A Compressive Sensing Perspective of Linguistic Information Recovery

Pranay Dighe[1,2], Afsaneh Asaei[1], Hervé Bourlard[1,2]
[1] Idiap Research Institute, Martigny [2] Ecole Polytechnique Fédérale de Lausanne(EPFL)

Automatic Speech Recognition (ASR)

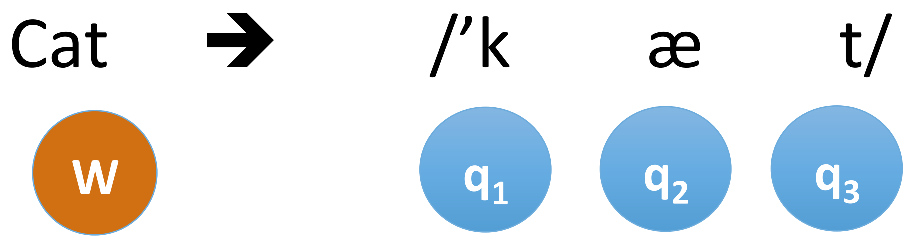
- Given acoustic observation $X = [x_1, x_2, \ldots, x_T]$, goal of ASR is to find a word sequence $\hat{W} = [w_1, w_2, \ldots, w_T]$ that has *Maximum a Posteriori* probability $P(W|X)$

x_t : *spectral features* q_k : *phones* w_ℓ : words t : *time*

Automatic Speech Recognition (ASR)

- Given acoustic observation $X = [x_1, x_2, \ldots, x_T]$, goal of ASR is to find a word sequence $\hat{W} = [w_1, w_2, \ldots, w_T]$ that has *Maximum a Posteriori* probability $P(W|X)$

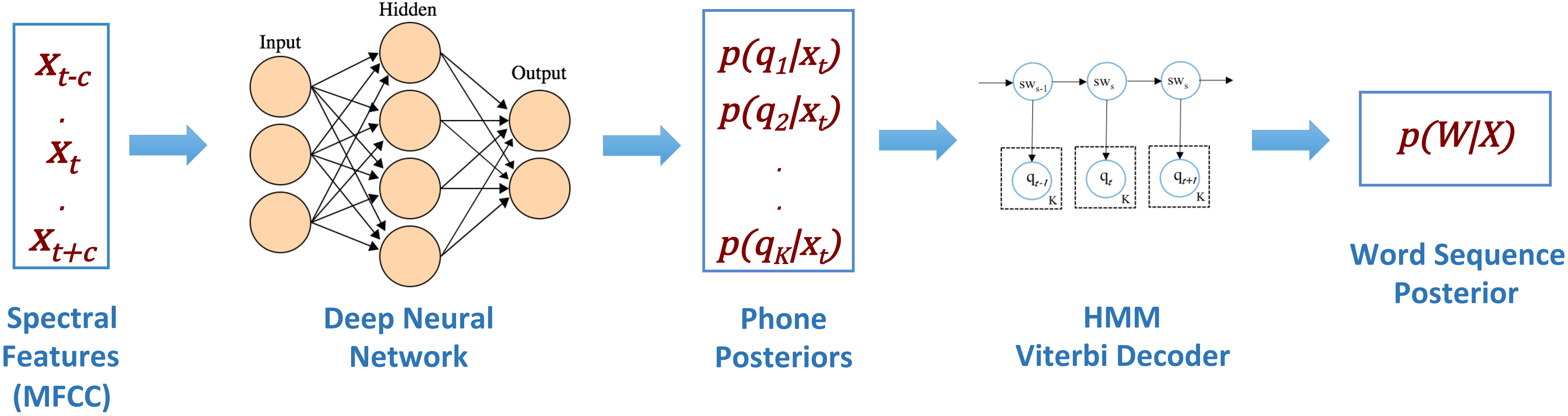
x_t : *spectral features* q_k : *phones* w_ℓ : words t : *time*



Automatic Speech Recognition (ASR)

- Given acoustic observation $X = [x_1, x_2, \ldots, x_T]$, goal of ASR is to find a word sequence $\hat{W} = [w_1, w_2, \ldots, w_T]$ that has *Maximum a Posteriori* probability $P(W|X)$

x_t : *spectral features* q_k : *phones* w_ℓ : words t : *time*



Automatic Speech Recognition (ASR)

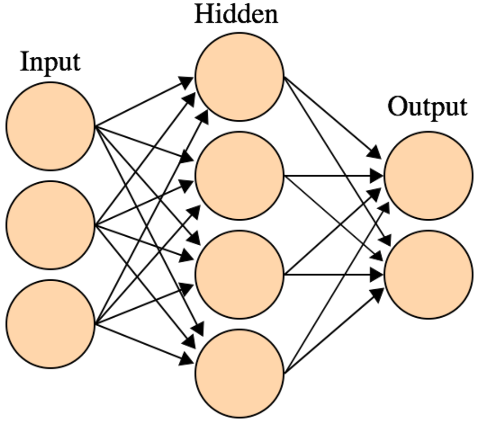
- Given acoustic observation $X = [x_1, x_2, \dots, x_T]$, goal of ASR is to find a word sequence $\hat{W} = [w_1, w_2, \dots, w_T]$ that has *Maximum a Posteriori* probability $P(W|X)$

x_t : *spectral features* q_k : *phones* w_ℓ : words t : *time*

Proposed Approach

x_{t-c}
.
x_t
.
x_{t+c}

Spectral Features (MFCC)



Deep Neural Network

$p(q_1|x_t)$
$p(q_2|x_t)$
.
.
$p(q_K|x_t)$

Phone Posteriors

Dictionary Learning & Sparse Recovery

Compressive Sensing Approach

$p(W|X)$

Word Sequence Posterior

Exploiting Posterior Features

- Given phone posterior probability $\mathrm{p}(\mathrm{q}_\mathrm{k}|\mathrm{x}_\mathrm{t})$ for phone q_k at frame x_t, the word level posterior probabilities can be generated by marginalization over L hidden variables w_l :

$$\boldsymbol{p(q_k|x_t)} \\ = \sum_{l=1}^{L} \boldsymbol{p(q_k, w_l|x_t)} \\ = \sum_{l=1}^{L} \boldsymbol{p(q_k|w_l, x_t)p(w_l|x_t)} \\ = \sum_{l=1}^{L} \boldsymbol{p(q_k|w_l)p(w_l|x_t)}$$

Exploiting Posterior Features

- Given phone posterior probability $\mathrm{p}(\mathrm{q_k}|\mathrm{x_t})$ for phone $\mathrm{q_k}$ at frame $\mathrm{x_t}$, the word level posterior probabilities can be generated by marginalization over L hidden variables $\mathrm{w_l}$:

$$\boldsymbol{p(q_k|x_t)}$$

Phone posteriors

$$= \sum_{l=1}^{L} \boldsymbol{p(q_k, w_l|x_t)}$$

$$= \sum_{l=1}^{L} \boldsymbol{p(q_k|w_l, x_t)p(w_l|x_t)}$$

$$= \sum_{l=1}^{L} \boldsymbol{p(q_k|w_l)p(w_l|x_t)}$$

Word posteriors

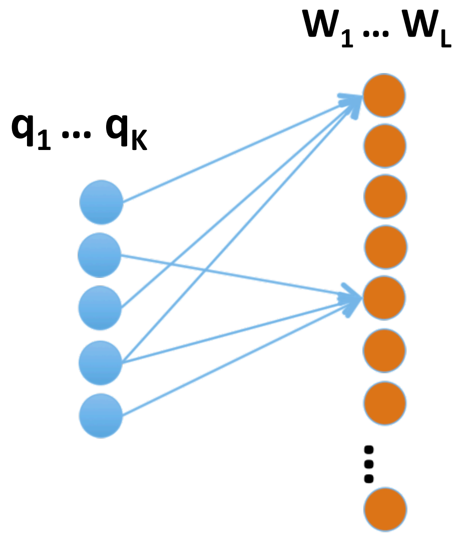
Exploiting Posterior Features

- Given phone posterior probability $\mathrm{p}(\mathrm{q_k}|\mathrm{x_t})$ for phone $\mathrm{q_k}$ at frame $\mathrm{x_t}$, the word level posterior probabilities can be generated by marginalization over L hidden variables $\mathrm{w_l}$:

Phone posteriors

$$
\begin{aligned}
\boldsymbol{p(q_k|x_t)} &= \sum_{l=1}^{L} \boldsymbol{p(q_k, w_l|x_t)} \\
&= \sum_{l=1}^{L} \boldsymbol{p(q_k|w_l, x_t)\,p(w_l|x_t)} \\
&= \sum_{l=1}^{L} \boldsymbol{p(q_k|w_l)\,p(w_l|x_t)}
\end{aligned}
$$

Word posteriors



# of Words, L

>>

# of Phones, K

A Compressive Sensing Approach
Phones to Word Posteriors

$$\boldsymbol{p(q_k|x_t) = \sum_{l=1}^{L} p(q_k|w_l)p(w_l|x_t) \quad where\ L \gg k}$$

$$\underbrace{\begin{bmatrix} p(q_1|x_t) \\ p(q_2|x_t) \\ \vdots \\ p(q_K|x_t) \end{bmatrix}}_{Z} = \underbrace{\begin{bmatrix} p(q_1|w_1) & \cdots & p(q_1|w_l) & \cdots & p(q_1|w_L) \\ p(q_2|w_1) & \cdots & p(q_2|w_l) & \cdots & p(q_2|w_L) \\ \vdots & \vdots & & & \\ p(q_K|w_1) & \cdots & p(q_K|w_l) & \cdots & p(q_K|w_L) \end{bmatrix}}_{\text{Dictionary Matrix: } D} \times \underbrace{\begin{bmatrix} p(w_1|x_t) \\ \vdots \\ p(w_l|x_t) \\ \vdots \\ p(w_L|x_t) \end{bmatrix}}_{\boldsymbol{\alpha}}$$

ASR can be cast as recovering high-dimensional sparse word posteriors from low-dimensional (phonetic) observations

A Compressive Sensing Approach
Phones to Word Posteriors

$$\boldsymbol{p(q_k|x_t) = \sum_{l=1}^{L} p(q_k|w_l)p(w_l|x_t) \quad where\ L \gg k}$$

From DNN Output

$$\underbrace{\begin{bmatrix} p(q_1|x_t) \\ p(q_2|x_t) \\ \vdots \\ p(q_K|x_t) \end{bmatrix}}_{Z} = \underbrace{\begin{bmatrix} p(q_1|w_1) & \cdots & p(q_1|w_l) & \cdots & p(q_1|w_L) \\ p(q_2|w_1) & \cdots & p(q_2|w_l) & \cdots & p(q_2|w_L) \\ \vdots & \vdots & & & \\ p(q_K|w_1) & \cdots & p(q_K|w_l) & \cdots & p(q_K|w_L) \end{bmatrix}}_{\text{Dictionary Matrix: } D} \times \underbrace{\begin{bmatrix} p(w_1|x_t) \\ \vdots \\ p(w_l|x_t) \\ \vdots \\ p(w_L|x_t) \end{bmatrix}}_{\boldsymbol{\alpha}}$$

Word Posteriors can directly be decoded into Word Sequences

ASR can be cast as recovering high-dimensional sparse word posteriors from low-dimensional (phonetic) observations

Modeling Word Manifold (Sub-dictionaries)

word w_ℓ

$$\underbrace{\begin{bmatrix} p(q_1|x_t) \\ p(q_2|x_t) \\ \vdots \\ p(q_K|x_t) \end{bmatrix}}_{z} = \underbrace{\begin{bmatrix} p(q_1|w_1) & \cdots & p(q_1|w_l) & \cdots & p(q_1|w_L) \\ p(q_2|w_1) & \cdots & p(q_2|w_l) & \cdots & p(q_2|w_L) \\ \vdots & \vdots & & & \\ p(q_K|w_1) & \cdots & p(q_K|w_l) & \cdots & p(q_K|w_L) \end{bmatrix}}_{\text{Dictionary Matrix: } D} \times \underbrace{\begin{bmatrix} p(w_1|x_t) \\ \vdots \\ p(w_l|x_t) \\ \vdots \\ p(w_L|x_t) \end{bmatrix}}_{\boldsymbol{\alpha}}$$

- A word w lies in a complex non-linear manifold.

Modeling Word Manifold (Sub-dictionaries)

word w_ℓ

$$\underbrace{\begin{bmatrix} p(q_1|x_t) \\ p(q_2|x_t) \\ \vdots \\ p(q_K|x_t) \end{bmatrix}}_{Z} = \underbrace{\begin{bmatrix} p(q_1|w_1) & \cdots & p(q_1|w_l) & \cdots & p(q_1|w_L) \\ p(q_2|w_1) & \cdots & p(q_2|w_l) & \cdots & p(q_2|w_L) \\ \vdots & \vdots & & & \\ p(q_K|w_1) & \cdots & p(q_K|w_l) & \cdots & p(q_K|w_L) \end{bmatrix}}_{\text{Dictionary Matrix: } D} \times \underbrace{\begin{bmatrix} p(w_1|x_t) \\ \vdots \\ p(w_l|x_t) \\ \vdots \\ p(w_L|x_t) \end{bmatrix}}_{\boldsymbol{\alpha}}$$

$$\underbrace{\begin{bmatrix} p(q_1|w_l) \\ p(q_2|w_l) \\ \vdots \\ p(q_K|w_l) \end{bmatrix}} = \underbrace{\begin{bmatrix} p(q_1|sw_1^{w_l}) & \cdots & p(q_1|sw_s^{w_l}) & \cdots & p(q_1|sw_{S_{w_l}}^{w_l}) \\ p(q_2|sw_1^{w_l}) & \cdots & p(q_2|sw_s^{w_l}) & \cdots & p(q_2|sw_{S_{w_l}}^{w_l}) \\ \vdots & \vdots & & & \\ p(q_K|sw_1^{w_l}) & \cdots & p(q_K|sw_s^{w_l}) & \cdots & p(q_K|sw_{S_{w_l}}^{w_l}) \end{bmatrix}} \times \begin{bmatrix} p(sw_1^{w_l}|w_l) \\ \vdots \\ p(sw_s^{w_l}|w_l) \\ \vdots \\ p(sw_{S_{w_l}}^{w_l}|w_l) \end{bmatrix}$$

Manifold of a word $\boldsymbol{w_\ell}$ modelled by union of subspaces in sub-dictionary: D_w

- A word w lies in a complex non-linear manifold defined by
- Sub-dictionary D_w models the *word manifold* as a union of subspaces (called subwords)

Modeling Word Manifold (Sub-dictionaries)

word w_ℓ

$$\underbrace{\begin{bmatrix} p(q_1|x_t) \\ p(q_2|x_t) \\ \vdots \\ p(q_K|x_t) \end{bmatrix}}_{z} = \underbrace{\begin{bmatrix} p(q_1|w_1) & \cdots & p(q_1|w_l) & \cdots & p(q_1|w_L) \\ p(q_2|w_1) & \cdots & p(q_2|w_l) & \cdots & p(q_2|w_L) \\ \vdots & \vdots & & & \\ p(q_K|w_1) & \cdots & p(q_K|w_l) & \cdots & p(q_K|w_L) \end{bmatrix}}_{\text{Dictionary Matrix: } D} \times \underbrace{\begin{bmatrix} p(w_1|x_t) \\ \vdots \\ p(w_l|x_t) \\ \vdots \\ p(w_L|x_t) \end{bmatrix}}_{\boldsymbol{\alpha}}$$

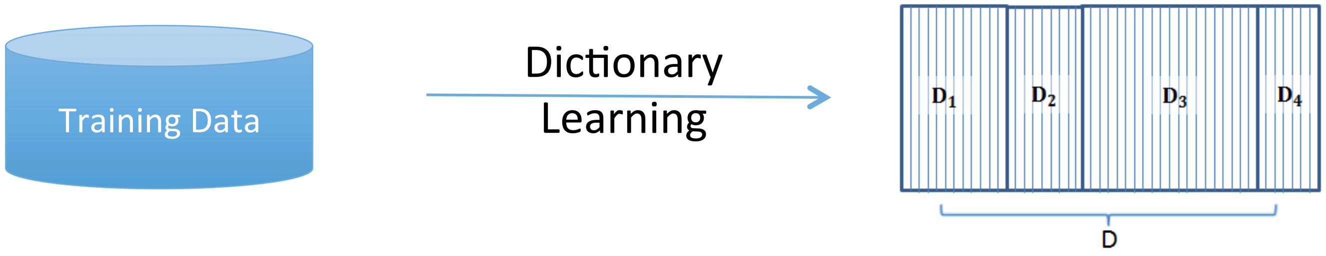
$$\begin{bmatrix} p(q_1|w_l) \\ p(q_2|w_l) \\ \vdots \\ p(q_K|w_l) \end{bmatrix} = \underbrace{\begin{bmatrix} p(q_1|sw_1^{w_l}) & \cdots & p(q_1|sw_s^{w_l}) & \cdots & p(q_1|sw_{S_{w_l}}^{w_l}) \\ p(q_2|sw_1^{w_l}) & \cdots & p(q_2|sw_s^{w_l}) & \cdots & p(q_2|sw_{S_{w_l}}^{w_l}) \\ \vdots & \vdots & & & \\ p(q_K|sw_1^{w_l}) & \cdots & p(q_K|sw_s^{w_l}) & \cdots & p(q_K|sw_{S_{w_l}}^{w_l}) \end{bmatrix}} \times \begin{bmatrix} p(sw_1^{w_l}|w_l) \\ \vdots \\ p(sw_s^{w_l}|w_l) \\ \vdots \\ p(sw_{S_{w_l}}^{w_l}|w_l) \end{bmatrix}$$

Manifold of a word modelled by union of subspaces in sub-dictionary: D_w

- A word w lies in a complex non-linear manifold defined by
- Sub-dictionary D_w models the *word manifold* as a union of subspaces (called subwords)
- Sparse recovery chooses a low-dimensional union of subspaces from an exponentially large number of such unions.

Dictionary Learning

- Dictionary Learning: Finding an over-complete basis set for sparse representation



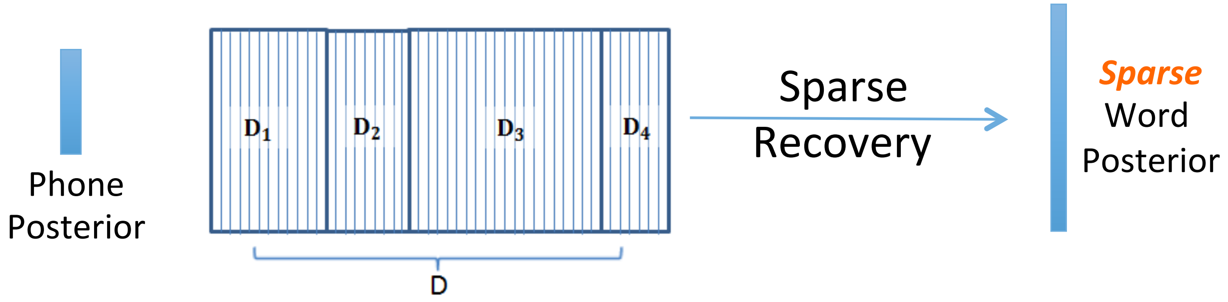
- Each sub-dictionary can be learnt *Independently* here

$$\hat{\mathbf{D}}_W = \arg\min_{\mathbf{D}} \left\{ \frac{1}{t} \sum_{i=1}^{t} \left(\frac{1}{2} \| z_i^W - \mathbf{D}_W \alpha_i \|_2^2 + \lambda \| \alpha_i \|_1 \right) \right\}$$

Prominent methods include Online algorithm (Mairal,2009), K-SVD algorithm (Aharon and Elad, 2005).

Sparse Recovery

- Sparse Recovery: Solving l_0 (or l_1)-norm sparse recovery minimization for α.



Given

- z : a phone posterior vector
- $\boldsymbol{D}$: an over-complete dictionary matrix for words

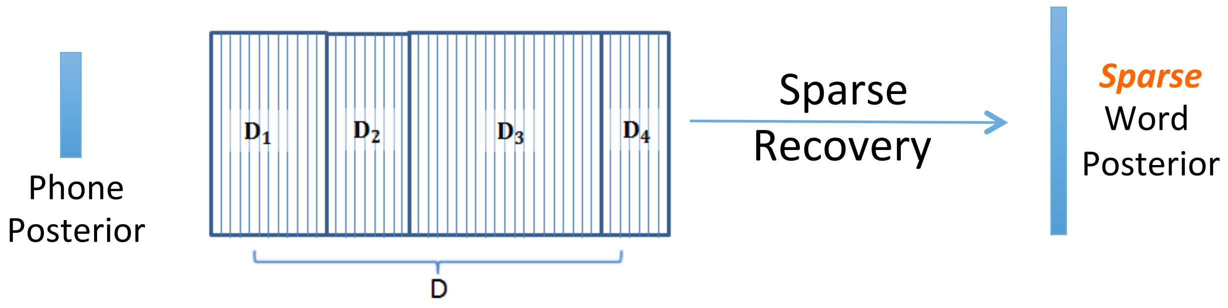
a sparse word posterior α is obtained

$$\hat{\alpha} = \arg\min_{\alpha} ||z - \boldsymbol{D}\alpha||_2^2 + \lambda||\alpha||_1$$

by solving the **Lasso** l_1-norm minimization problem.

Sparse Recovery

- Sparse Recovery: Solving l_0 (or l_1)-norm sparse recovery minimization for *α.*



Given

- z : a phone posterior vector
- $\boldsymbol{D}$: an over-complete dictionary matrix for words

a sparse word posterior α is obtained

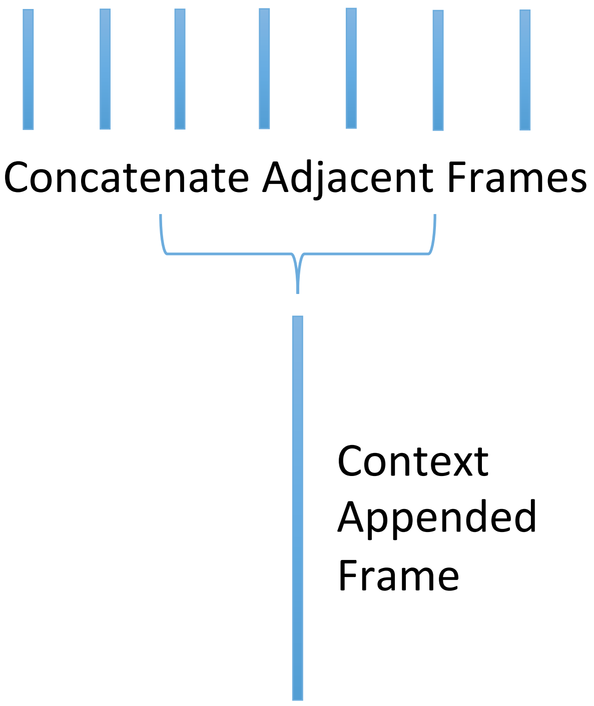
How to handle sequential information inherent in Speech signals ?

$$\hat{\alpha} = \underset{\alpha}{arg\,min}\ \left|\left|z - \boldsymbol{D}\alpha\right|\right|_2^2 + \lambda\left|\left|\alpha\right|\right|_1$$

by solving the **Lasso** l_1-norm minimization problem.

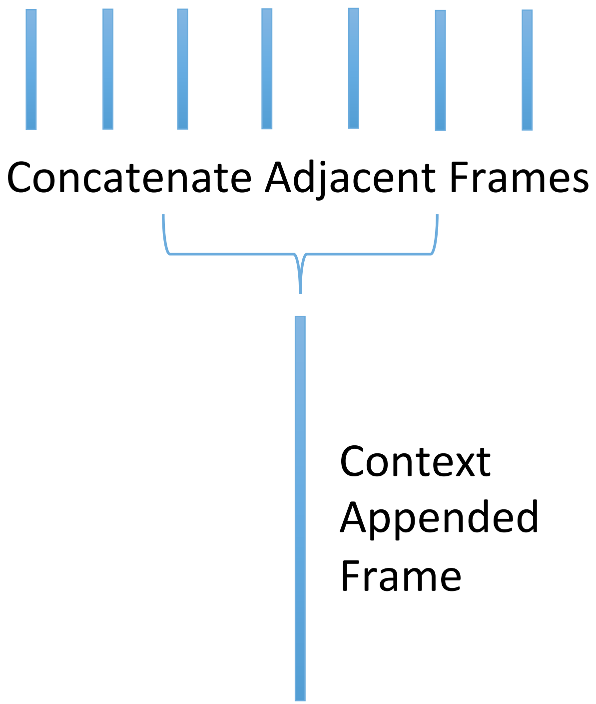
Capturing Sequential Information

- **Contextual Embedding**: Phone posterior feature vectors are used in the framework after appending a context (vertically concatenating adjacent frames)

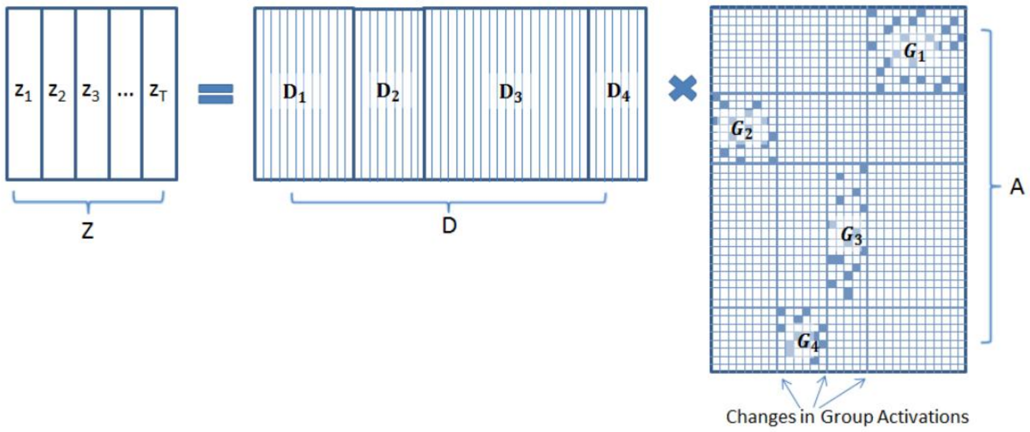


Capturing Sequential Information

- **Contextual Embedding**: Phone posterior feature vectors are used in the framework after appending a context (vertically concatenating adjacent frames)

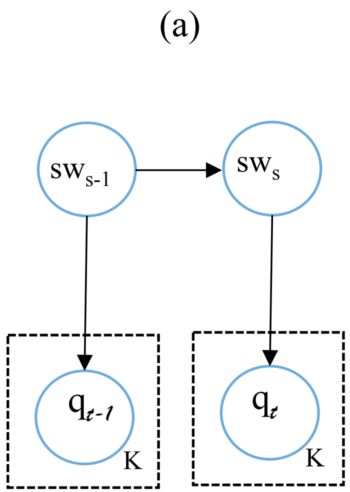


- **Structured Sparsity**: presence of ***group*** and ***hierarchical*** sparsity can be exploited using techniques like Collaborative-Hierarchical Lasso Sprechmann (or C-HiLasso) [Sprechmann, 2011].

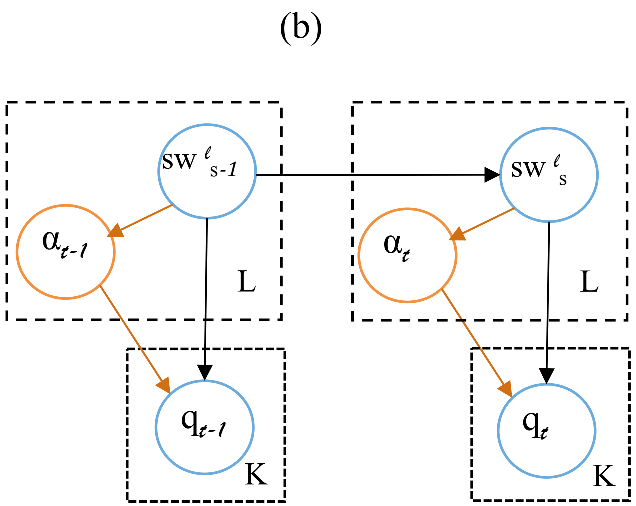


$$\min_{\alpha} \frac{1}{2} ||Z - DA||_F^2 + \lambda_2 \sum_{g \in G} ||A^g||_F + \lambda_1 \sum_{t=1}^{T} ||\alpha||_1$$

Posterior-based Sparse Modeling for ASR



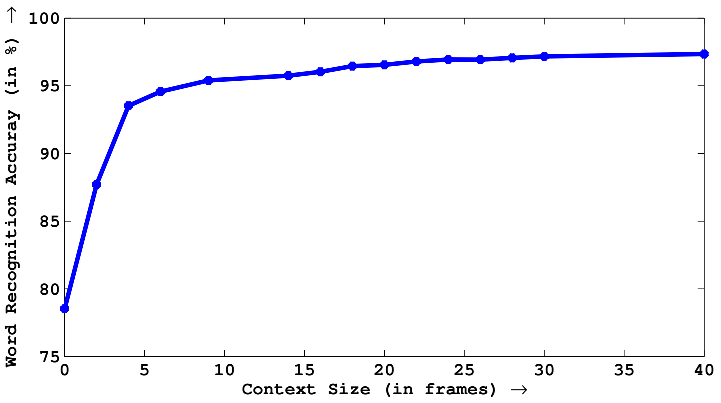
(a) Graphical model for the conventional acoustic modeling



(b) Graphical model for posterior-based sparse modeling framework.

Experimental Results

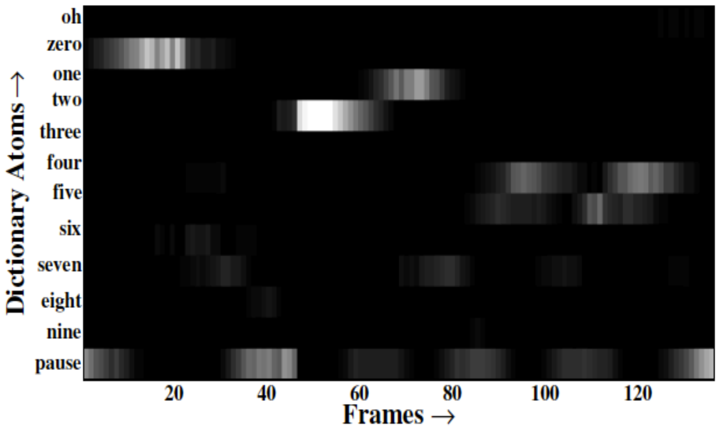
Effect of Increasing Context Size on Performance



Isolated Word Recognition (Accuracy %)

Task	DTW	Compressive Sensing
Phonebook (75)	84.7%	97.8%
Phonebook (600)	73.5%	93.2%

Structured sparsity of continuous speech
Digit Sequence 0-2-1-4-4



Connected Word Recognition (100- Word Error Rate %)

ASR Task	Collection of Exemplars	Compressive Sensing	Dimensionality reduction
Numbers	78.6%	87.5%	97%

Conclusions

- ✓ We propose a *compressing sensing* based alternative to traditional HMM for ASR.
- ✓ Working in posterior domain directly gives word sequence posteriors.
- ✓ Learning a dictionary alleviates the need of huge database of exemplars needed in previous sparse representation approaches.
- ✓ Sequence Information can be tackled with variants of Lasso that enforce group and collaborative sparsity.

Future Work:

- ⌘ Evaluate the approach on Large Vocabulary continuous speech recognition (LVCSR) tasks.
- ⌘ Improving sequential information processing by integrating *Language Modeling*.

References

- G. Aradilla, H. Bourlard et al., "Posterior features applied to speech recognition tasks with user-defined vocabulary," in Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009, pp. 3809–3812.
- P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar,"C-HiLasso: A collaborative hierarchical sparse modeling framework," Signal Processing, IEEE Transactions on, vol. 59, no. 9, pp. 4183–4198, 2011.
- J. Pitrelli, C. Fong, S. Wong, J. Spitz, and H. Leung, "Phonebook: a phonetically-rich isolated-word telephone-speech database," in Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on, vol. 1, May 1995, pp. 101–104 vol.1.
- R. A. Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at cslu," 1995.
- J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," Journal of Machine Learning Research (JMLR), vol. 11, pp. 19–60, 2010.
- B. Efron, T. Hastie, I. Johnstone, R. Tibshirani et al., "Least angle regression," The Annals of statistics, vol. 32, no. 2, pp. 407–499, 2004.

Thank You ☺

Questions ?