Articulatory Phonological Code for Speech Recognition

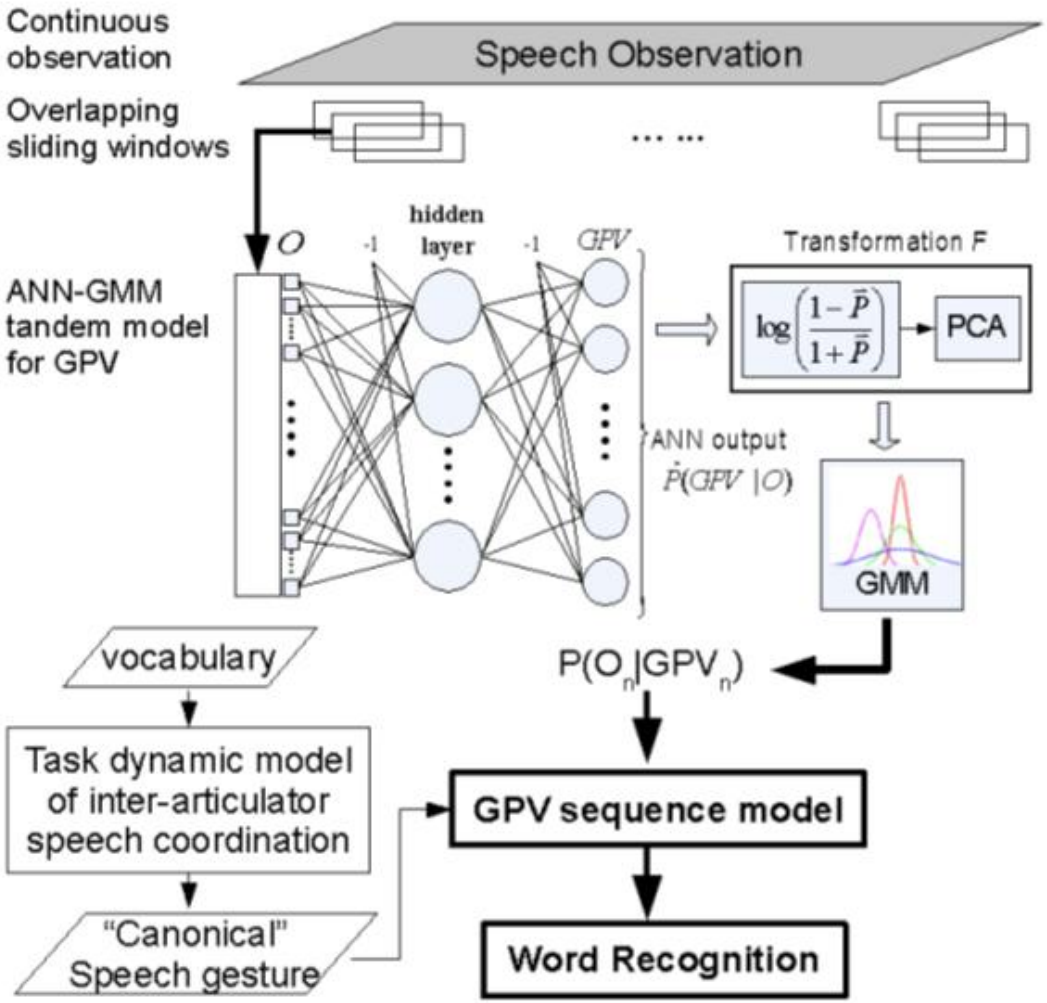
FSM-based Word Classification

FSM-based Gestural Score Variation

Chi Hu

OCT, 2009

GPV-based speech recognition



GPV-based speech recognition

❖ The proposed framework leverages speech gesture as the invariant representation of human speech.

❖To classify words, we leverage finite state machines that encode the plausible gestural scores for each vocabulary word.

❖Each GPV sequence is also weighted by the likelihood for all the recognized individual GPVs involved.

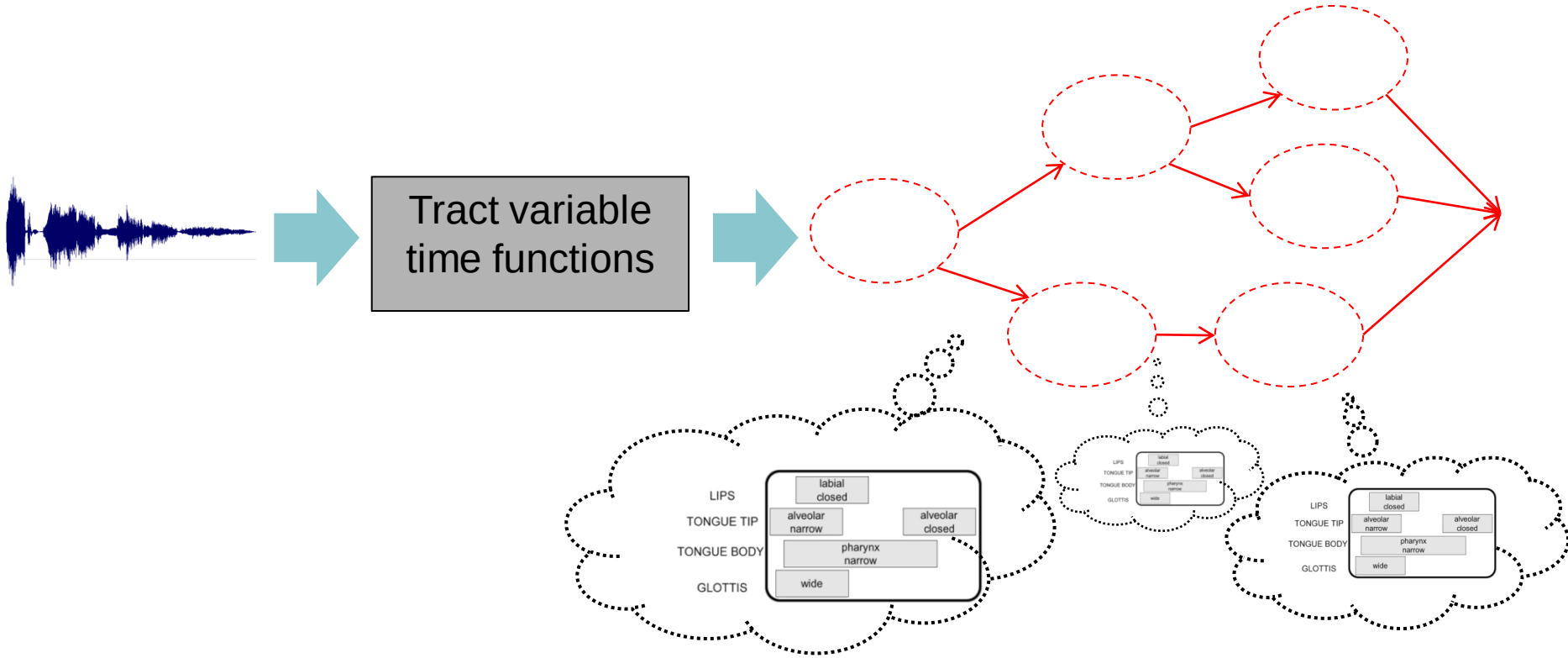
GPV-based speech recognition

- GPV lattice / FSM
  a compact representation of possible gestural scores, given an utterance

- FSM-based word classification
  finding scores for each dictionary entry

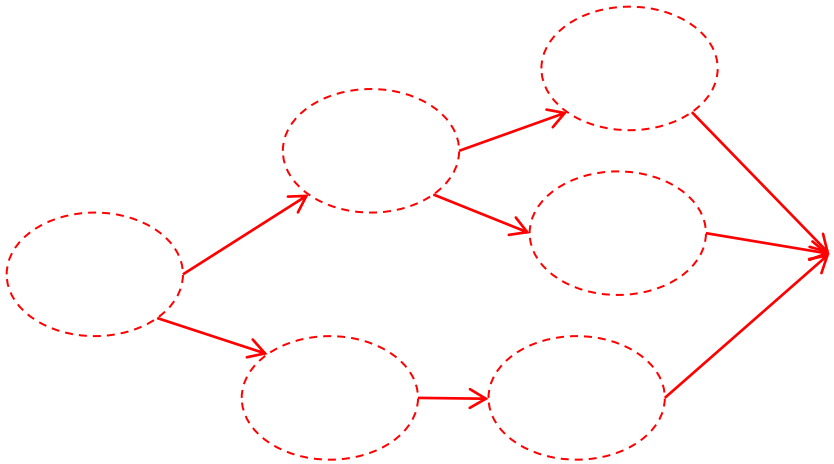
- FSM-based gestural score variation
  inducing plausible gestural score variation from the canonical gestural score

GPV lattice / finite state automata

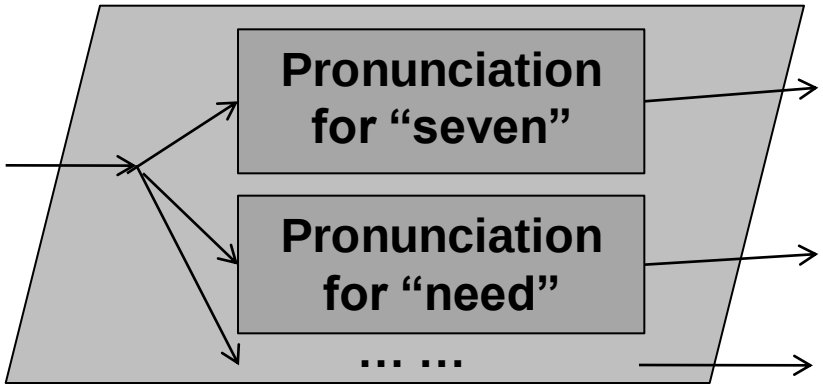


- a compact representation of possible gestural scores (GPV sequences), given an utterance

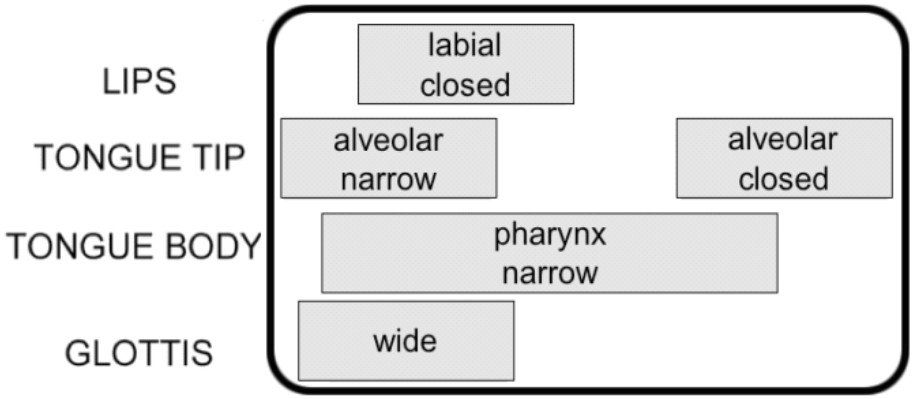
FSM-based word classification



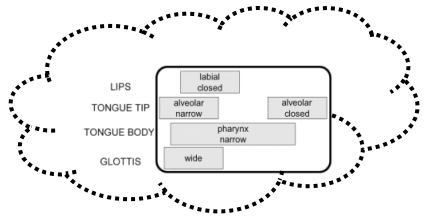
WORD	SCORE
• seven	3.2
• culmination	4.5
• need	1.3
• ingredients	2.1
• this	0.7
• dresses	1.5
• … …	

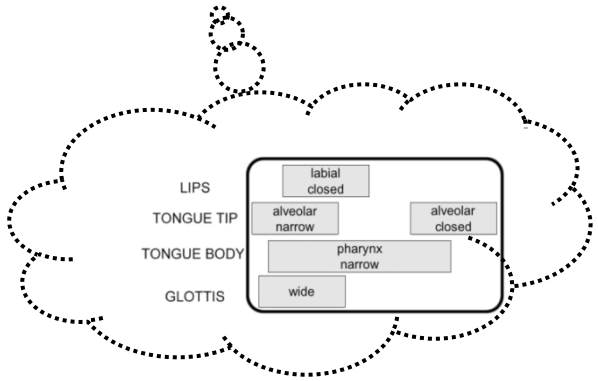


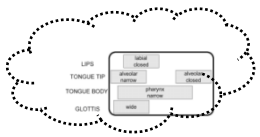
FSM-based gestural score variation



Canonical gestural score

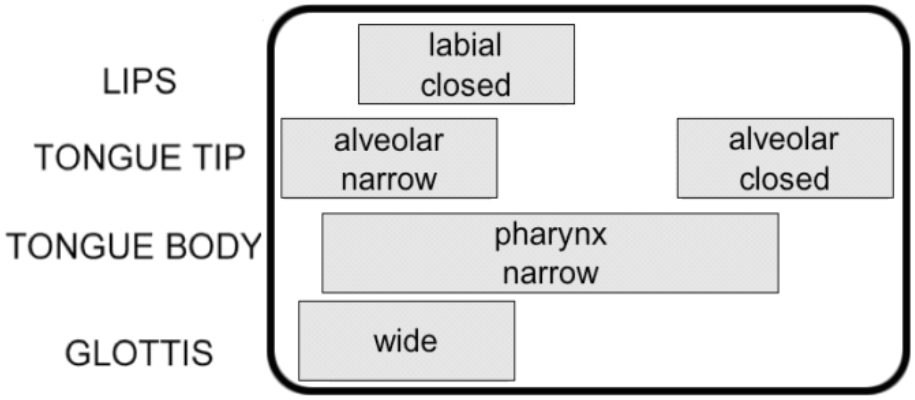






Plausible Variations Pronunciation for an utterance

FSM-based gestural score variation



How to factor gestural activations?

- Labial closed
- Alveolar
- Alveolar closed
- Pharynx
- Glottis wide

Ensemble of gestural activations (together with start and end times)

- The alternative gestural scores should have the same ensemble of gestural activations, but with possibly shifted start/end times.

FSM-based gestural score variation

- How to factor gestural activations

Dim	T_s	T_e	====against==== Target Value	Stiffness Value
3	1	61	2.61799387799149	355.305758439217
3	51	131	1.74532925199433	2526.61872667888
3	51	109	1.65806278939461	355.305758439217
3	115	131	1.74532925199433	2526.61872667888
4	1	61	0.0650000000000000	355.305758439217
4	51	145	-0.0200000000000000	2526.61872667888
4	51	109	0.115000000000000	631.654681669719
4	63	77	0.0600000000000000	2526.61872667888
4	115	131	0.100000000000000	2526.61872667888
6	51	145	-0.100000000000000	2526.61872667888
6	89	113	0.200000000000000	2526.61872667888
6	115	145	-0.100000000000000	2526.61872667888
6	127	145	-0.100000000000000	2526.61872667888
7	115	131	0.400000000000000	2526.61872667888
8	101	145	0.977384381116825	2526.61872667888
8	115	145	0.977384381116825	2526.61872667888
8	127	141	0.418879020478639	2526.61872667888
8	127	145	0.977384381116825	2526.61872667888
8	139	141	0.418879020478639	2526.61872667888
9	101	145	-0.0200000000000000	2526.61872667888
9	115	131	0.0100000000000000	2526.61872667888
9	127	141	0.110000000000000	2526.61872667888
9	127	145	-0.0200000000000000	2526.61872667888
9	139	141	0.110000000000000	2526.61872667888

Tract Variable Chart

Dim	TVs
1	LP
2	LA
3	TBCL
4	TBCD
6	VEL
7	GLO
8	TTCL
9	TTCD

Decoupling of TVs

- Observation
  - TBCL/TBCD, TTCL/TTCD ends at different time
    - E.g. 416 TADA words, 7457 activations, 146 decoupling TT gestures
  - Num(TBCD) > Num(TBCL)
    - 1213 Vs 1314

- Methods
  - Treat them as different activations to do the shifting
  - Bundle coupled activations and then do shifting in order to save computation complexity

FSM-based gestural score variation

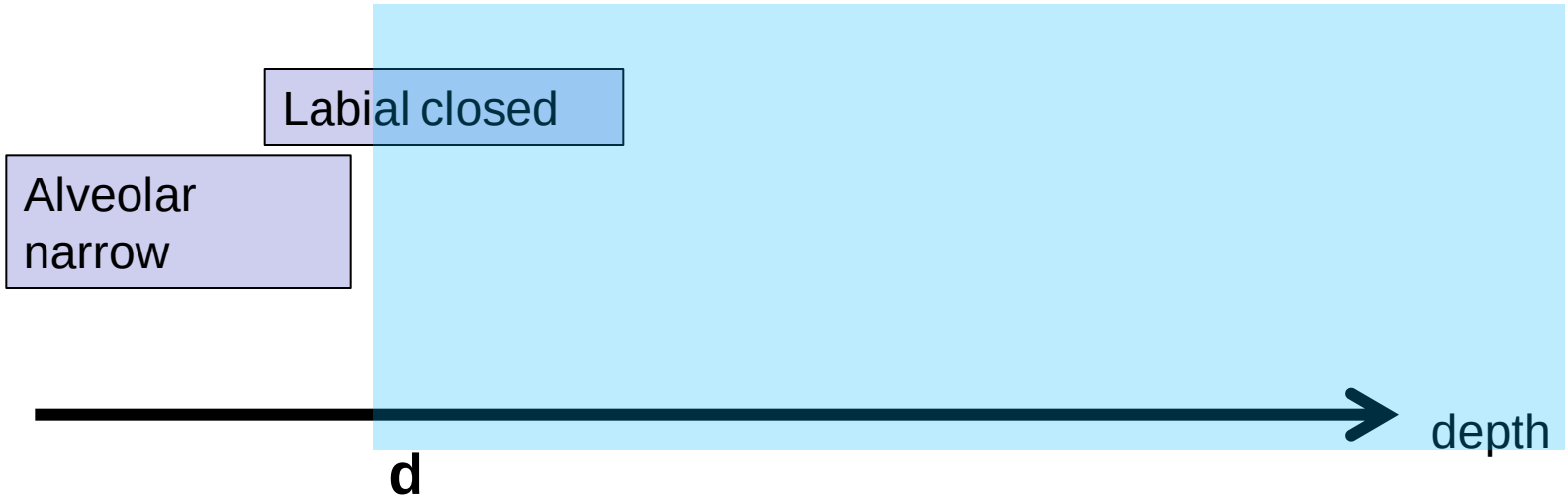
- Plausible gestural scores are "generated" according to changes of activation without actual time limitation
  - At each node
    - Allow only one change from the previous node
    - all possible combinations of instantaneous gestural activation targets/stiffnesses are proposed
    - a "plausibility" score is assigned to each combination
  - Given each one combination, move to the next node

- This should grow a 'tree'
  - each node defined by the "time" (depth of the tree) and gestural activations up to the current time
  - Prune the tree on the fly
  - one alternative gestural score at each leave

FSM-based gestural score variation

- Prior knowledge
  - Canonical gestural score
  - constraints
    - Unused activation should be started
    - Started activation should be ended
    - Ended activation should not be started again
    - Linguistic rules
- What to keep track of …
  - 'CurrentNode': decide the next action of growing according to previous node
  - '*Gended*': gestural activations that have ended
  - '*Gunused*': gestural activations that have not yet started
  - '*Gstarted*': gestural activations that have started but not yet ended
  - '*Slocal*': Local 'plausibility' costs up to current tree node

FSM-based gestural score variation



Gstarted

Labial closed

Gunused

Alveolar closed

Pharynx

Glottis wide

Gended

Alveolar narrow

Slocal

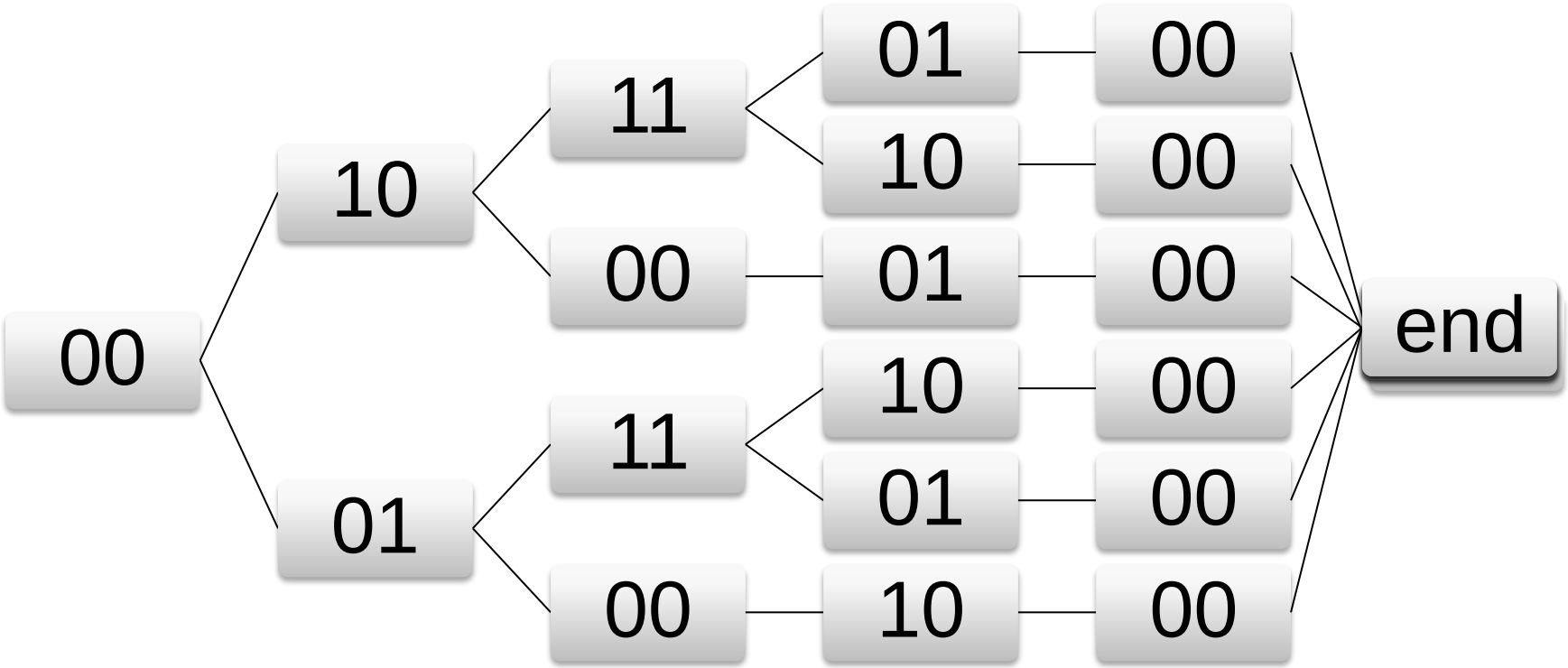
Recursive Function to Grow the Tree

- Function *FSMgrow(CurrentNode, Gstarted, Gunused, Gended, Slocal)*
  - Ending condition:
    - (isempty(Gunused) && isempty(Gstarted))
      - Print out this valid path
      - Get the corresponding gesture score
    - Elseif (*Gended* doesn't contain all activations)
      - Invalid path
    - Else
      - continue
  - Identify all instantaneous gestural activation combinations
    - End activations in *Gstarted*
    - Start activation in *Gunused*
    - One action per node

Recursive Function to Grow the Tree

- Function *FSMgrow(CurrentNode, Gstarted, Gunused, Gended, Slocal)*
  - For each combination
    - **provide the 'plausibility' cost**, update *Slocal*
    - Set a threshold C of *Slocal,* if >C, break
    - Update *CurrentNode,Gstarted,Gunused,Gended,Slocal*
    - **Call** *FnGrow(CurrentNode,Gstarted,Gunused,Gended, Slocal)*

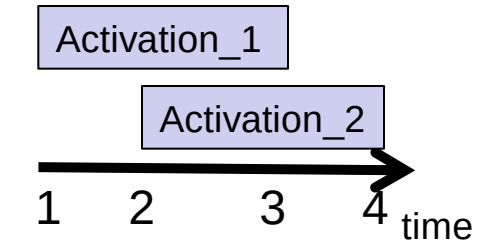
FSM-based gestural score variation



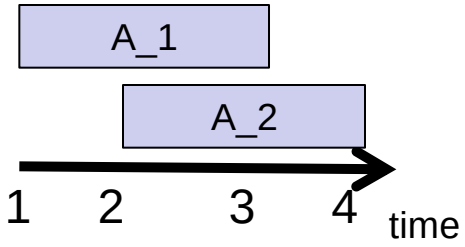
- Each path represents a possible gesture score of a given utterance
- Each node represents the combination of activations.
  E.g. 01 <=> activation 2

FSM-based gestural score variation

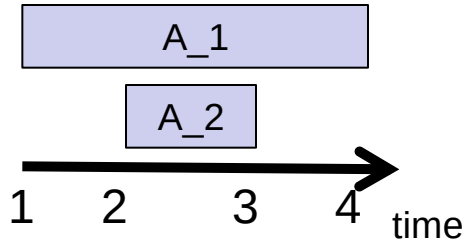
- **Example of two gestural activations in an utterance**



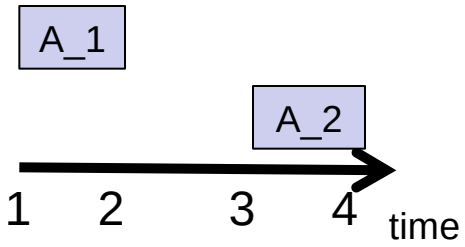
Canonical Gesture Score



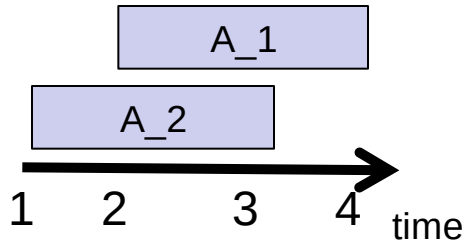
Gesture Score 1 (Path 1)



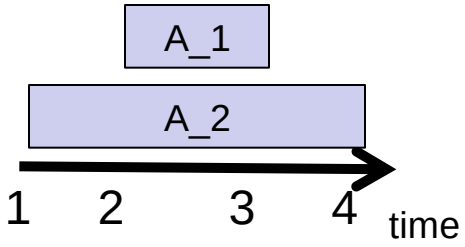
Gesture Score 2 (Path 2)



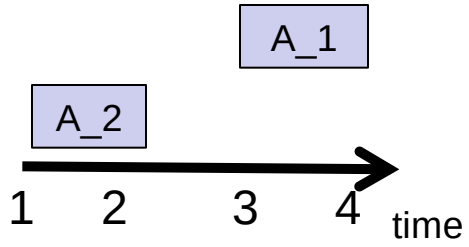
Gesture Score 3 (Path 3)



Gesture Score 4 (Path 4)

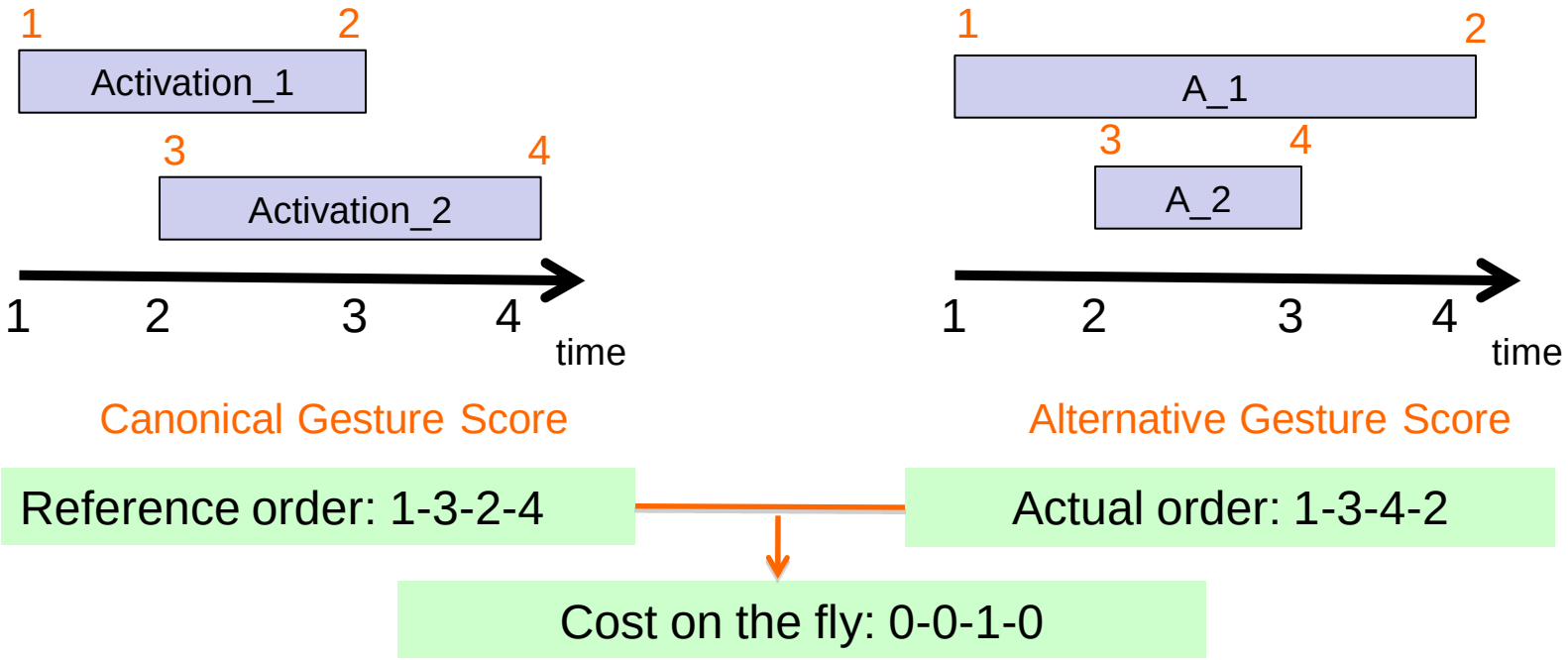


Gesture Score 5 (Path 5)



Gesture Score 6 (Path 6)

- **How to provide the 'plausibility' cost for the alternative gestural scores?**
  - Favor gestural scores that:
    - Similar to canonical gestural score
    - Assign an edge sequence for every activation
    - Edge crossing number get cost



Following Steps

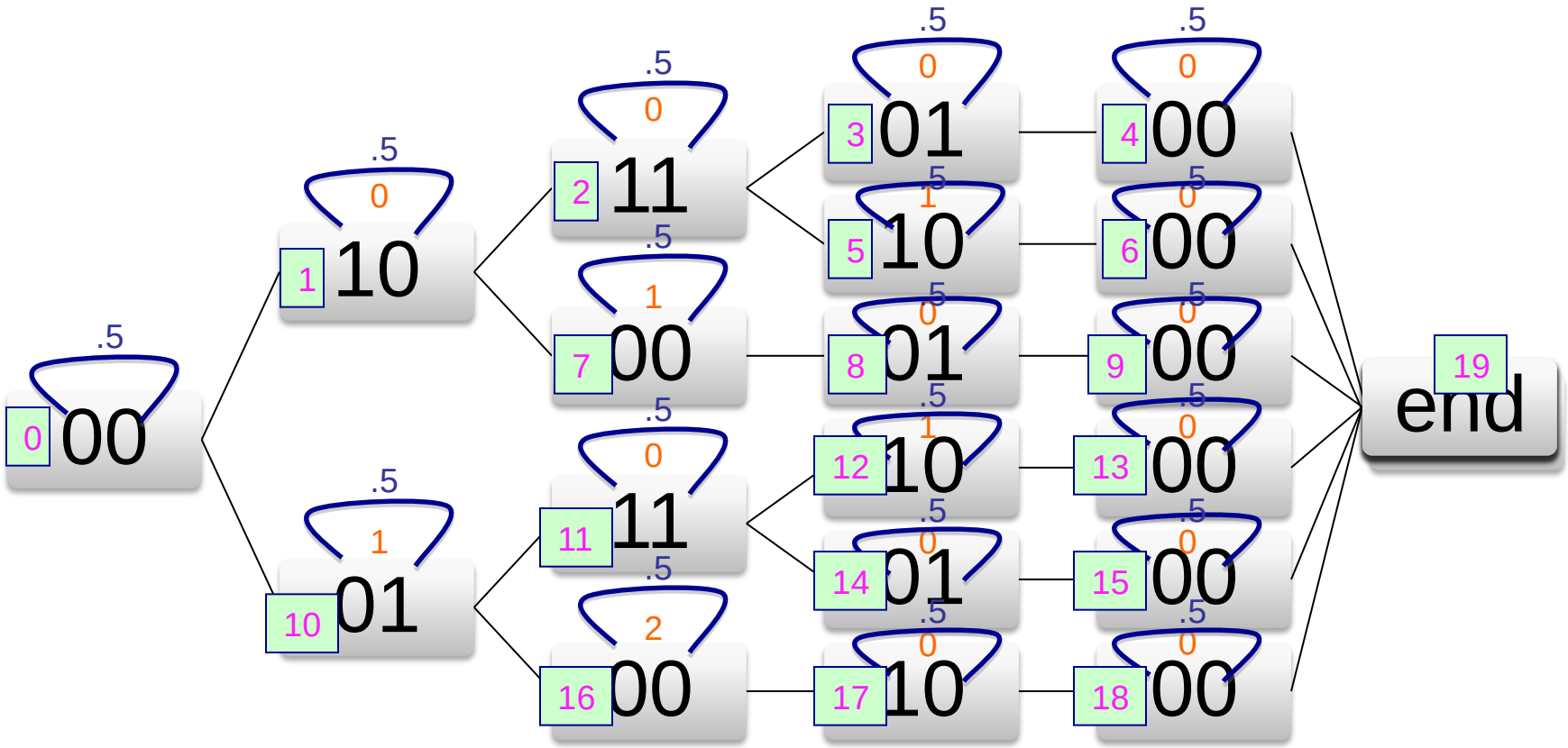
- Add on-the-fly costs on each node
- Add self-loop to each node
  - Add duration of each state
  - Assign them with uniform costs of value .5
- Discretizing gesture scores
  - Mapping each node with previous GPV types

    e.g.

00	Class 0
01	Class 3
10	Class 1
11	Class4

- Add state number on each node

Tree with Costs on Each Node



FSA Representation

Source State	Destination State	Arc Symbol Number	Arc Costs
0	0	class0	.5
0	1	class0	1
0	10	class0	0
1	1	class3	.5
1	2	class3	2
1	5	class3	0
2	2	class0	.5
2	3	class0	0
3	3	class1	.5
3	4	class1	0
4	4	class0	.5
4	19	class0	0
•••	•••	•••	•••
•••	•••	•••	•••
18	19	class0	0
19			

Improvements & Discussion

- Rules/constraints for shifting
  - Allow more changes at each node
  - Phonological rules according to C-V & C-C relations
- Cost function (Simko & Cummins, *09 InterSpeech*)

$$C = E + \omega_P P + \omega_D D \quad (1)$$

where E is a measure of articulatory effort, P is a measure of communicative efficacy, or parsing cost for the listener, and D is the overall utterance duration. The weights, ω_P and ω_D allow differential weighting of the cost components and are scaled so that the corresponding weight for the effort term, ω_E, has a value of one.

- Computational optimization

The End

~Thank you~