Estimation of articulatory gesture patterns from speech acoustics

Prasanta Kumar Ghosh[1], *Shrikanth Narayanan*[1], *Pierre Divenyi*[2], *Louis Goldstein*[3], *Elliot Saltzman*[4]

[1]Department of Electrical Engineering, University of Southern California, LA, CA, 90089

[2]EBIRE, Martinez, CA 94553

[3]Department of Linguistics, University of Southern California, LA, CA, 90089

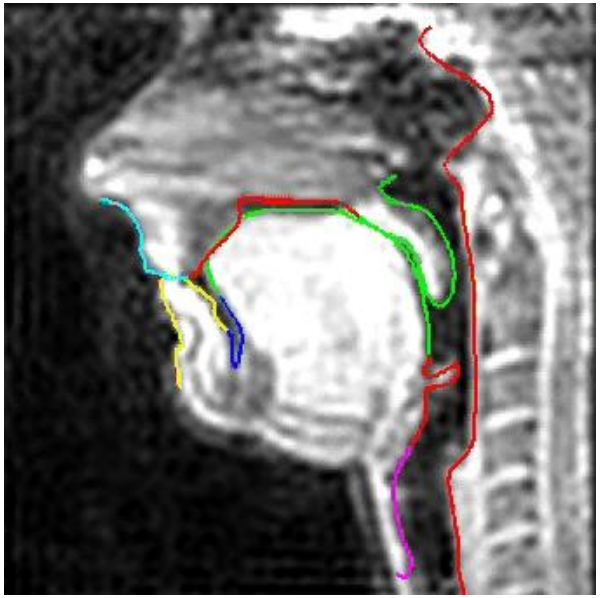
[4]Haskins Laboratories, New Haven, CT 06511

October 7, 2009

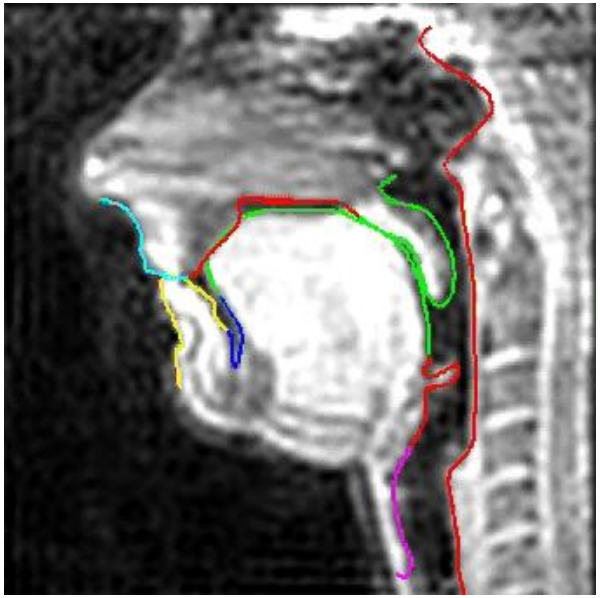
Outline

- Introduction to articulatory gesture pattern.
- Why is estimation of articulatory gesture patterns from speech important?
- Problem definition
- Proposed approaches for gesture pattern estimation
  1. Dynamic programming (DP) approach
  2. State model (SM) approach
- Dataset
- Experiment and evaluation
- Discussion and future works

Articulatory gesture pattern



Articulatory gesture pattern



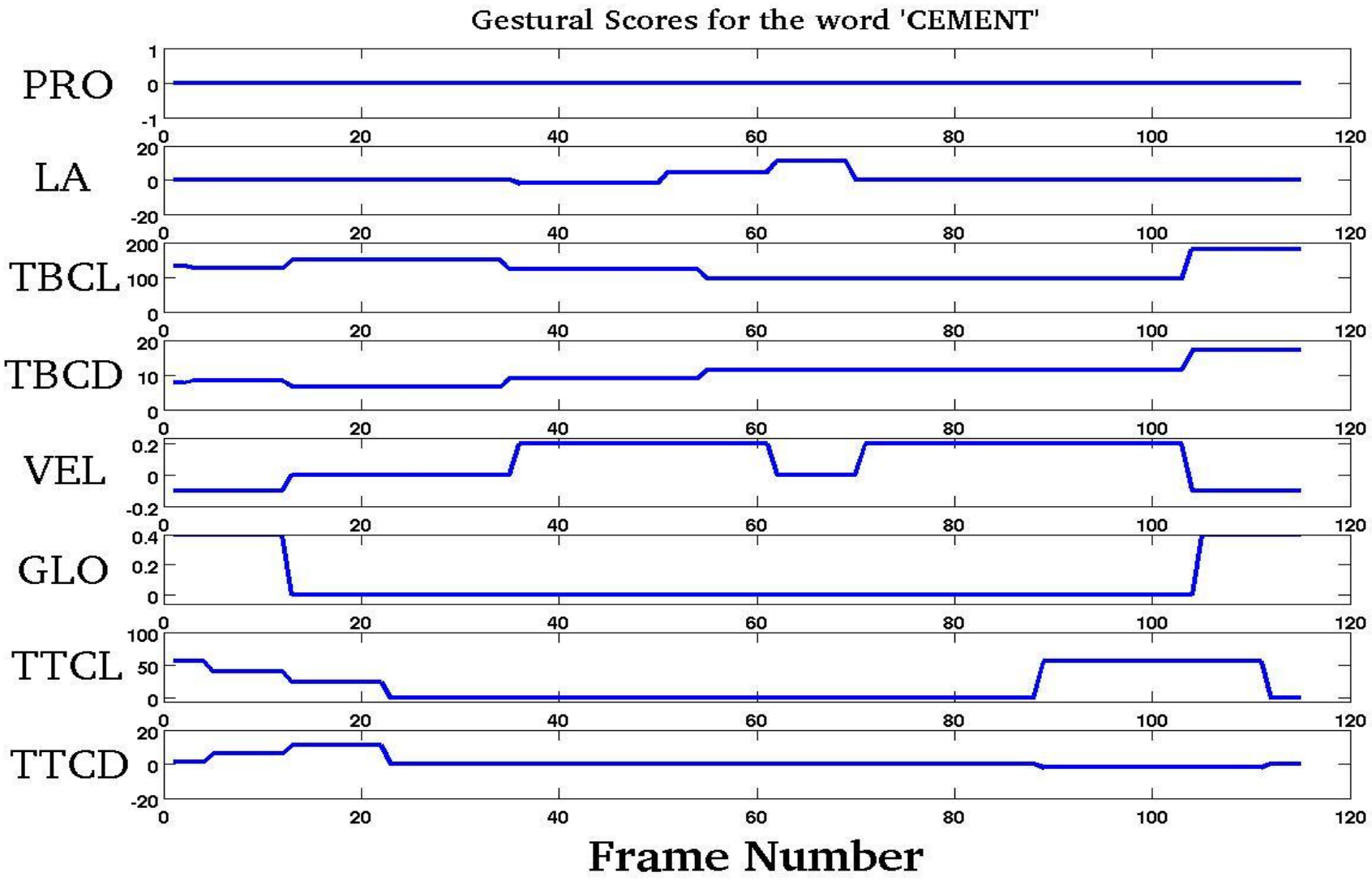
- Articulators move in coordinated fashion to produce speech and estimating articulation from speech is useful for many applications – acoustic-to-articulatory inversion.
- Unlike other descriptions of speech articulation, articulatory gesture pattern is a description which captures phonological information in speech.

Articulatory gesture pattern

Gestures are control regimes (functions overlapping in time) for

5 constriction degrees and 3 constriction locations

LA, TBCD, TTCD, VEL, GLO LP, TTCL, TBCL


Gestural Scores for the word 'CEMENT'
PRO
LA
TBCL
TBCD
VEL
GLO
TTCL
TTCD
Frame Number

Articulatory gesture pattern

Gestures are control regimes (functions overlapping in time) for

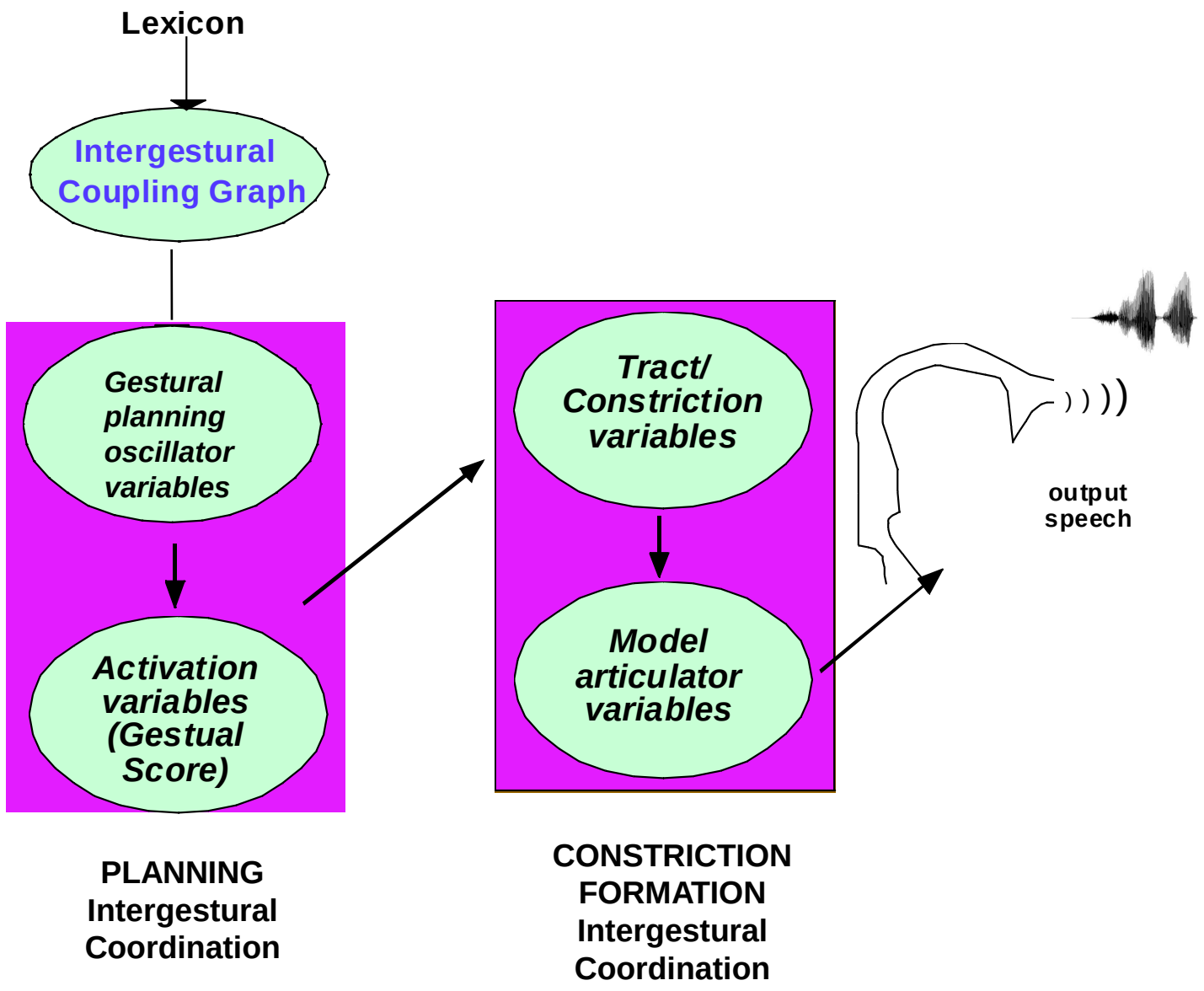
5 constriction degrees and 3 constriction locations

LA, TBCD, TTCD, VEL, GLO LP, TTCL, TBCL

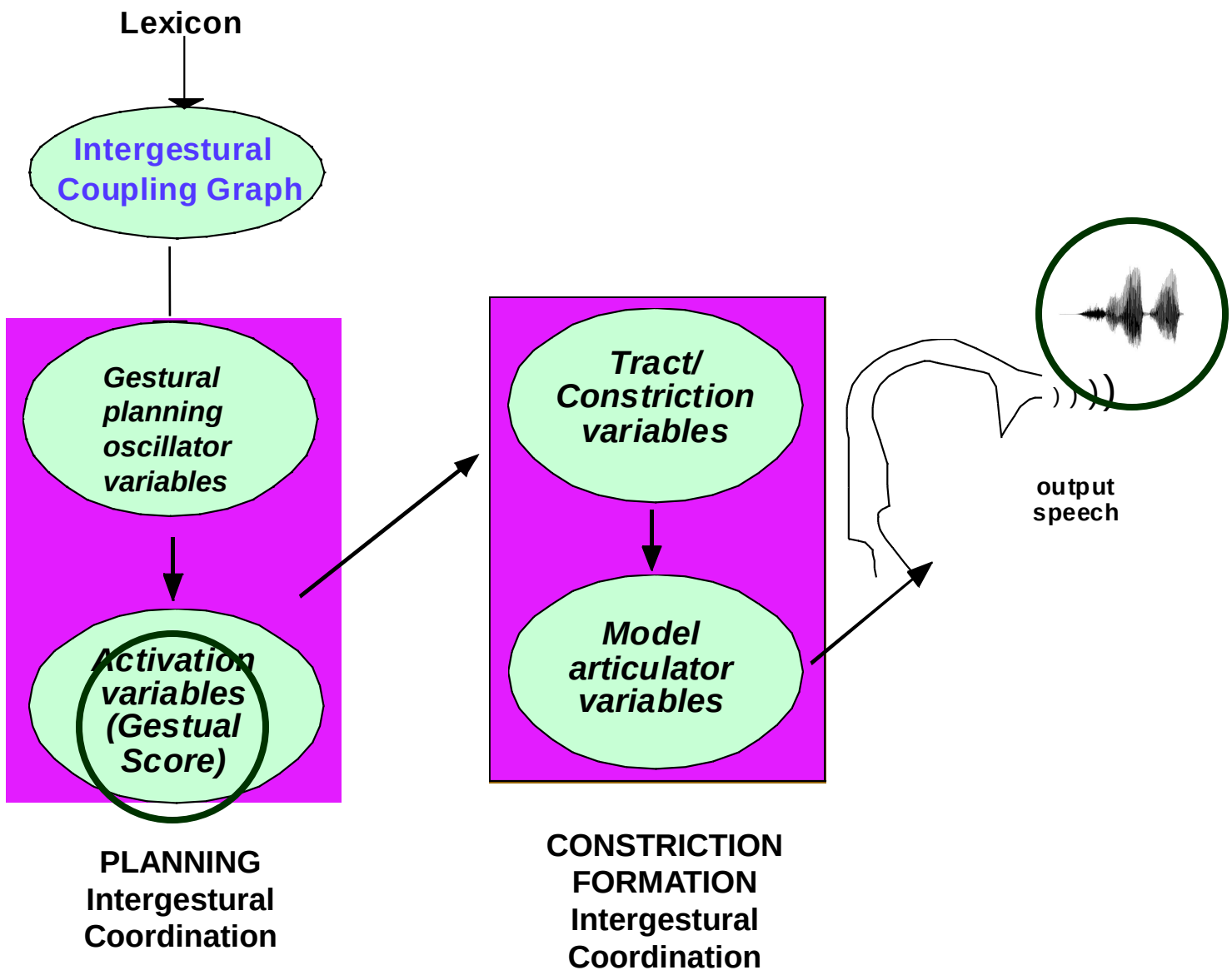
- According to articulatory phonology the plan of an utterance is formatted as a gestural score, which regulates the tract variable, which in turn shape the acoustic
- Speech synthesizer like TADA [1] is built on such principle

[1] Nam H., Goldstein L., Saltzman E., and Byrd D., "Tada: An enhanced, portable task dynamics model in matlab", J. Acoust. Soc. Am., vol 115, issue 5, pp 2430, 2004.

Task Dynamic Application (TADA)



Task Dynamic Application (TADA)



So, estimation of articulatory gesture pattern from speech

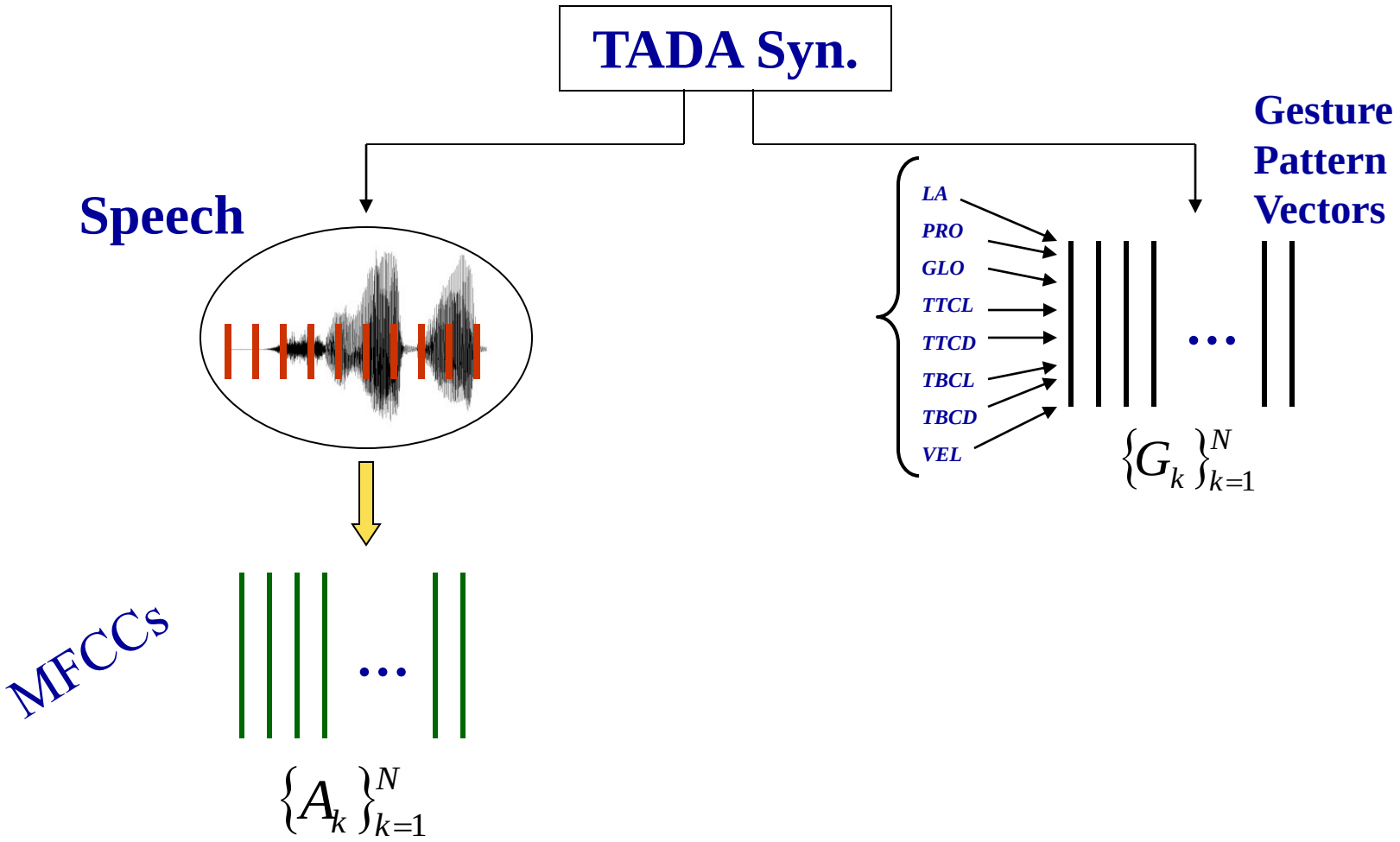
|||

estimation of the inverse of the function which maps the gesture pattern to the speech signal in TADA

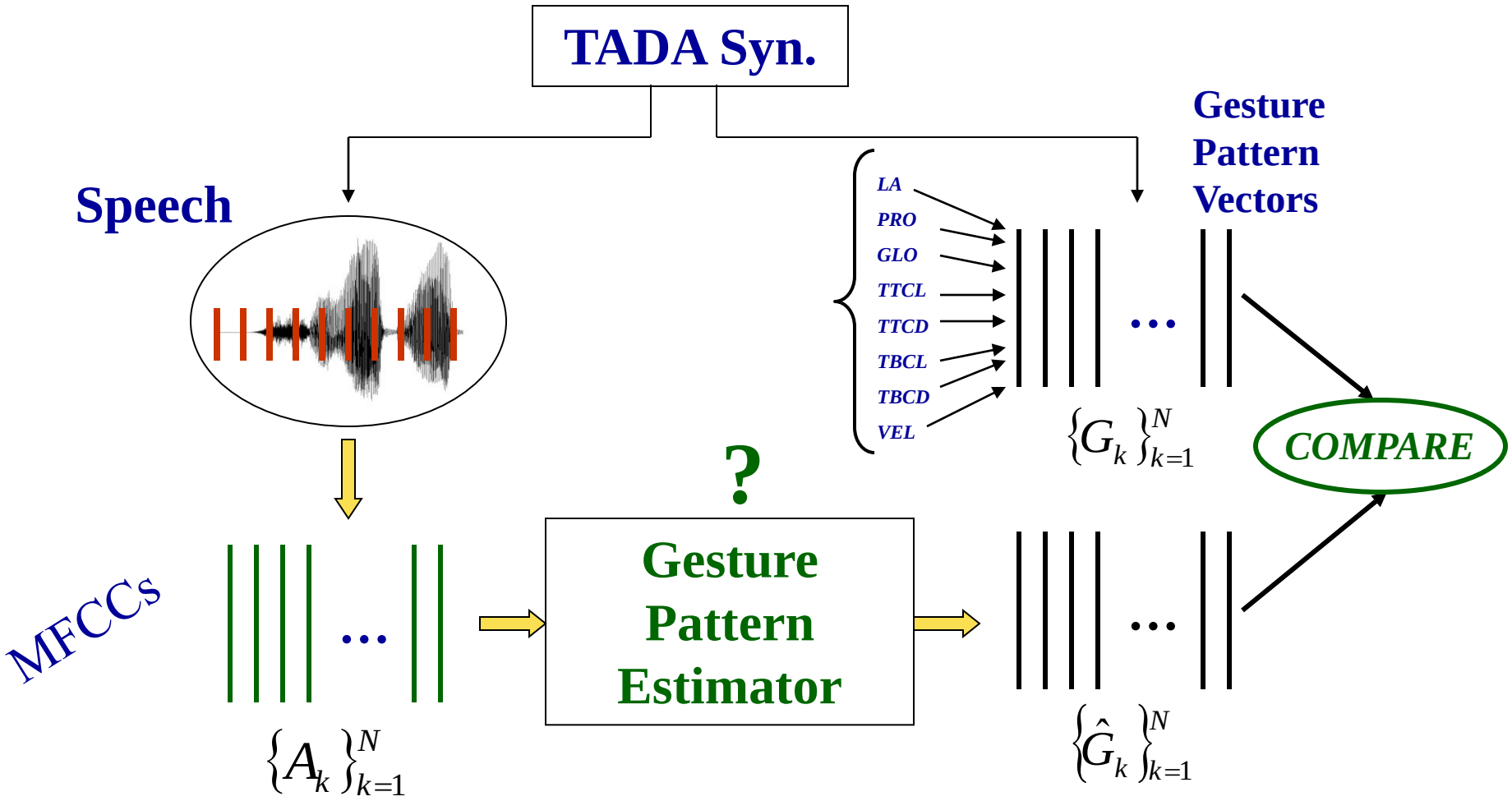
Motivations

- The gestural patterns constitute the (reduced dimensional) phonological *information* in speech.
- Gestures provide an invariant representation of the acoustic signal, which can be potentially useful to identify the spoken utterance; also useful for speech recognition, speech synthesis.
- Articulatory gestures provides a description of speech planning; hence could be useful for analyzing speech disorder or disfluency.

Problem Definition

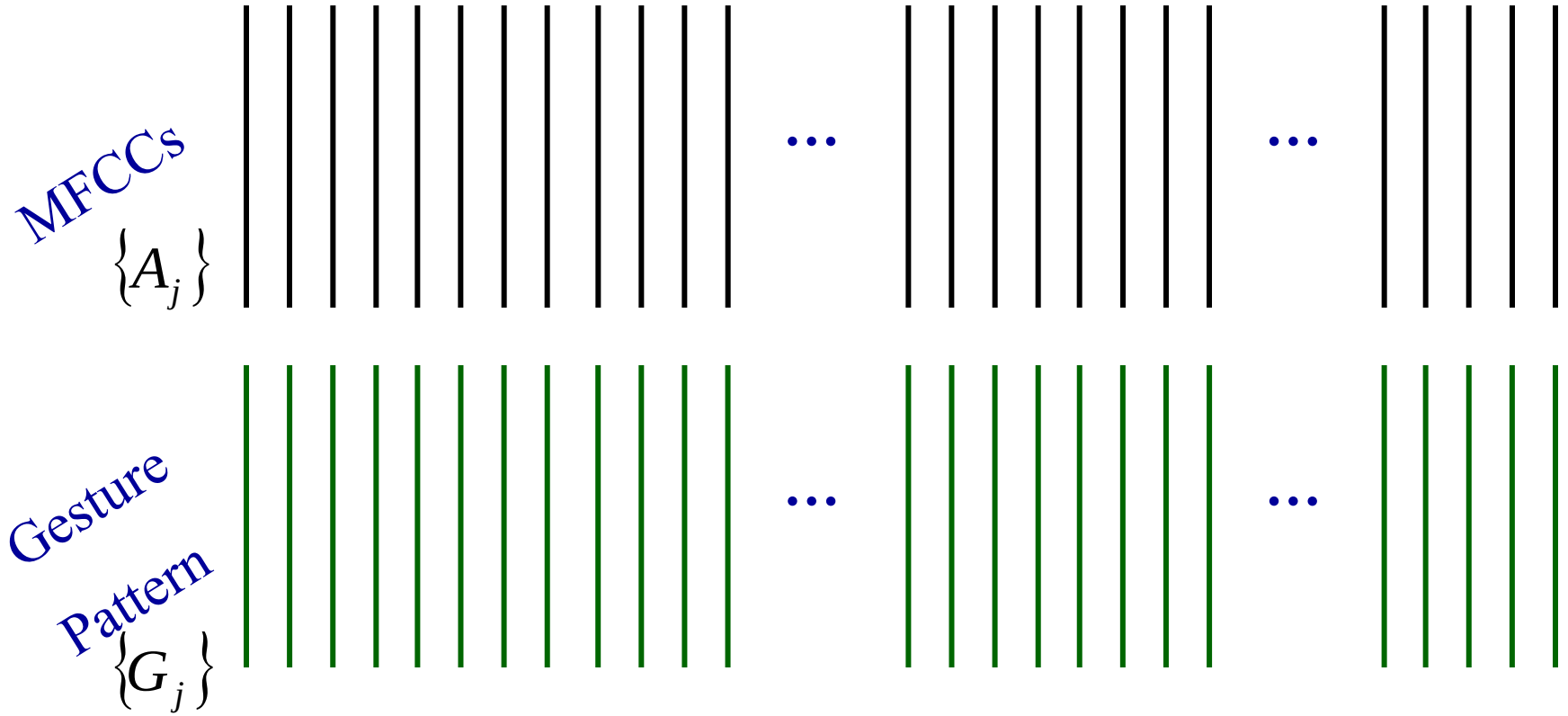


Problem Definition



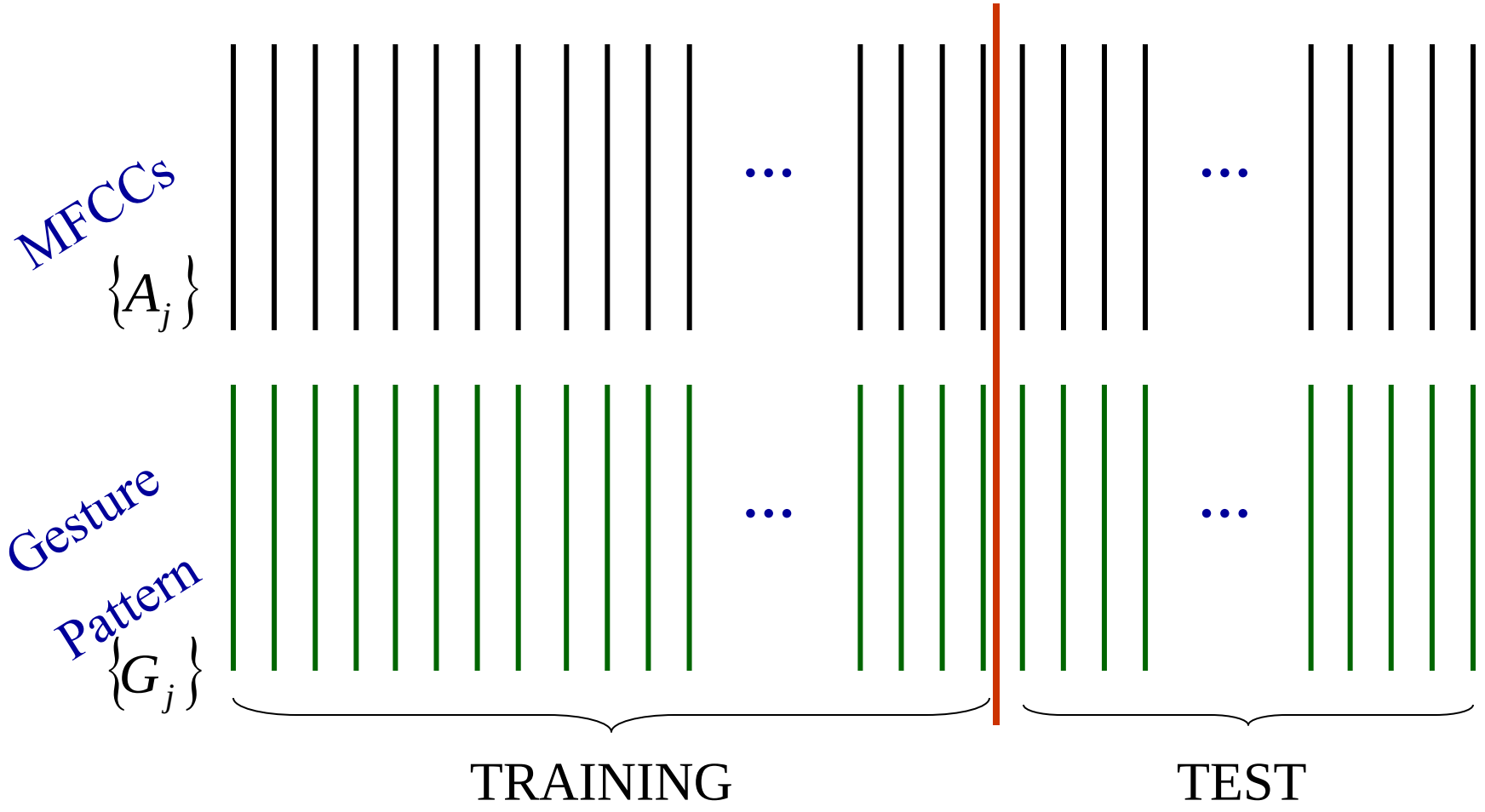
Proposed Approach

- The basic idea -- From TADA, we obtain parallel sequences of



Proposed Approach

- The basic idea -- From TADA, we obtain parallel sequences of



Learn the mapping statistically from <u>training</u> set and evaluate on <u>test</u> set

Proposed Approach

1. Dynamic programming (DP) approach *[2]*

$$\left\{\hat{G}_k\right\}_{k=1}^{N} = \arg\max_{\{G_k\}_{k=1}^{N}} \prod_{j=1}^{N} p\left(G_j \mid A_j\right) p\left(G_j \mid G_{j-1}\right)$$

$$= \frac{p\left(G_j, A_j\right)}{p\left(A_j\right)}$$

Joint Likelihood in Two Spaces (acoustic & gesture)

Smoothing Constraint

$$p\left(G_j \mid G_{j-1} = g\right) = N\left(g, \Lambda\right)$$

Λ - **diagonal covariance matrix, learnt from training data**

2. State model (SM) approach

[2] Lammert A., Ellis D. P. W., Divenyi P., "Data-driven articulatory inversion incorporating articulator priors", ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition, SAPA 2008, 21 September 2008, Brisbane, Australia.

Proposed Approach

1. Dynamic programming (DP) approach *[2]*

$$\left\{\hat{G}_k\right\}_{k=1}^{N} = \arg\max_{\{G_k\}_{k=1}^{N}} \prod_{j=1}^{N} p\left(G_j \mid A_j\right) p\left(G_j \mid G_{j-1}\right)$$

$$= \frac{p\left(G_j, A_j\right)}{p\left(A_j\right)}$$

Joint Likelihood in Two Spaces (acoustic & gesture)

Smoothing Constraint

$$p\left(G_j \mid G_{j-1} = g\right) = N\left(g, \Lambda\right)$$

Λ - **diagonal covariance matrix, learnt from training data**

*** The best sequence of G_k is obtained using DP search from a finite gesture pattern vectors, whose corresponding acoustic vectors are in the neighborhood of A_j**

2. State model (SM) approach

[2] Lammert A., Ellis D. P. W., Divenyi P., "Data-driven articulatory inversion incorporating articulator priors", ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition, SAPA 2008, 21 September 2008, Brisbane, Australia.

Proposed Approach

1. Dynamic programming (DP) approach *[2]*

$$\{\hat{G}_k\}_{k=1}^{N} = \arg\max_{\{G_k\}_{k=1}^{N}} \prod_{j=1}^{N} p(G_j \mid A_j)\, p(G_j \mid G_{j-1})$$

$$= \frac{p(G_j, A_j)}{p(A_j)}$$

Joint Likelihood in Two Spaces (acoustic & gesture)

Smoothing Constraint

$$p(G_j \mid G_{j-1} = g) = N(g, \Lambda)$$

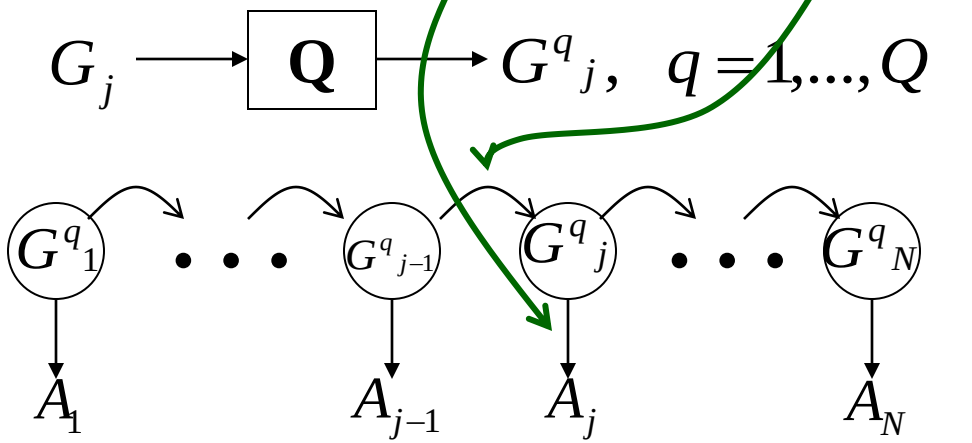
Λ - **diagonal covariance matrix, learnt from training data**

*** The best sequence of G_k is obtained using DP search from a finite gesture pattern vectors, whose corresponding acoustic vectors are in the neighborhood of A_j**

2. State model (SM) approach

$$\{\hat{G}_k\}_{k=1}^{N} = \arg\max_{\{G^q{}_k\}_{k=1}^{N}} \prod_{j=1}^{N} p(G^q{}_j \mid A_j)\, p(G^q{}_j \mid G^q{}_{j-1})$$

$$= \arg\max_{\{G^q{}_k\}_{k=1}^{N}} \prod_{j=1}^{N} p(A_j \mid G^q{}_j)\, p(G^q{}_j)\, p(G^q{}_j \mid G^q{}_{j-1})$$



[2] Lammert A., Ellis D. P. W., Divenyi P., "Data-driven articulatory inversion incorporating articulator priors", ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition, SAPA 2008, 21 September 2008, Brisbane, Australia.

Proposed Approach

1. Dynamic programming (DP) approach *[2]*

$$\{\hat{G}_k\}_{k=1}^{N} = \arg\max_{\{G_k\}_{k=1}^{N}} \prod_{j=1}^{N} p(G_j \mid A_j)\, p(G_j \mid G_{j-1})$$

$$= \frac{p(G_j, A_j)}{p(A_j)}$$

Joint Likelihood in Two Spaces (acoustic & gesture)

Smoothing Constraint

$$p(G_j \mid G_{j-1} = g) = N(g, \Lambda)$$

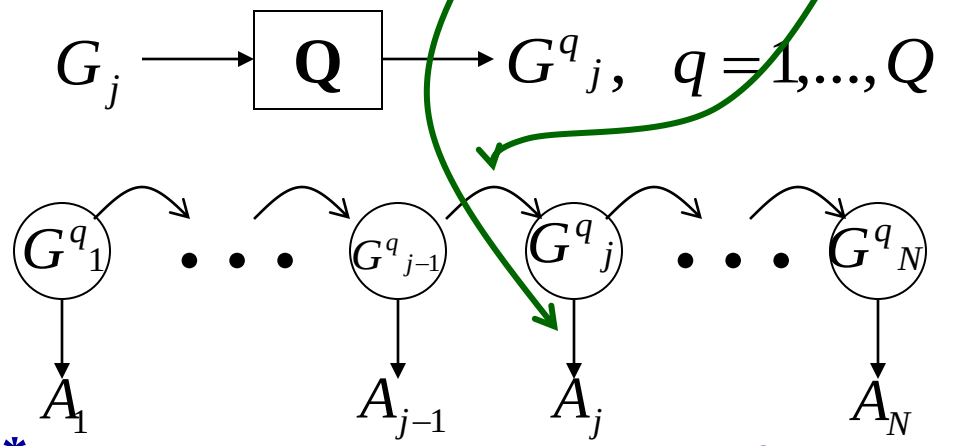
Λ - **diagonal covariance matrix, learnt from training data**

* **The best sequence of G_k is obtained using DP search from a finite gesture pattern vectors, whose corresponding acoustic vectors are in the neighborhood of A_j**

2. State model (SM) approach

$$\{\hat{G}_k\}_{k=1}^{N} = \arg\max_{\{G^q{}_k\}_{k=1}^{N}} \prod_{j=1}^{N} p(G^q{}_j \mid A_j)\, p(G^q{}_j \mid G^q{}_{j-1})$$

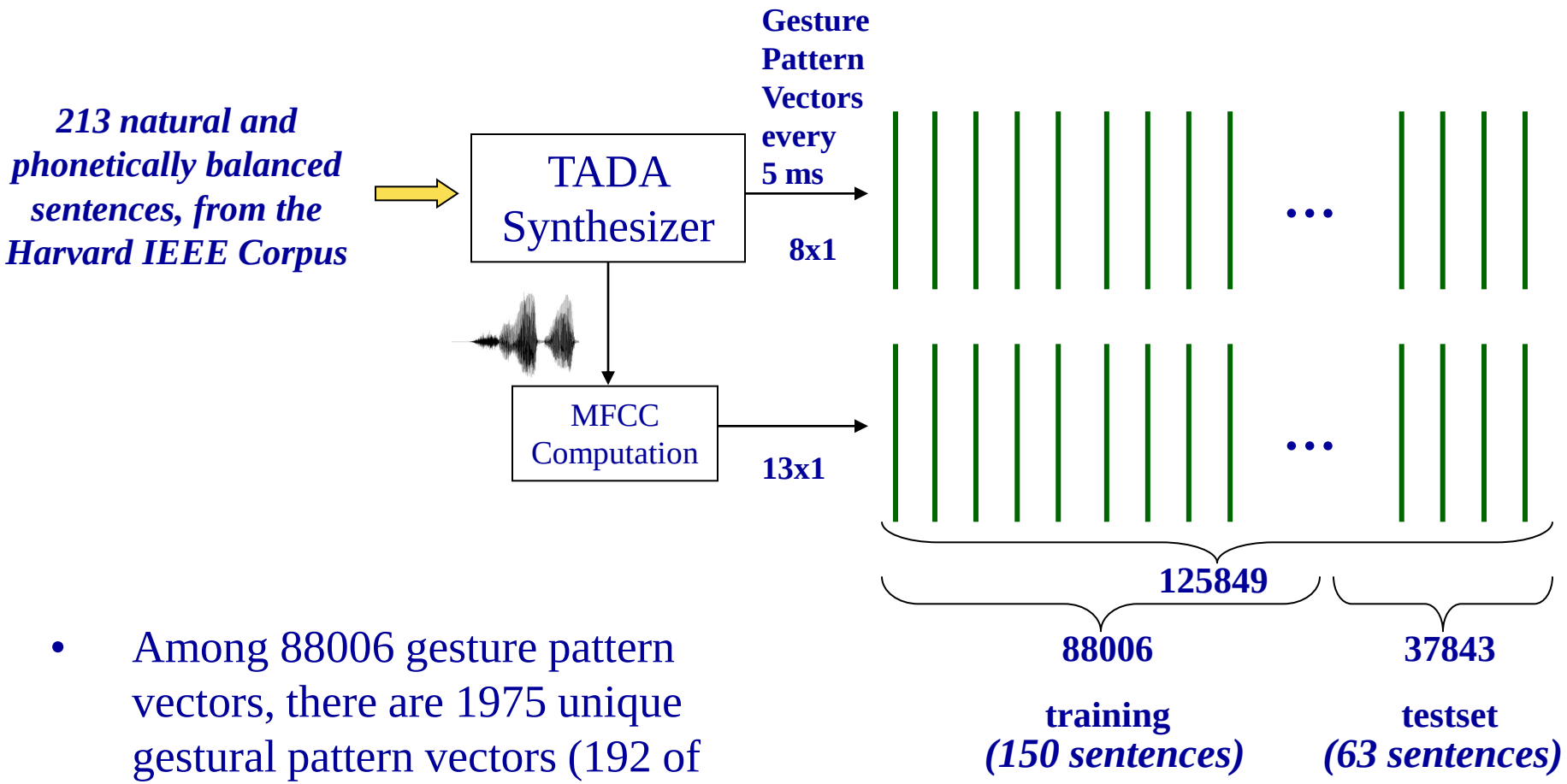
$$= \arg\max_{\{G^q{}_k\}_{k=1}^{N}} \prod_{j=1}^{N} p(A_j \mid G^q{}_j)\, p(G^q{}_j)\, p(G^q{}_j \mid G^q{}_{j-1})$$



* **The transition probability and probability of emission are learnt from training data. The Best sequence of gesture pattern vectors for a test utterance are obtained using an approach similar to viterbi-decoding**

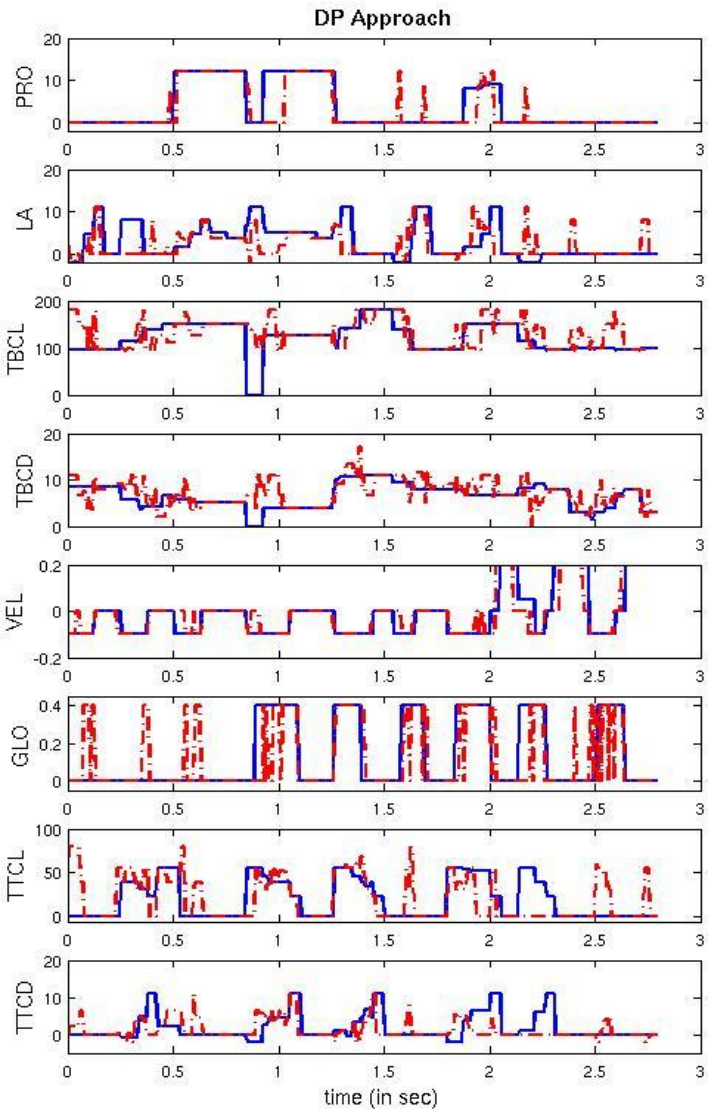
[2] Lammert A., Ellis D. P. W., Divenyi P., "Data-driven articulatory inversion incorporating articulator priors", ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition, SAPA 2008, 21 September 2008, Brisbane, Australia.

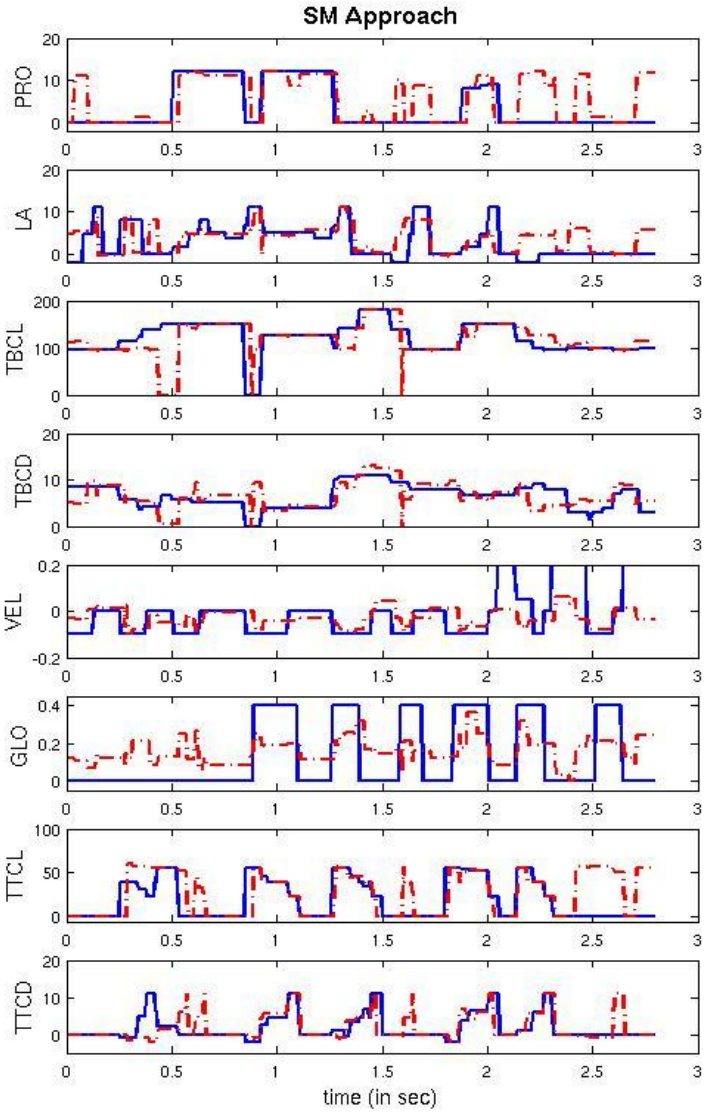
Dataset



- Among 88006 gesture pattern vectors, there are 1975 unique gestural pattern vectors (192 of them cover 70%)
- *Q* for SM chosen 150, 180, 210

Example of estimates

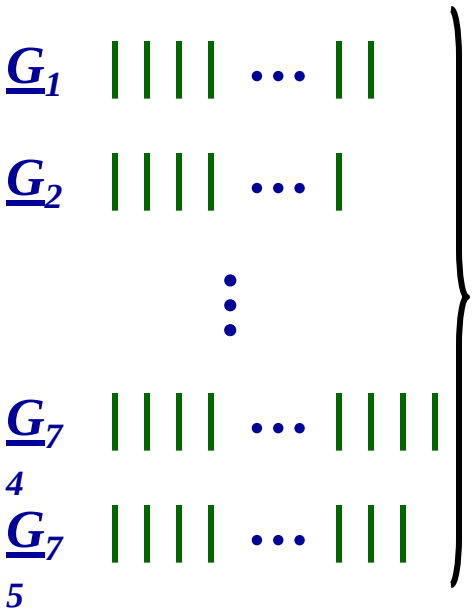




Evaluation

- **Word Identification**

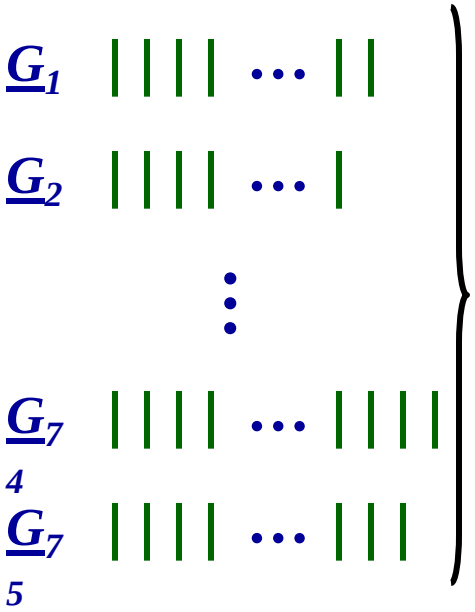
  75 words were randomly picked from the testset. Let $\underline{G_i}$ denote the sequence of gesture pattern vectors for *i*-th word.

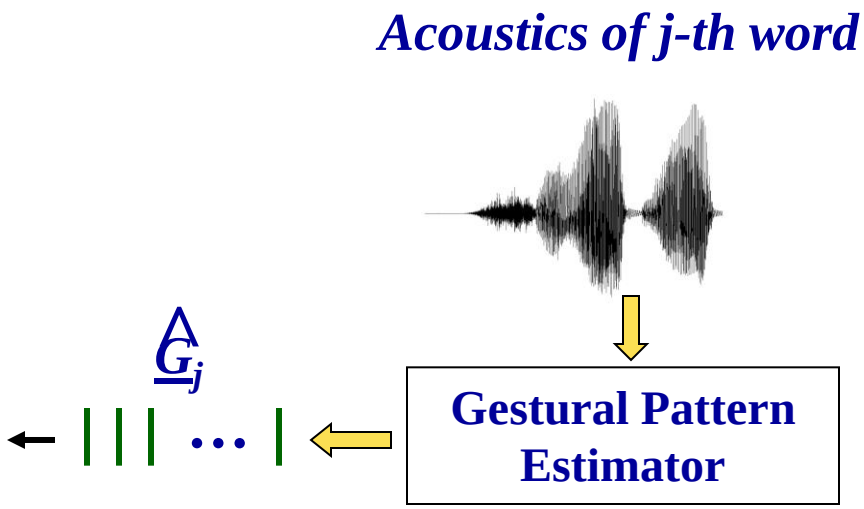


Evaluation

- **Word Identification**

  75 words were randomly picked from the testset. Let $\underline{G_i}$ denote the sequence of gesture pattern vectors for *i*-th word.

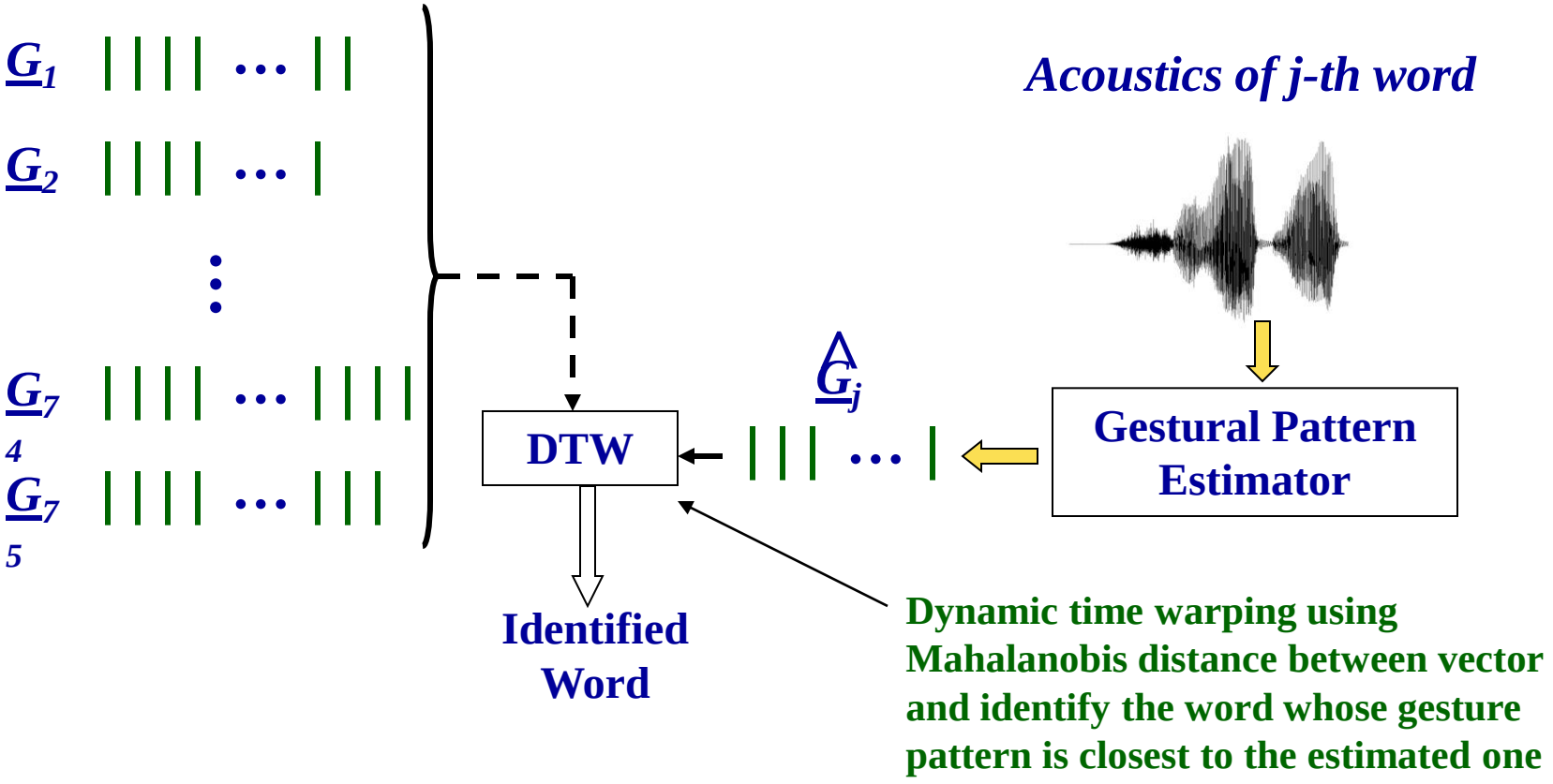




Evaluation

- **Word Identification**

  75 words were randomly picked from the testset. Let G_i denote the sequence of gesture pattern vectors for *i*-th word.



Word Identification Result

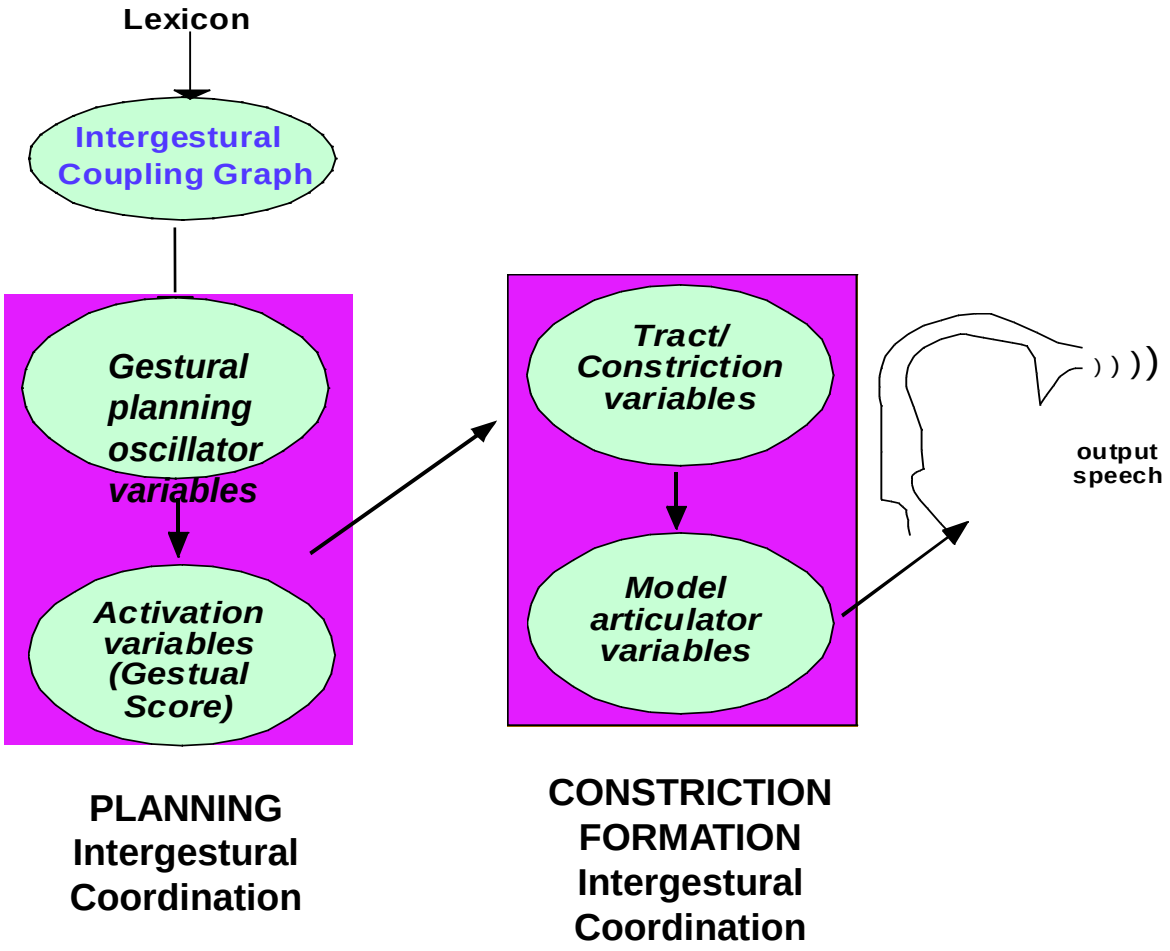
Schemes	Identification Accuracy
DP	66.67%
SM(Q=150)	52.00%
SM(Q=180)	61.33%
SM(Q=210)	60.00%

- It appears that the quantization error can diminish the performance of the SM approach.
- DP approach constrains the estimated sequence to be as smooth as possible which belies the quantal nature of gestural pattern; still it performs the best.

Discussion and Future Works

- Experimental result suggests that gestural recovery from acoustics using DP approach provides considerable information for identifying a word from a set of words.
- It will be interesting to investigate how accurate are the inter-gestural timings in the estimated gesture patterns.
- Investigation of the accuracy of the proposed estimator in the case of noisy acoustics.

Discussion and Future Works



- In addition to gestural patterns, tract variables or articulatory trajectory also provide description about speech articulation. It will be interesting to analyze which one can be estimated with greater accuracy from speech acoustic.

QUESTIONS?

COMMENTS?

SUGGESTIONS?