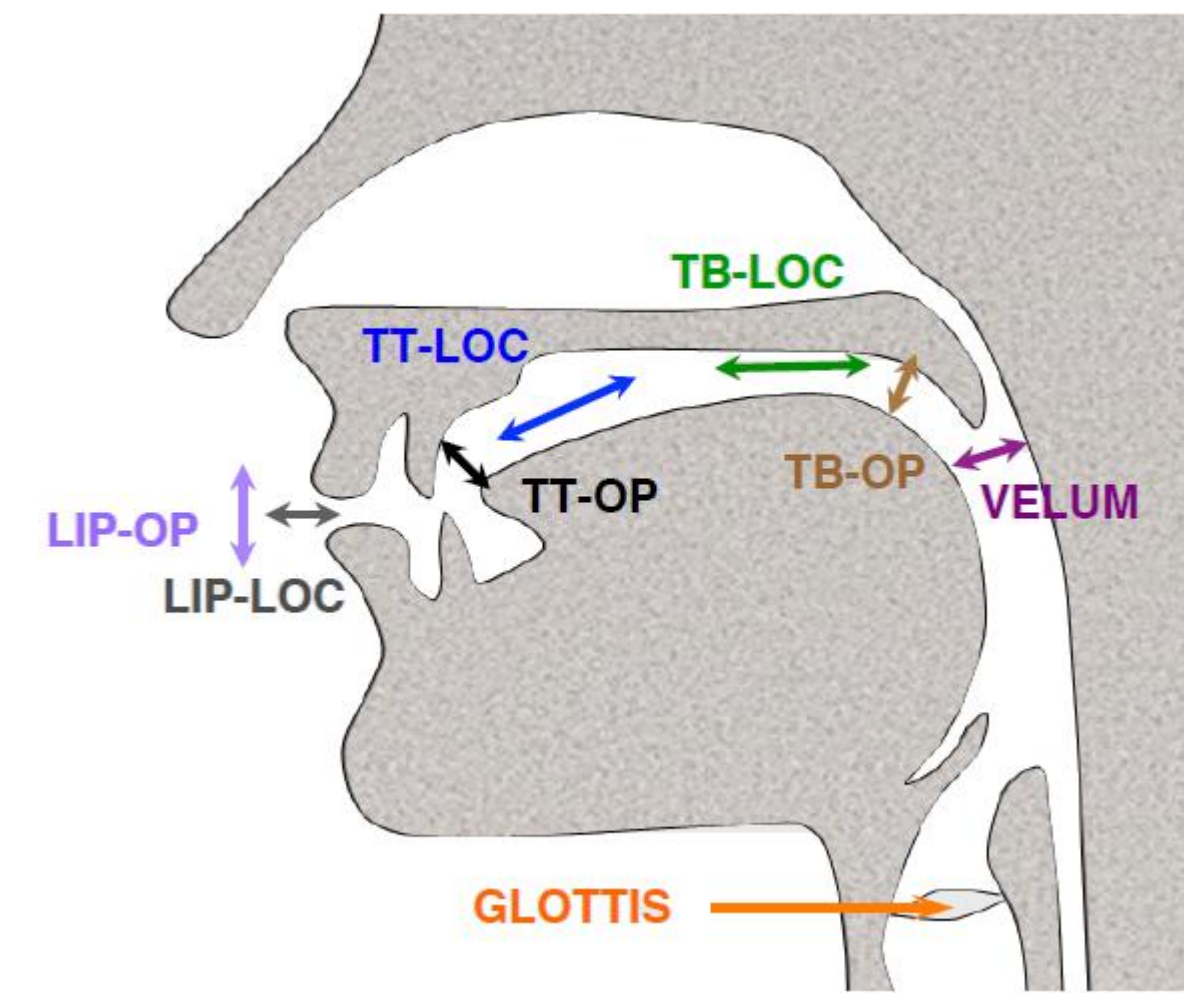


## 1. Introduction

- We consider factored models of the articulatory state space with explicit articulatory asynchrony modeling, applying these to the task of automatically generating feature transcriptions given word transcripts
- This task is motivated by the need for larger amounts of labeled data for ASR and linguistics research, which is expensive and hard to obtain
- We compare directed graphical models – dynamic Bayesian networks (DBNs) – based on previous work [2], and undirected graphical models – conditional random fields (CRFs) [3] – which are developed here
- The CRF-based models outperform the DBN-based models on the transcription task with relative improvements of 2.2%-10.0% in frame error rate

## 2. Articulatory Features

- The articulatory features used in this study are based on the tract variables of Articulatory Phonology [1]



Articulatory Feature	Values
LIP-LOC	protruded, labial, dental
LIP-OPEN	closed, critical, narrow, wide
TT-LOC	inter-dental, alveolar, palato-alveolar, retroflex
TT-OPEN	closed, critical, narrow, mid-narrow, mid, wide
TB-LOC	palatal, velar, uvular, pharyngeal
TB-OPEN	closed, critical, narrow, mid-narrow, mid, wide
VEL	closed, open
GLOT	closed, critical, wide

## 3. Articulatory Feature-based Pronunciation Modeling

- Following previous work [2], we represent word pronunciations in terms of articulatory feature targets by mapping from a phone-based dictionary
- The articulatory features may move asynchronously from one target to the next

Feature	closed (1)	closed (2)	open (3)	closed (4)
VEL	closed (1)	closed (2)	open (3)	closed (4)
TB	uvular/medium (1)	palatal/medium (2)	uvular/medium (3)	uvular/medium (4)
TT	alveolar/critical (1)	alveolar/medium (2)	alveolar/closed (3)	alveolar/critical (4)
LIPS	wide/labial (1)	wide/labial (2)	wide/labial (3)	wide/labial (4)
GLO	wide (1)	critical (2)	critical (3)	wide (4)
Phone	s	eh	n	s

Synchronous transitions correspond to the canonical pronunciation.

Asynchronous transitions can account for some variant pronunciations.

## 4. Articulatory Feature Forced Transcription Task

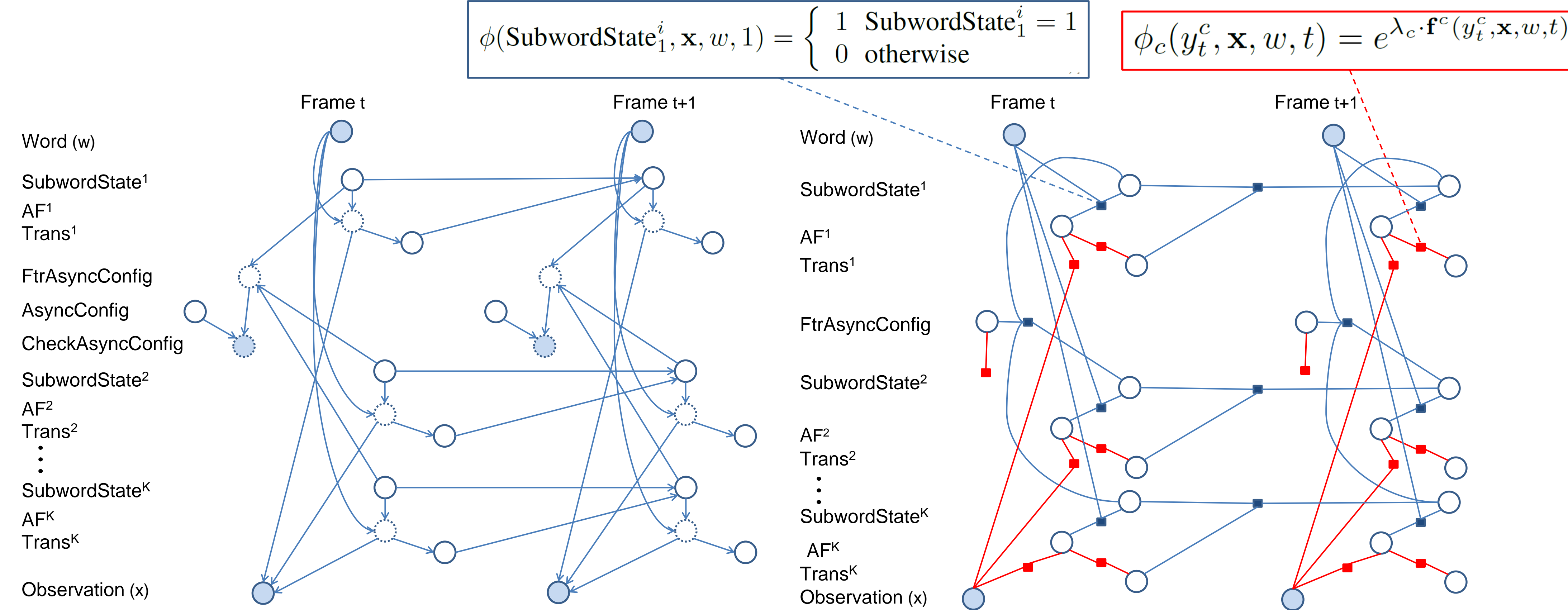
- Given a parameterization of the acoustics ( $\mathbf{x}$ ) for the word ( $w$ ), predict the most likely values for the articulatory features at each time frame

$$\mathbf{AF}^{1*}, \mathbf{AF}^{2*}, \dots, \mathbf{AF}^{K*} = \underset{\mathbf{AF}^1, \mathbf{AF}^2, \dots, \mathbf{AF}^K}{\operatorname{argmax}} p(\mathbf{AF}^1, \mathbf{AF}^2, \dots, \mathbf{AF}^K | w, \mathbf{x})$$

## References

- C. P. Browman and L. Goldstein, "Articulatory Phonology: An overview," *Phonetica*, vol. 49, pp.155-180, 1992.
- K. Livescu, "Feature-based pronunciation modeling for automatic speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2005.
- C. Sutton, K. Rohanimanesh, and A. McCallum, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," in Proc. ICML, 2004.
- S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," in Proc. ICSLP, 1996.

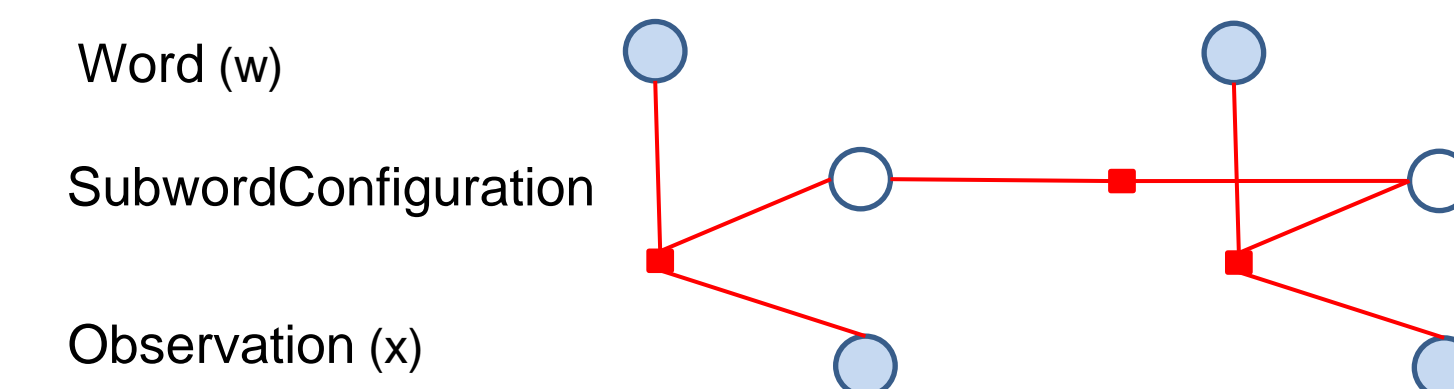
## 5. Graphical Models for Forced Transcription



$$p(\mathbf{v}) = \prod_t \prod_i p(v_t^i | \pi_{v_t^i})$$

$$p(\mathbf{y} | \mathbf{x}, w) = \frac{1}{Z(\mathbf{x}, w)} \prod_t \prod_{c \in \mathcal{C}} \phi_c(y_t^c, \mathbf{x}, w, t)$$

DBN (left) and CRF (right) models for articulatory feature forced transcription. Shaded nodes represent values that we condition on. Blue squares represent factors that enforce deterministic constraints in the CRF; red squares represent factors with associated learnable parameters.



Equivalent linear-chain CRF obtained by exploiting deterministic constraints. Sub-word state variables for the individual feature streams are combined into a single SubwordConfiguration variable.

Articulatory Stream	State Space Size	Example
Lips (L)	8	protruded/narrow
Tongue (T)	25	alveolar/closed/uvular/wide
Glottis/Velum (G)	4	wide/open

Details of the articulatory features used in the experiments.

## 6. CRF for Articulatory Feature Forced Transcription

- In addition to learnable parameters, both models include deterministic constraints specific to the problem domain
- Learnable parameters are associated with:
  - Transition probabilities for articulatory feature streams
  - Articulatory feature identity given word's pronunciation
  - Asynchronous state configurations
- Deterministic constraints encode:
  - Restrictions on maximum amount of asynchrony between articulatory feature streams
  - Constraints that ensure that all articulatory targets in the word's pronunciation are achieved (no substitution)
- Inference in original CRF model can be performed very efficiently in an equivalent linear-chain CRF model by exploiting deterministic constraints
  - Sub-word state variables for individual feature streams are combined into a single SubwordConfiguration variable
  - The linear-chain CRF retains non-deterministic (log-linear) factors; deterministic factors are used as constraints in the dynamic programming algorithm

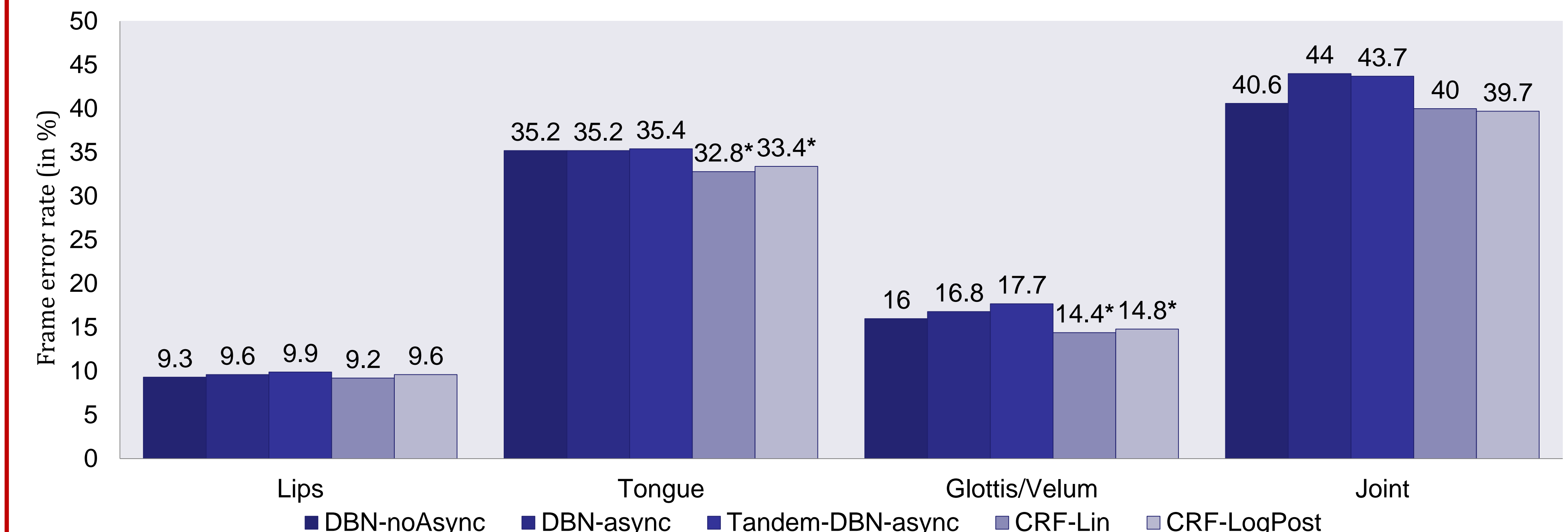
## 7. Experiments

- We evaluate the proposed method on a subset of the Switchboard Transcription Project (STP) data [4] – a subset of Switchboard manually transcribed at a detailed phonetic level
- We use the same train, development and test sets as in [2]
- Transcribed phones are mapped to articulatory features which serve as ground truth labels
- We assume that all tongue features are synchronized, the lip features are synchronized, and the glottis and velum are synchronized. Thus, we have three effective articulatory feature streams

Set	Number of words	Number of frames
Train	2941	89,748
Development	165	5,365
Test	236	7,037

## 8. Experimental Setup

- The output distributions in the DBN models are modeled as mixtures of Gaussians
- We consider both a baseline that allows no articulatory asynchrony (DBN-noasync) and one that allows up to 1 state of relative asynchrony (DBN-async)
- After training the asynchronous DBN system (DBN-async), it is used in forced-alignment mode to produce training labels for the CRF-based system
- The feature functions for the CRF are based on multilayer perceptrons (MLPs) trained to predict L, T, G configurations and phones
- We experiment with using either log-posteriors from the MLPs (CRF-LogPost) and linear outputs from the MLPs with the final softmax output layer removed (CRF-Lin)
- For comparison, we also consider a "tandem" asynchronous DBN system



Frame error rates (in %) of forced transcription using the various models. (\*) Indicates a statistically significant improvement over the DBN-noasync system.