Analysis of audio intercepts: Can we identify and locate the speaker?

K V Vijay Girish, PhD Student
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

kv@ee.iisc.ernet.in

April 29, 2016

Audio intercepts

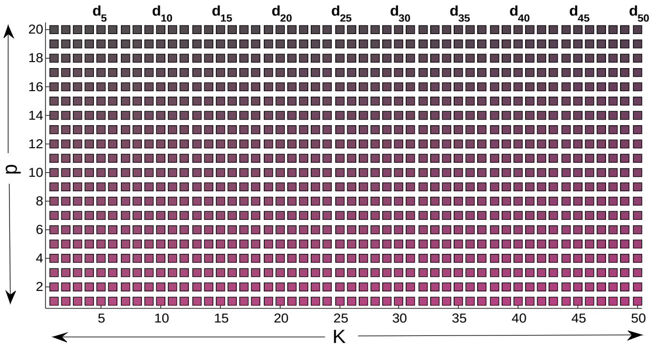
- Any audio intercept is a mixture of
  - Environmental sounds: any non-speech sounds
  - Human speech: single or multi-speaker
- Analyzing the intercepts of conversations is of importance to forensics and other investigations
- Identifying the probable geographical location and the speaker is useful for tracking the suspect
- Background environmental noise => **Probable geographical location**
- **Speaker =** Possible suspect

Note:

Every audio intercept includes environmental sound in the background and speech at specific intervals

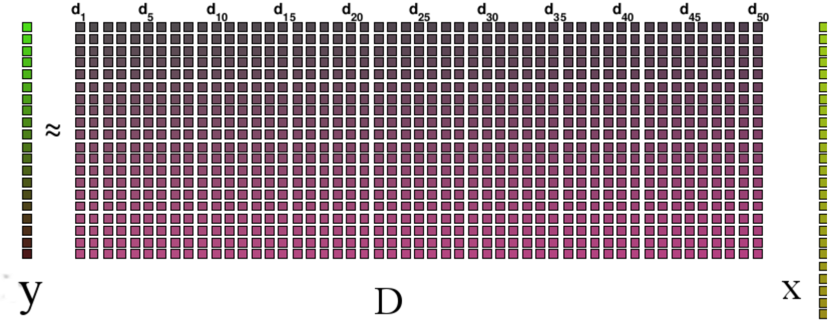
Dictionary

- 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 $\mathbf{d_n}, 1 \leq n \leq K$
- $K > p$ for an overcomplete dictionary



Visualization of a sample dictionary, D with $p = 20, \ K = 50$

Source Recovery



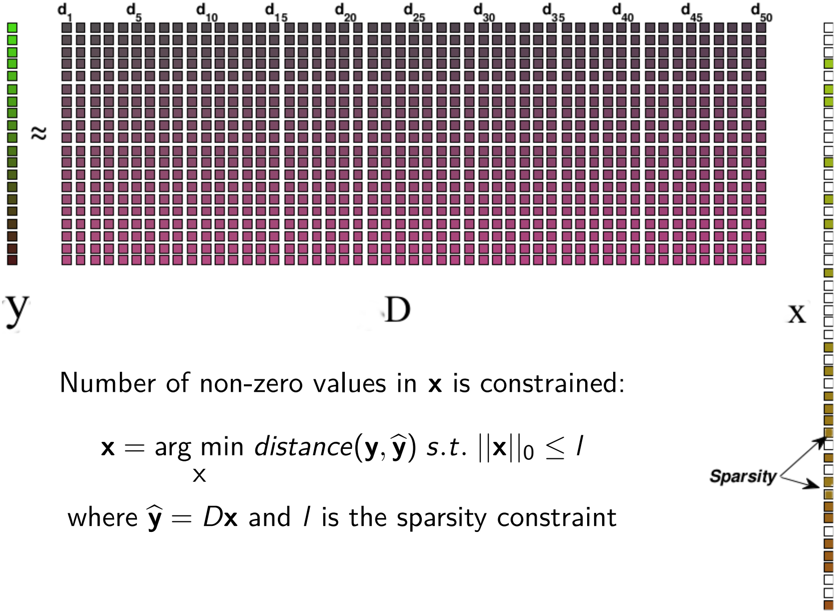
Any known real valued feature vector, $\mathbf{y}$ can be represented as

$$\mathbf{y} \approx \widehat{\mathbf{y}} = D\mathbf{x}$$

D is known and $\mathbf{x} \in \mathbb{R}^{K}$ is the weight vector

$$\mathbf{x} = \arg\min_{\mathrm{x}} \mathit{distance}(\mathbf{y}, D\mathbf{x})$$

Sparsity



Number of non-zero values in $\mathbf{x}$ is constrained:

$$\mathbf{x} = \arg\min_{\mathbf{x}} \textit{distance}(\mathbf{y}, \hat{\mathbf{y}}) \ s.t. \ ||\mathbf{x}||_0 \leq l$$

where $\hat{\mathbf{y}} = D\mathbf{x}$ and l is the sparsity constraint

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

Sample noisy speech signal

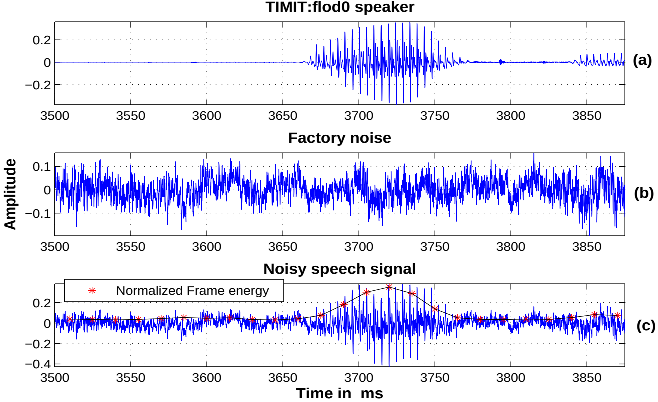
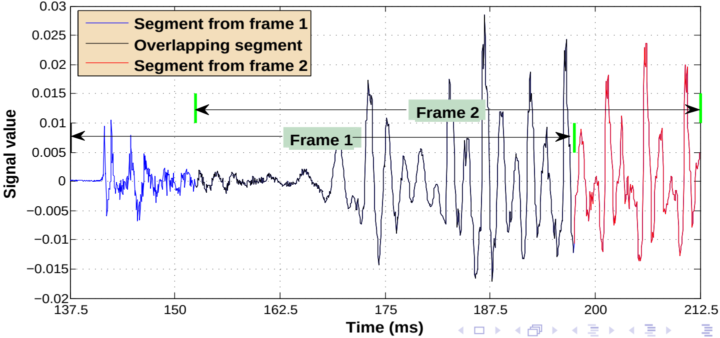


Figure: Illustration of utterance from a female speaker, factory noise and the noisy speech signal at an SNR of 0 dB. Star marks in (c) indicate framewise energies.

Feature vector extraction and dictionary learning

- 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

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

- 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 speaker source dictionaries $[D_{sp}^{1} \ldots . D_{sp}^{N_{sp}}]$

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

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

- The weight vector, x is estimated by minimizing the distance, $distance(y, Dx)$ using *ASNA* by Virtanen et. al. "Active-Set Newton Algorithm for Overcomplete Non-Negative Representations of Audio", 2013

Speaker classification stage: Measure evaluated

- 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,

$$TSW_i = \sum_{j} ||x_i||_1, \forall y = y_j, \; 1 \leq j \leq L \tag{5}$$

  Here, L is equal to the count of the 60% of the feature vectors having high energy

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

$$\hat{n} = \arg\max_i TSW_i \tag{6}$$

Implementation details

- 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
- All the audio signals are sampled at 16 kHz
- Speech segments are added to the noise test signal at randomly selected intervals ensuring a minimum of 200 ms gap between consecutive speech segments
- The test signal is simulated at a local SNR of -10, 0, 10 and 20 dB.

Confusion matrix and classification accuracy

Table: Confusion matrix showing the estimated sources for all combinations of speaker and noise sources at a SNR of 0 dB. Only misclassified speakers are shown in the table.

Noise > Speaker	*white*	*factory1*	*hf-channel*	*f16 cockpit*	*jet cockpit*
fsms1	*	*	*	*	*
flod0	fbjl0	*	*	*	*
mmab1	fbjl0	mmwb0	*	mtat0	mtat0
fbjl0	*	fsms1	fsms1	*	fsms1
mmwb0	*	*	*	*	*
mmdm1	*	*	*	*	*
mges0	fbjl0	mmwb0	*	mtat0	mmwb0
mtat0	fbjl0	*	*	*	*
ftbw0	fbjl0	*	*	*	*
mram0	fbjl0	mmwb0	mges0	*	*

Table: Classification accuracy of speaker and noise sources

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

Thank you,

K V Vijay Girish
PhD Student,
Medical Intelligence & Language Engineering (MILE) Lab,
Department of Electrical Engineering,
Indian Institute of Science, Bangalore

kv@ee.iisc.ernet.in