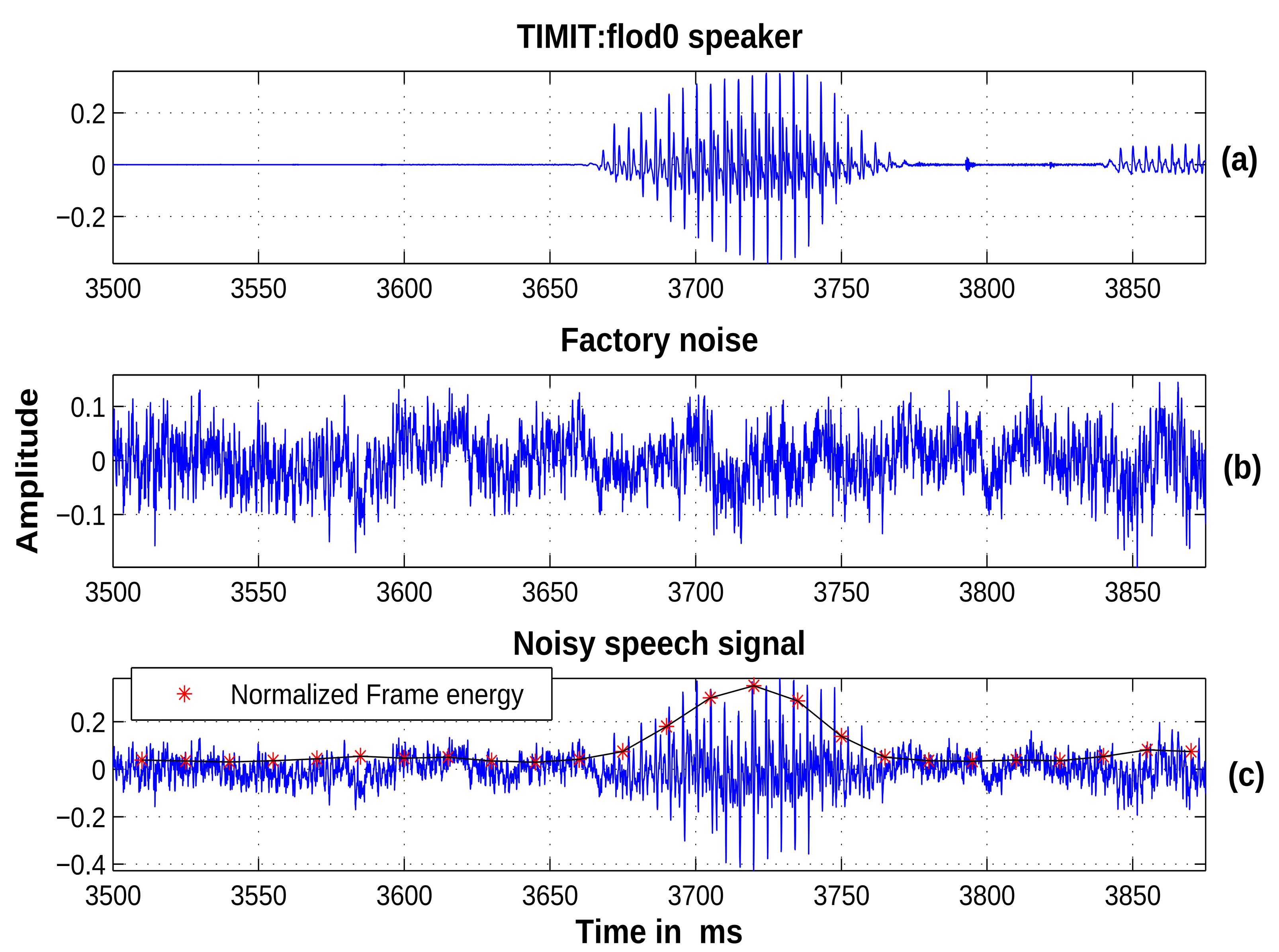
$$\hat{\mathbf{y}} = [D_1\ D_2 ... D_N] \begin{vmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \mathbf{x_N} \end{vmatrix}$$

## Problem definition

- Noisy speech signal, $s[n]$ is simulated as a linear combination of two sources, a speech, $s_{sp}[n]$ and a noise source, $s_{ns}[n]$.

$$s[n] = s_{sp}[n] + s_{ns}[n] \tag{1}$$

- The speech and noise are constrained to belong to a specific set of speakers and noise sources
- The signal is classified as belonging to one of the predefined speaker and noise sources




- A frame size of 60 ms shifted by 15 ms is considered for feature vector extraction
- Fourier transform (FT) of each frame, called as STFT (short time FT) is computed and its magnitude is used as the feature vector $\mathbf{y_i}$
- Dictionary learning involves random selection of $K = 500$ number of feature vectors

## Noise classification stage

- Given $T$ frames from $s[n]$, and the corresponding feature vectors $y_i$, $1 \leq i \leq T$, the energy of each frame is $E_y(i) = ||y_i||_2^2$
- Ten feature vectors having the lowest energy are extracted as $Y_{min} = [y_{(1)} \cdots y_{(10)}]$
- A concatenated dictionary is constructed from the individual noise source dictionaries as $D_{ns} = [D_{ns}^1 .. D_{ns}^{N_{ns}}]$
- The $j^{th}$ column of $Y_{min}$ can be represented as

$$y_{(j)} \approx [D_{ns}^1 .. D_{ns}^{N_{ns}}][x_1' ... x_{N_{ns}}']' \tag{2}$$

- The noise source is estimated as the index $\hat{m}$ which gives maximum absolute sum of correlation :

$$\hat{m} = \arg\max_i \sum_{j=1}^{10} ||(D_{ns}^i)^T y_{(j)}||_1 \tag{3}$$

## Speaker classification stage

- The test feature vectors $y_i$ from $s[n]$ (60% of the feature vectors, whose energies are higher than those of the other 40%) are represented as linear combination of the dictionary atoms from the estimated noise source, $D_{ns}^{\hat{m}}$ and concatenation of speech source dictionaries $[D_{sp}^1 .... D_{sp}^{N_{sp}}]$

$$y \approx [D_{sp}^1 .... D_{sp}^{N_{sp}} D_{ns}^{\hat{m}}][x_1' ... x_{N_{sp}}' x_{\hat{m}}']' = Dx \tag{4}$$

where $D = [D_{sp}^1 .... D_{sp}^{N_{sp}} D_{ns}^{\hat{m}}]$, $x = [x_1' ... x_{N_{sp}}' x_{\hat{m}}']'$

- The weight vector, $x$ is estimated by minimizing the distance, $distance(y, Dx)$ using Active Set Newton Algorithm ($ASNA$):

$$\underset{\mathbf{x}}{\text{minimize}}\ KL(\mathbf{y}||\hat{\mathbf{y}}),\ \hat{\mathbf{y}} = D\mathbf{x}\ s.t.\ x \geq 0 \tag{5}$$

where $KL(\mathbf{y}||\hat{\mathbf{y}}) = \mathbf{y} \log\left(\frac{\mathbf{y}}{\hat{\mathbf{y}}}\right) - \mathbf{y} + \hat{\mathbf{y}}$ is the distance measure used.

- $ASNA$ is based on iteratively updating a set of active atoms, with the weights updated using the Newton method:

$$\mathbf{x_{k+1}} = \mathbf{x_k} - \alpha_\mathbf{k} \mathbf{H_k^{-1}} \mathbf{g_k} \tag{6}$$

where $\mathbf{g_k}$ is the gradient of the minimization function ($KL(\mathbf{y}||\hat{\mathbf{y}})$) with respect to $\mathbf{x}$, $\mathbf{H_k}$ is the corresponding Hessian matrix and $\alpha_\mathbf{k}$ is the step size parameter

- A new measure *Total Sum of Weights (TSW)* is defined as the total absolute sum of elements of $x_i$, $1 \leq i \leq N_{sp}$ for all selected feature vectors $y_j$,

$$TSW_i = \sum_j ||x_i||_1, \forall y = y_j,\ 1 \leq j \leq L \tag{7}$$

- The speaker source is estimated as the index $\hat{n}$

$$\hat{n} = \arg\max_i TSW_i \tag{8}$$

## Implementation and Results

- Database used
  - Ten different noise sources taken from Noisex database: *http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex*
  - Data from ten randomly selected speakers from dialect 5 of training set of TIMIT database
- Classification accuracy of speaker and noise sources:

| SNR (dB) | -10 | 0 | 10 | 20 |
|---|---|---|---|---|
| Speaker ( %) | 37 | 83 | 99 | 100 |
| Noise (%) | 100 | 100 | 100 | 100 |

## Conclusion and Future work

- We have shown speaker and noise classification in a noisy speech signal with good classification accuracy using a simple dictionary learning method and sparse representation
- We are exploring other dictionary and discriminative learning methods

ster