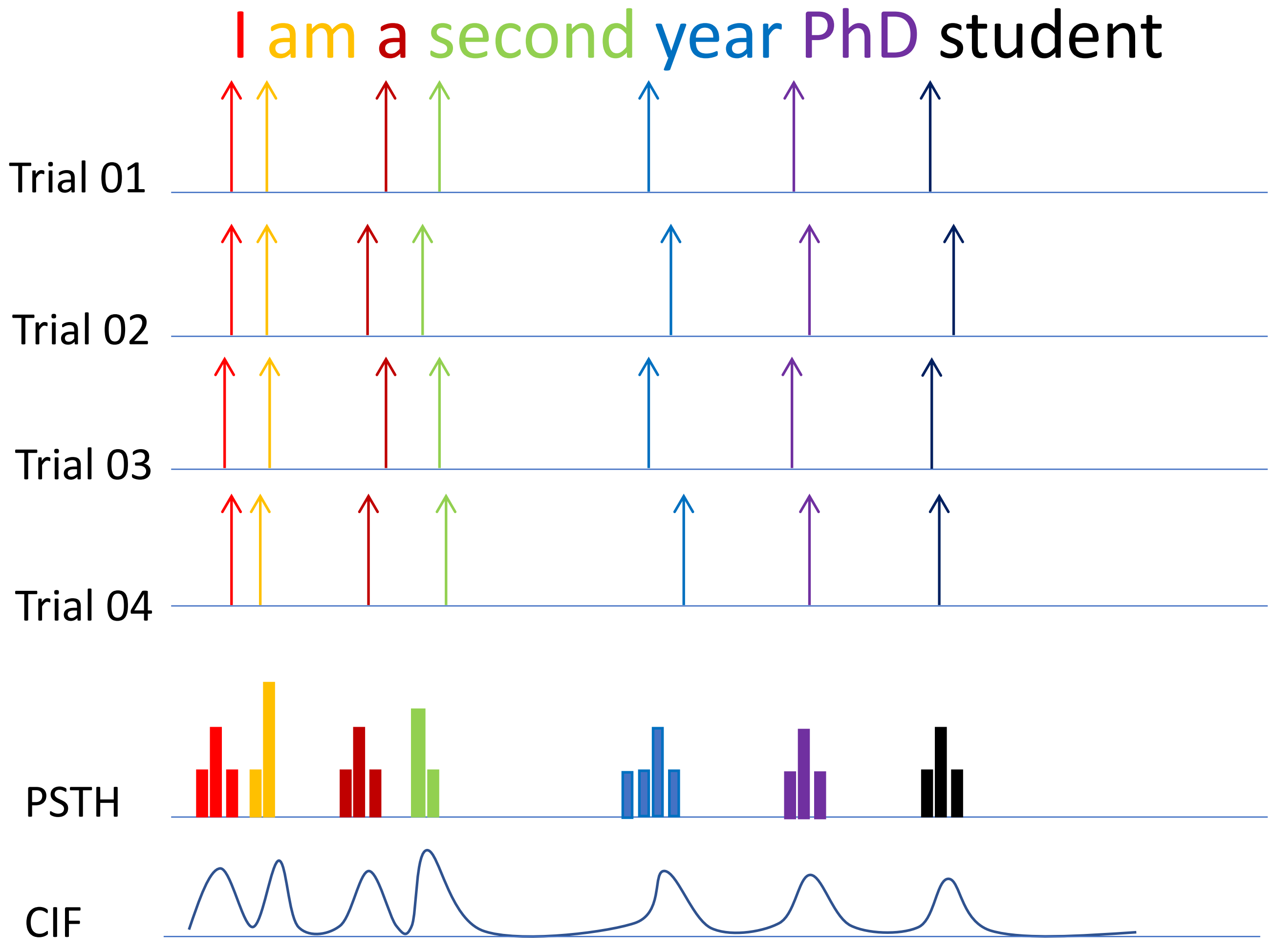
Specific word onset prediction

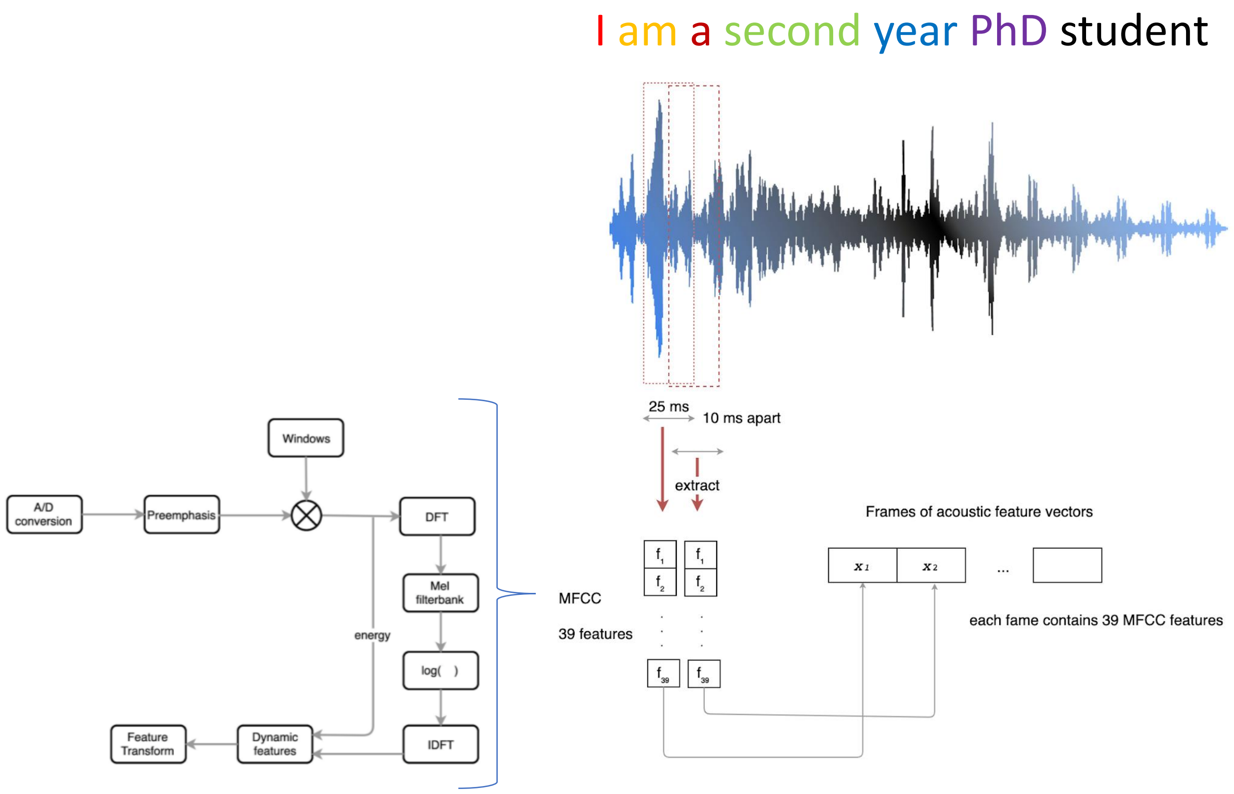
Mohammad Reza Rezaei

April 1st 2022

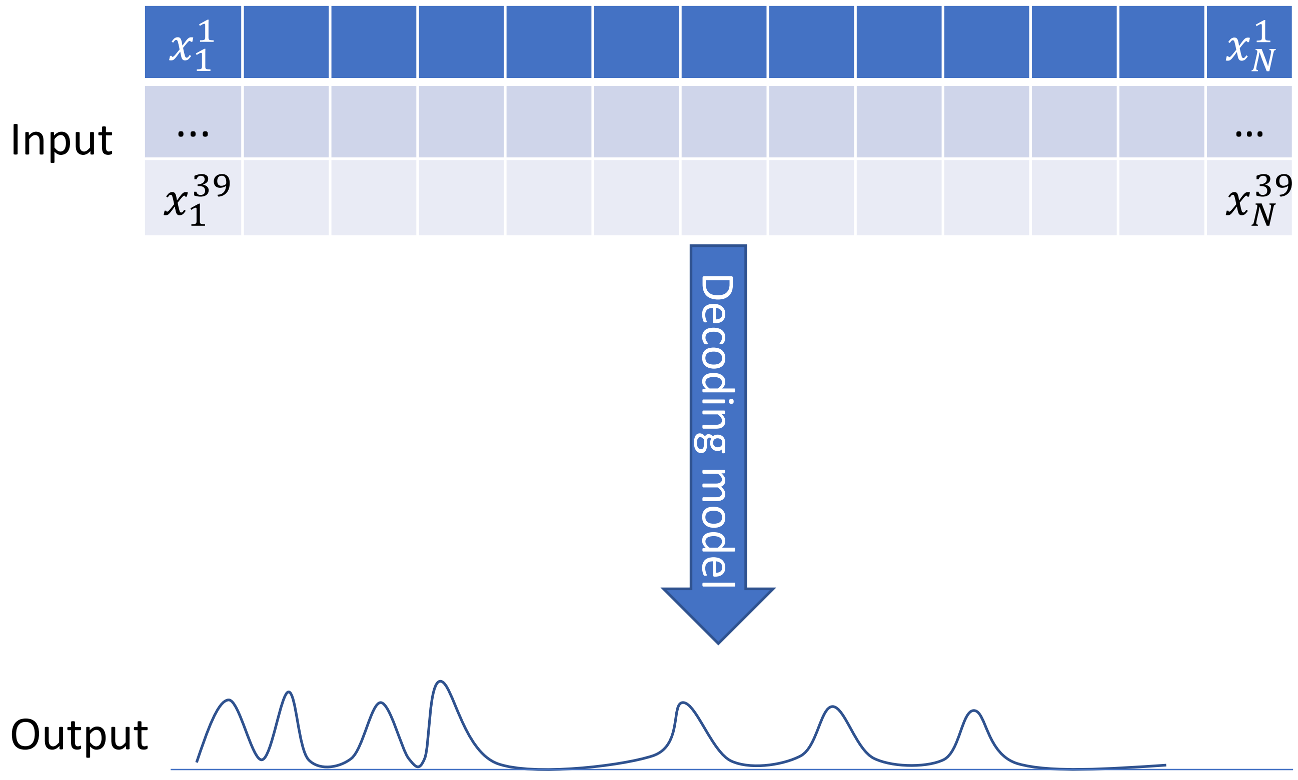
The behavioral model



Acoustic feature extraction



Decoding the behavior from acoustic features



Predict a specific phoneme/word onset z_t^k

I am a second year PhD student and I am really enjoying it

- Goal

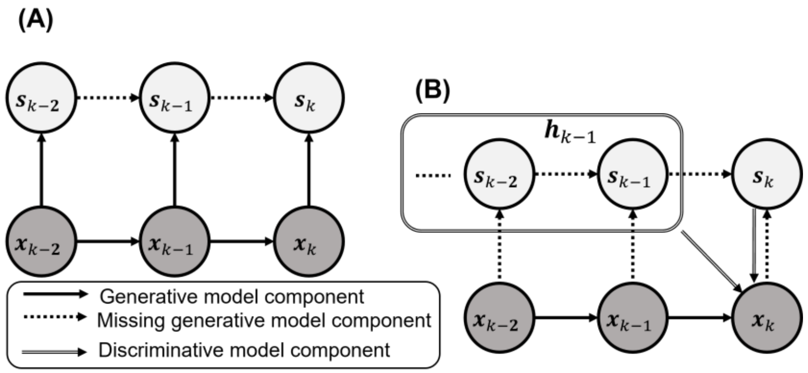
$$P\left(z_t^k \middle| X_{1:t}, \boldsymbol{l}_{1:k}, \boldsymbol{d}_{1:k}, \boldsymbol{w}_{1:k}\right) = P\left(z_t^k \middle| X_{1:t}, \boldsymbol{l}_{1:k}, \boldsymbol{d}_{1:k}\right)$$
$$P\left(z_t^k \middle| X_{1:t}, \boldsymbol{l}_{1:k}, \boldsymbol{d}_{1:k}\right) \propto P\left(\boldsymbol{x}_t \middle| X_{1:t-1}, z_t^k, \boldsymbol{l}_{1:k}, \boldsymbol{d}_{1:k}\right)$$

Predict a specific phoneme/word onset z_t^k

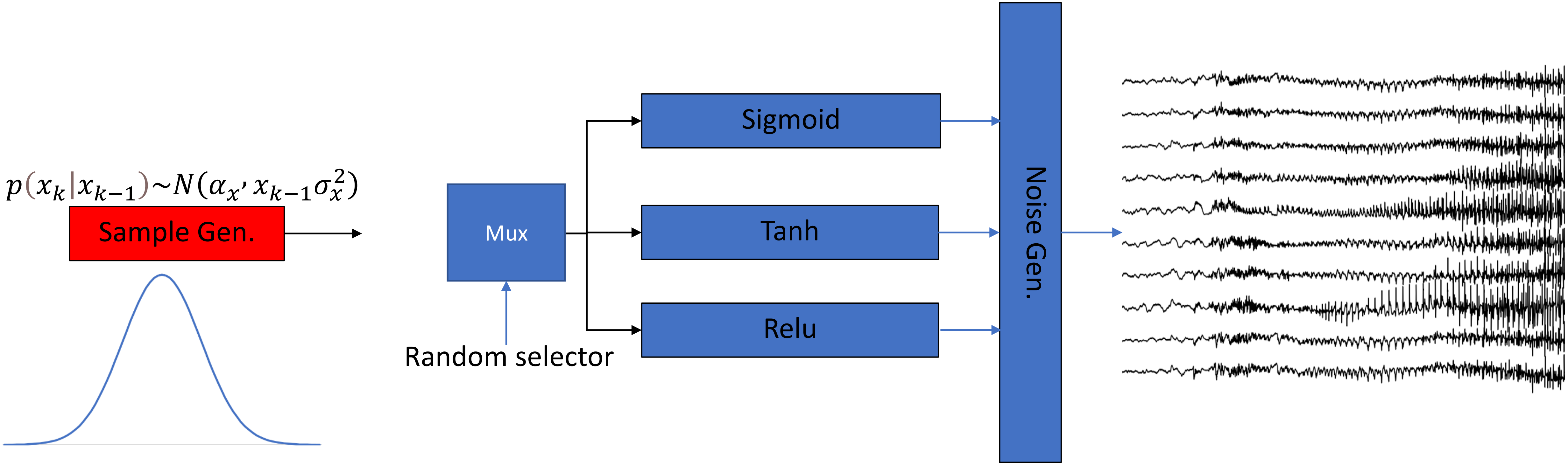
I am a second year PhD student

- $z_t^k \in R^1$defined as observation of k^{th} phoneme of the sequence $W_{1:K}$ at time t, $k = 1, \ldots, K, t = 1, \ldots, T$
- $\boldsymbol{x}_t \in R^{39}$ is the acoustic features at time t
- $\boldsymbol{w}_{1:K}$ word/phoneme sequence (deterministic)
- l_k defines as duration of k^{th} phoneme (can be learned by a statistical model)
- $d_k | d_{k-1}$,1st order autoregressive model, defines as the time delay between oberving k^{th} phoneme given previous one (can be learned by a statistical model)

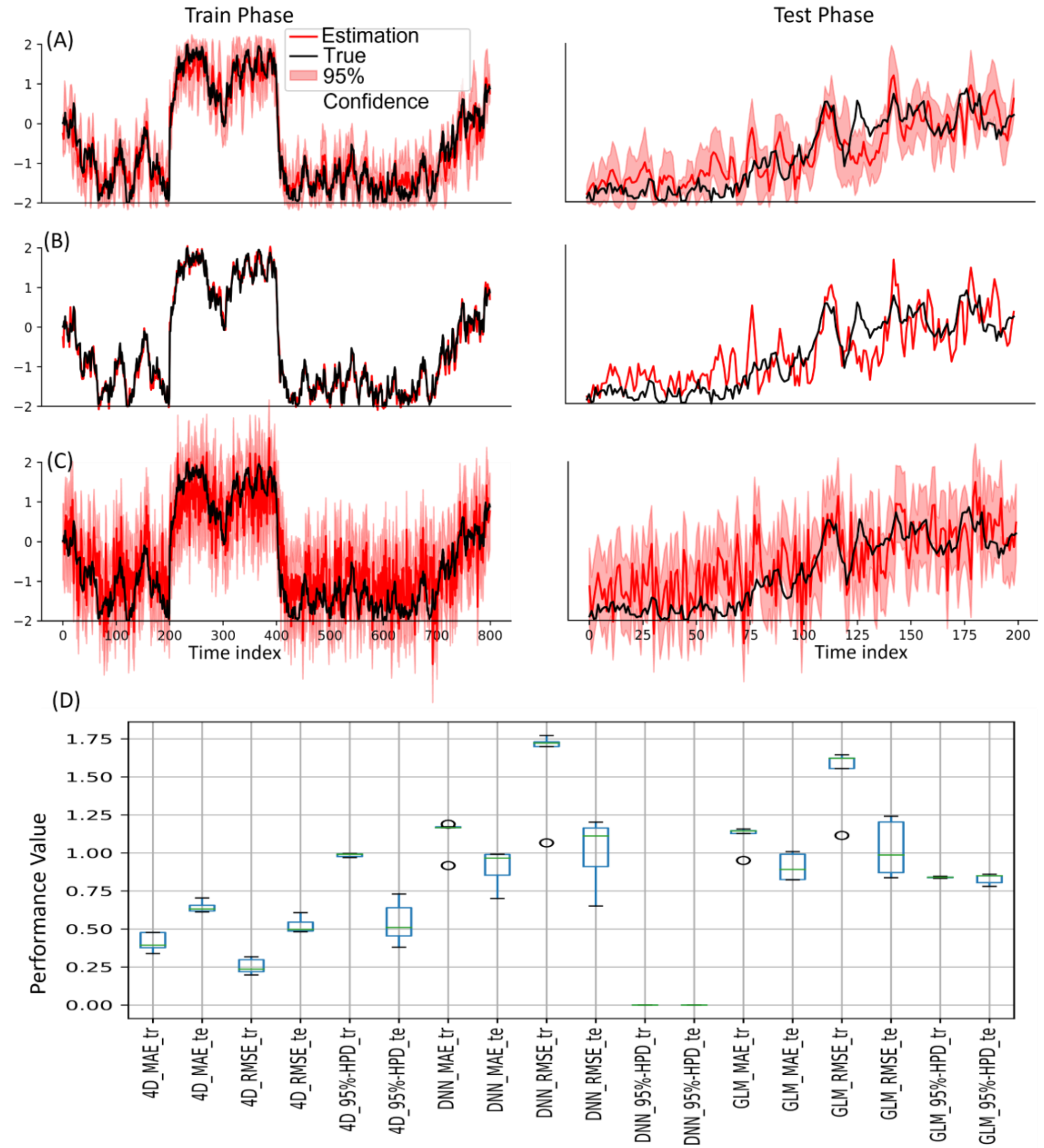
Synthetic data



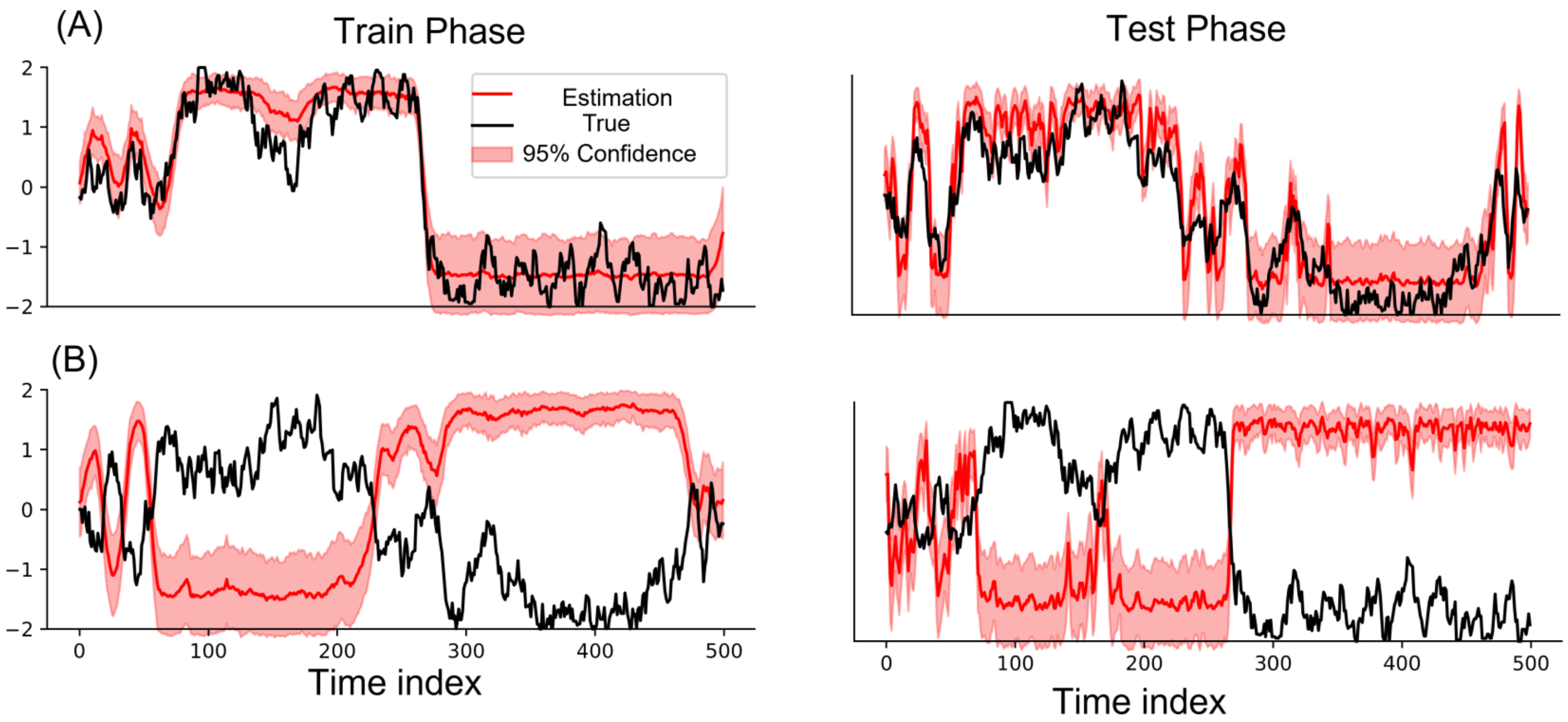
$$p(x_k|x_{k-1}) \sim N(\alpha_x, x_{k-1}\sigma_x^2)$$



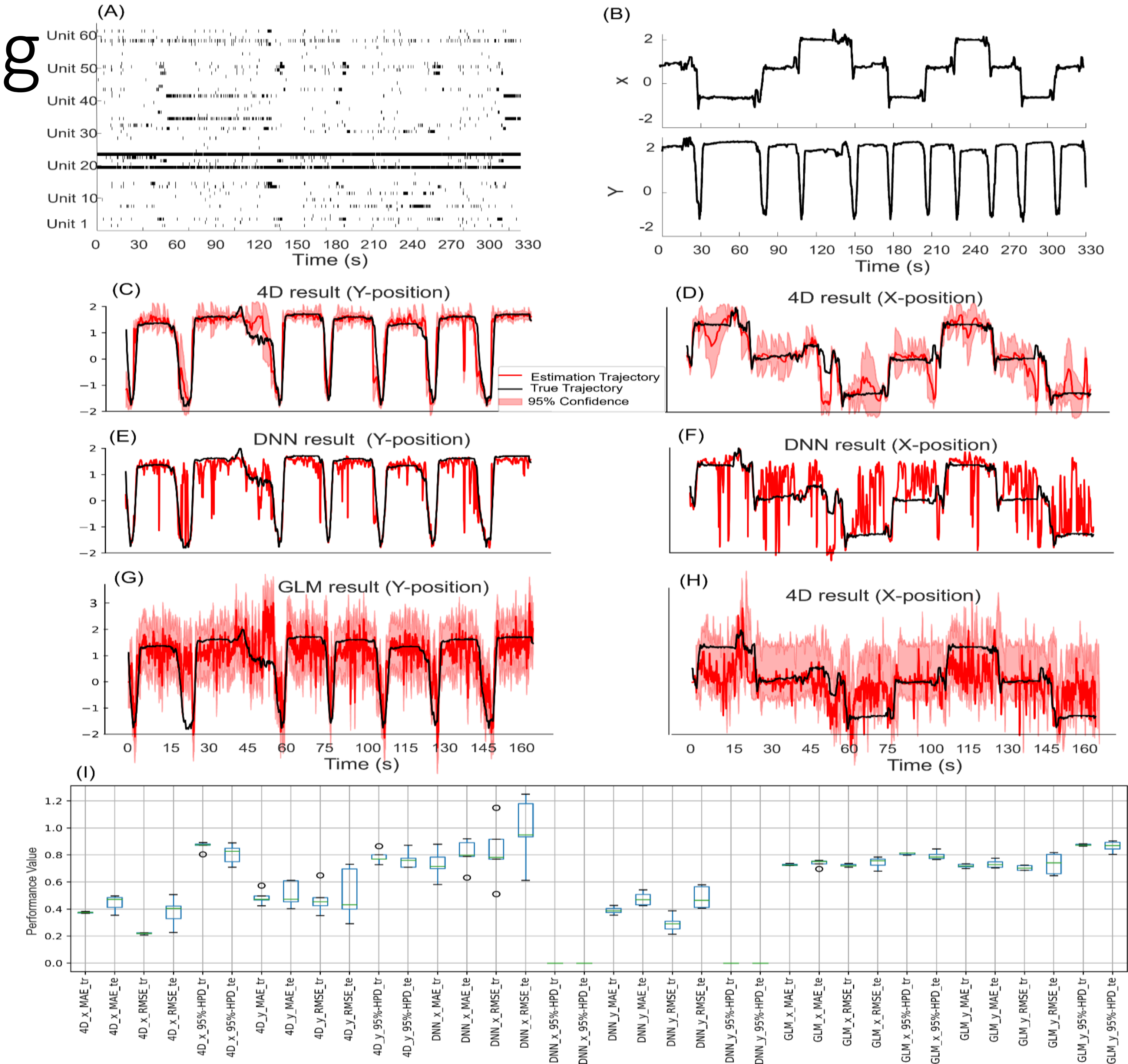
D4 supervised learning for synthetic data



D4 unsupervised- learning for synthetic data



D4 supervised learning for Wmazde data



Automatic speech recognition with D4

- Traditional generative model

$$P(w_k \mid X_k) \propto P(w_k | x_k) \sum_{w_{k-1}} P(w_k \mid w_{k-1}) P(w_{k-1} \mid X_{k-1})$$

- By using the **Deep Discriminative direct decoders (4D)** theory we can represent the ASR as

$$P(w_k | X_k) \propto \frac{P(w_k | x_k, h_k)}{\sum_{w_{k-1}} P(w_k | w_{k-1}) P(w_{k-1} | x_{k-1}, h_{k-1})} \sum_{w_{k-1}} P(w_k | w_{k-1}) P(w_{k-1} | X_{k-1})$$

- Here we assumed a **bigram** as the language model.

Need for a sophisticated prior model (language model)

Solution:

- Marked-point process language model?
  - See phones as the marks
    - Structures in word sequence can reduce complexity of marks model
    - Dirichlet process for marks
  - Language model embedded in mark process (ordinal marks?)

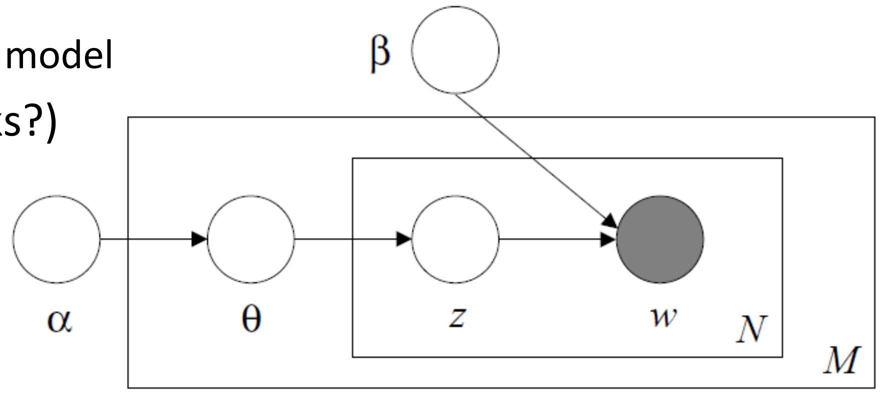
- Latent Dirichlet Language models?
  - Finite Mixture Models

Need for a sophisticated prior model (language model)

Solution:

- Marked-point process language model?
  - See phones as the marks
    - Structures in word sequence can reduce complexity of marks model
  - Language model embedded in mark process (ordinal marks?)

- Latent Dirichlet Language models?
  - Finite Gaussian Mixture Models
  - **Latent Dirichlet allocation LDA**



1. Choose $N \sim \mathrm{Poisson}(\xi)$.
2. Choose $\theta \sim \mathrm{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \mathrm{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Need for a sophisticated prior model (language model)

Solution:

- **Dirichlet Mixtures of Bayesian Linear Gaussian State-Space**

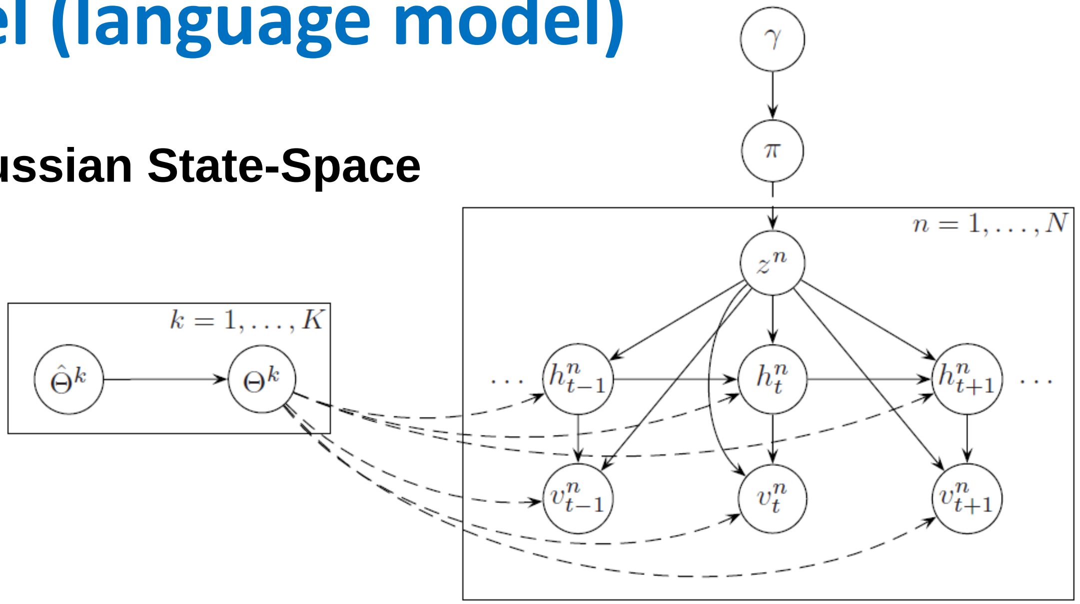


Figure 3: Graphical representation of the Dirichlet Mixture of Bayesian LGSSMs for performing clustering based on global similarity.

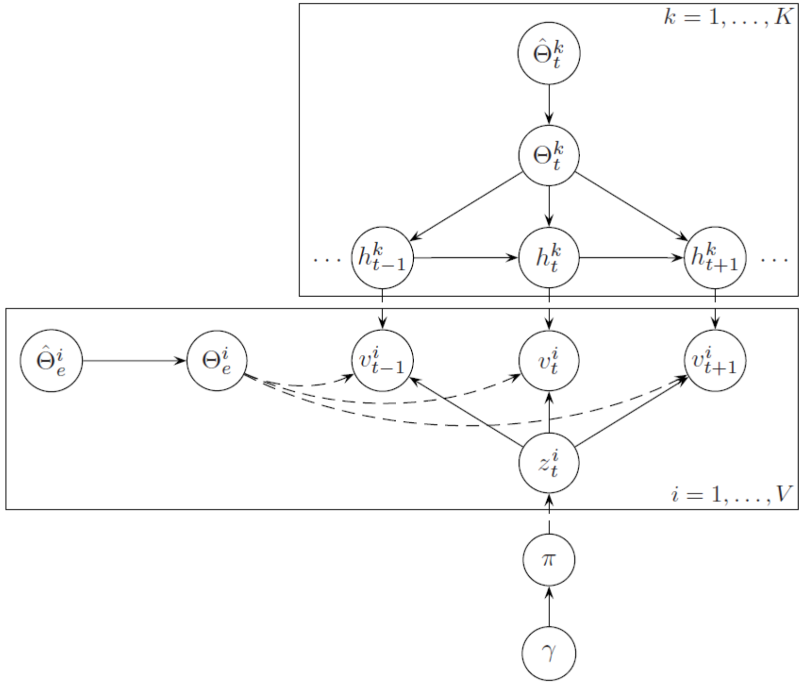


Figure 5: Graphical representation of the Dirichlet Mixture of Bayesian LGSSMs for performing clustering based on simultaneous similarity.

Simultaneous Similarity Our simultaneous similarity clustering approach assigns two time-series to the same cluster if they are derived from the *same realization* of a dynamical process.

Global Similarity The global similarity method will assign two time-series to the same cluster if they are generated by *different realizations* of the same dynamical process.

Speech recognition

- Predict a sequence of words (or phoneme sequence) $w_{1:k}^{*} \equiv W_k^{*}$ by observing a sequence of acoustics features $x_{1:k} \equiv X_k$.

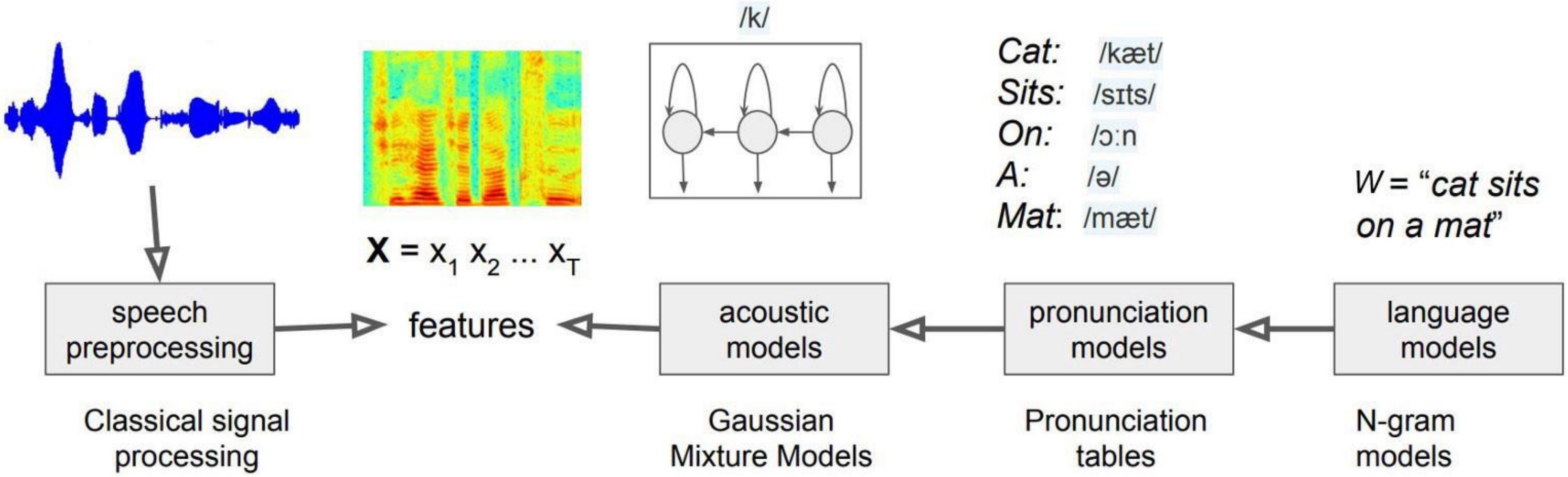
$$W_k^{*} = argmax_{W_k}\, P(W_k \,|\, X_k\,) \text{ Discriminative Models}$$

- By using **Bayes rule** we can rewrite it as

$$W_k^{*} = argmax_{W_k} \frac{P(X_k|W_k\,)P(W_k\,)}{P(X_k\,)} = argmax_{W_k}\, P(X_k \,|\, W_k\,)P(W_k) \text{ Generative Models}$$

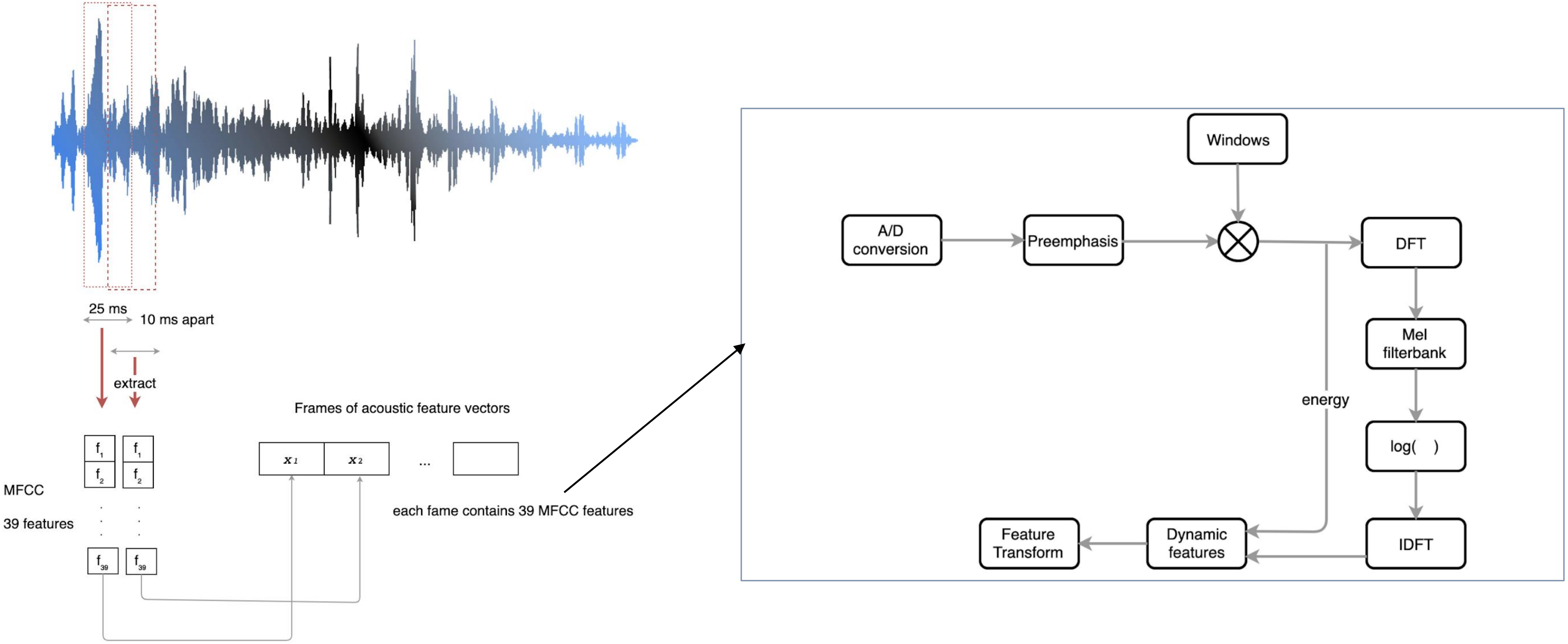
- $P(X_k\,|W_k)$ is the acoustic model: represents how the speech may sound given a sequence of words.
- $P(W_k)$ is the language model: describes the likelihood of the word sequence
- The generative models were more convenient to build in speech recognition before the deep neural networks (**DNNs**) entrance to this field

Speech recognition



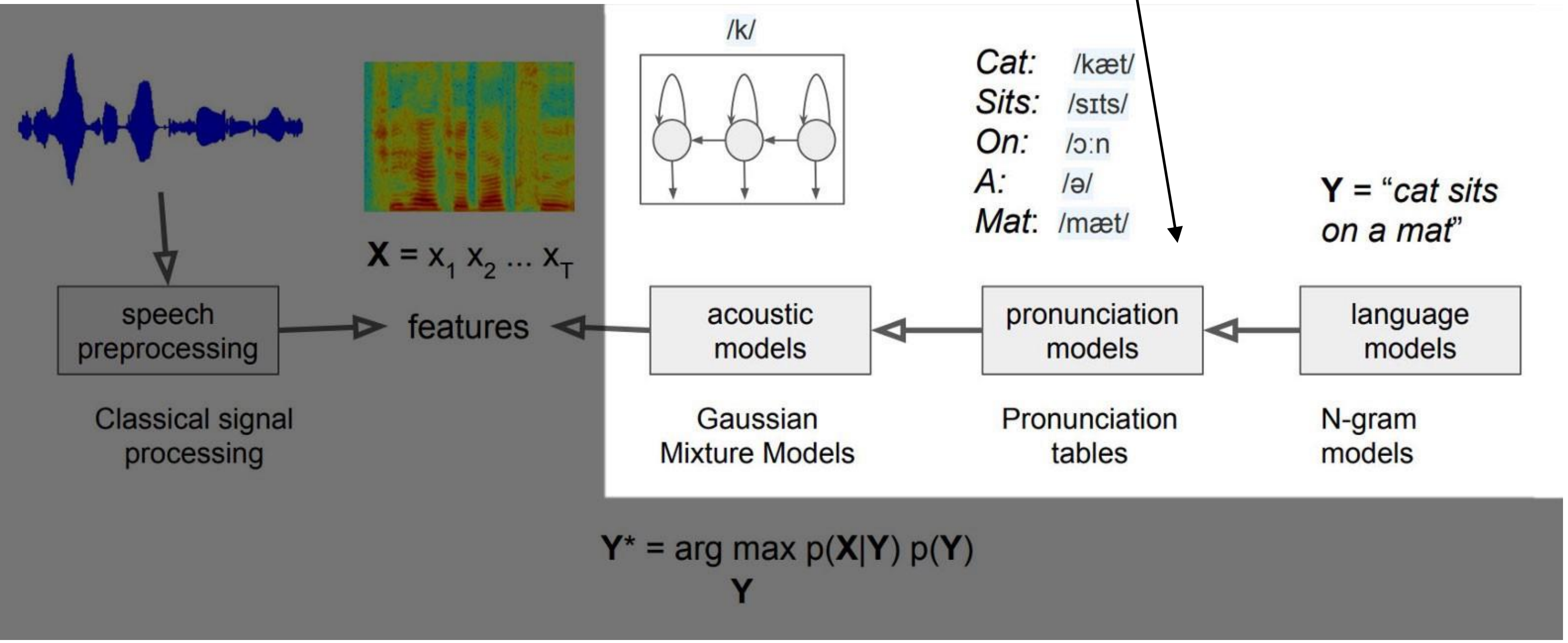
$$W_k^* = argmax_{W_k} P(X_k \mid W_k) P(W_k)$$

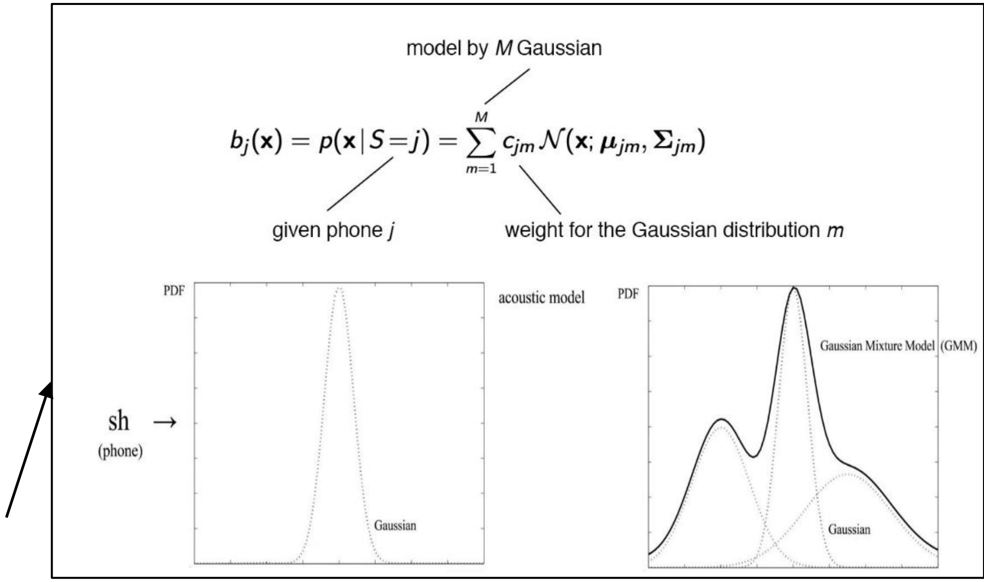
Speech feature extraction with Mel-frequency cepstral coefficients (MFCC)



Acoustic model

pronunciation table to produce the phones for the text sequence *Y (**Deterministic**)*



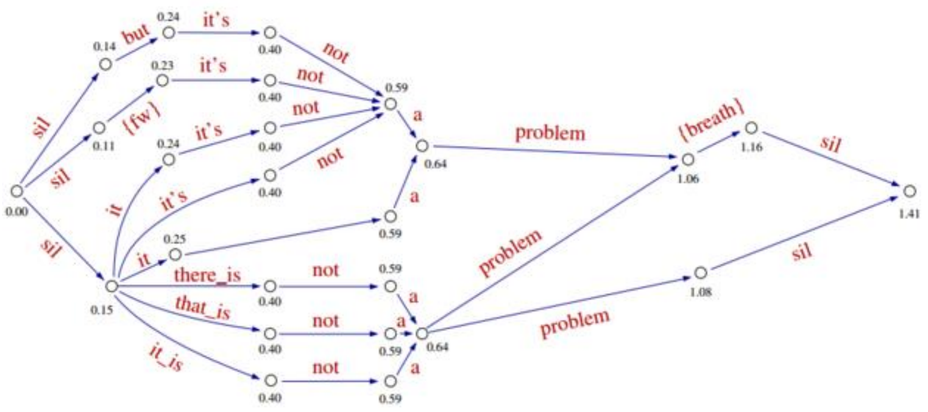


Language Model (N-gram)

Models probabilities of sequences of words

- Estimating the probability of a word w given a history of words, or the probability of an entire word sequence W (documents).

- N-gram: instead of computing the probability of a word given its entire history, we can approximate the history by just the last N words .



$$P(w_n|w_{1:n-1}) \approx P(w_n|w_{n-N+1:n-1})$$

Bi-gram

	i	**want**	**to**	**eat**	**chinese**	**food**	**lunch**	**spend**
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Automatic Speech recognition (ASR) summary