A Segment-wise Hybrid Approach for Improved Quality Text-to-Speech Synthesis

MSc Research

Stas Tiomkin

Supervised by Professor David Malah

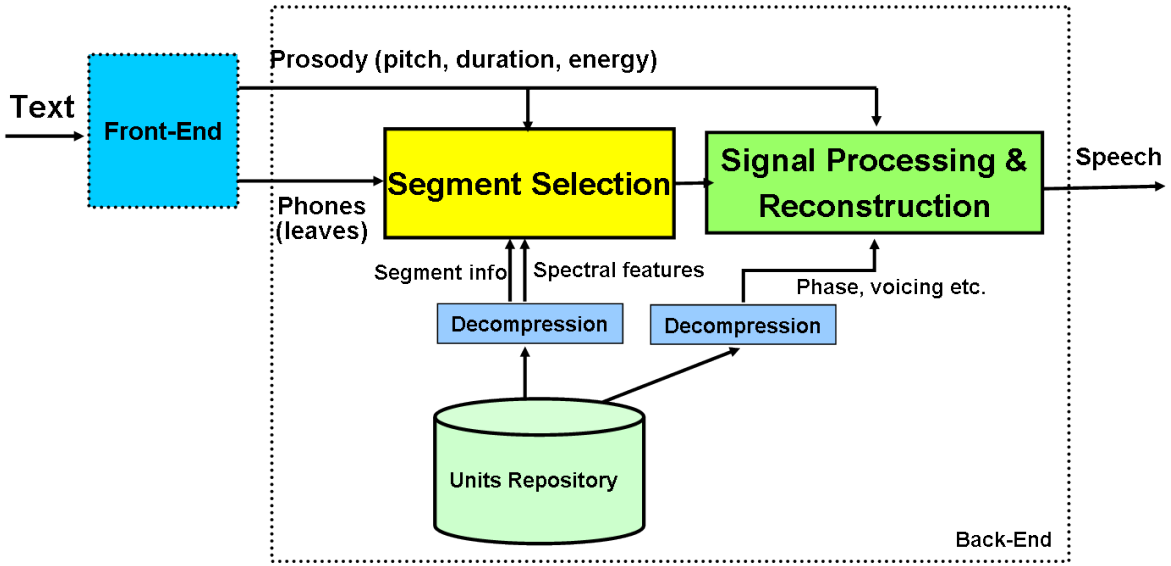
Dept. of Electrical Engineering Technion

In Collaboration with IBM - HRL

Outline

Outline

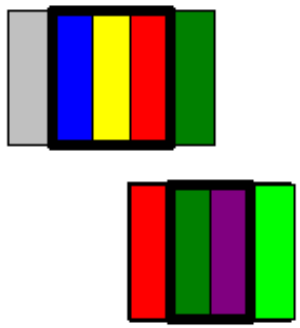
IBM's Concatenative Text-to-Speech System Overview



- "Natural" speech quality by using using natural speech units.
- Possible unpleasant audible discontinuities.

Concatenation Of Segments By Viterbi Search

- Cost function
  - Concatenative Cost
    - Spectral distance between adjacent segments



  - Prosody Target Cost
    - Difference between segments prosody to target prosody.

Statistical Text-to-Speech

- Modeling Speech Features
  - Hidden Markov Model
    - Independent Output Probabilities
- Modeling in Augmented Space
  - Combining Dynamic Features with Static Features
- Smooth transitions
  - Over-smoothing
    - Tackled in Future Work
- Lower footprint compared to CTTS

Statistical Text-to-Speech

- Modeling Speech Features
  - Hidden Markov Model
    - Independent Output Probabilities
- Modeling in Augmented Space
  - Combining Dynamic Features with Static Features
- Smooth transitions
  - Over-smoothing
    - Natural or Unnatural speech
- Lower footprint compared to CTTS

Statistical Text-to-Speech

- Modeling Speech Features
  - Hidden Markov Model
    - Independent Output Probabilities
- Modeling in Augmented Space
  - Combining Dynamic Features with Static Features
- Smooth transitions
  - Over-smoothing
- Lower footprint compared to CTTS

Statistical Text-to-Speech

- Modeling Speech Features
  - Hidden Markov Model
    - Independent Output Probabilities
- Modeling in Augmented Space
  - Combining Dynamic Features with Static Features
- Smooth transitions
  - Over-smoothing
- Lower footprint compared to CTTS

Statistical Text-to-Speech

- Modeling Speech Features
  - Hidden Markov Model
    - Independent Output Probabilities
- Modeling in Augmented Space
  - Combining Dynamic Features with Static Features
- Smooth transitions
  - Over-smoothing
    - Muffled and buzzy speech
- Lower footprint compared to CTTS

Statistical Text-to-Speech

- Modeling Speech Features
  - Hidden Markov Model
    - Independent Output Probabilities
- Modeling in Augmented Space
  - Combining Dynamic Features with Static Features
- Smooth transitions
  - Over-smoothing
    - Muffled and buzzy speech
- Lower footprint compared to CTTS

Statistical Text-to-Speech

- Modeling Speech Features
  - Hidden Markov Model
    - Independent Output Probabilities
- Modeling in Augmented Space
  - Combining Dynamic Features with Static Features
- Smooth transitions
  - Over-smoothing
    - Muffled and buzzy speech
- Lower footprint compared to CTTS

Model Features Representation

Static Features

- $\mathbf{c} = [\mathbf{c}_1^T, \mathbf{c}_2^T, \ldots, \mathbf{c}_N^T]_{dN\times 1}^T$
  - $\mathbf{c}_i = (c_i(1), c_i(2), \ldots, c_i(d))^T$

Dynamics Features

- $\Delta_i^1 = \frac{1}{2}(\mathbf{c}_{i+1} - \mathbf{c}_{i-1})$
- $\Delta_i^2 = -\mathbf{c}_{i-1} + 2\mathbf{c}_i - \mathbf{c}_{i+1}$

Augmented Space

- $\mathbf{o} = W\mathbf{c}$

$$W_i = \begin{pmatrix} \mathbf{0}_{d\times d} & \mathbf{1} & \mathbf{0} \\ -\frac{1}{2} & \mathbf{0} & \frac{1}{2} \\ -\mathbf{1} & \mathbf{2} & -\mathbf{1} \end{pmatrix}_{3d\times 3d}$$

Model Features Representation

Static Features

- $\mathbf{c} = [\mathbf{c}_1^T, \mathbf{c}_2^T, \ldots, \mathbf{c}_N^T]^T_{dN \times 1}$
  - $\mathbf{c}_i = (c_i(1), c_i(2), \ldots, c_i(d))^T$

Dynamics Features

- $\Delta_i^1 = \frac{1}{2}(\mathbf{c}_{i+1} - \mathbf{c}_{i-1})$
- $\Delta_i^2 = -\mathbf{c}_{i-1} + 2\mathbf{c}_i - \mathbf{c}_{i+1}$

Augmented Space

- $\mathbf{o} = W\mathbf{c}$

$$W_i = \begin{pmatrix} \mathbf{0}_{d\times d} & \mathbf{1} & \mathbf{0} \\ -\frac{1}{2} & \mathbf{0} & \frac{1}{2} \\ -\mathbf{1} & \mathbf{2} & -\mathbf{1} \end{pmatrix}_{3d\times 3d}$$

Model Features Representation

Static Features

- $\mathbf{c} = [\mathbf{c}_1^T, \mathbf{c}_2^T, \ldots, \mathbf{c}_N^T]^T_{dN\times 1}$
  - $\mathbf{c}_i = (c_i(1), c_i(2), \ldots, c_i(d))^T$

Dynamics Features

- $\Delta_i^1 = \frac{1}{2}(\mathbf{c}_{i+1} - \mathbf{c}_{i-1})$
- $\Delta_i^2 = -\mathbf{c}_{i-1} + 2\mathbf{c}_i - \mathbf{c}_{i+1}$

Augmented Space

- $\mathbf{o} = W\mathbf{c}$
  - $\mathbf{o} = [\mathbf{o}_1^T, \mathbf{o}_2^T, \ldots, \mathbf{o}_N^T]^T_{3dN\times 1}$
  - $\mathbf{o}_i = (\mathbf{c}_i, \Delta_i^1, \Delta_i^2)$

$$W_i = \begin{pmatrix} \mathbf{0}_{d\times d} & \mathbf{1} & \mathbf{0} \\ -\frac{\mathbf{1}}{\mathbf{2}} & \mathbf{0} & \frac{\mathbf{1}}{\mathbf{2}} \\ -\mathbf{1} & \mathbf{2} & -\mathbf{1} \end{pmatrix}_{3d\times 3d}$$

Model Description

- HMMs model phonemes in **o** - space
  - 3 states per phoneme
- Probability Density Function - GMM
  - $\mathbf{o} \sim \sum_{i=1}^{K} \omega_i \mathcal{N}(\mathbf{o}; \mathbf{m}_i, \mathbf{U}_i)$
    - $\mathbf{m}_i = [\mathbf{m}^c, \mathbf{m}^{\Delta^1}, \mathbf{m}^{\Delta^2}]_{3d \times 1}$
    - $\mathbf{U}_i = diag[\mathbf{U}^c, \mathbf{U}^{\Delta^1}, \mathbf{U}^{\Delta^2}]_{3d \times 3d}$
- Using one Gaussian per state
  - $P(\mathbf{o}) = \frac{1}{\sqrt{(2\pi)^{3dN}|U|}} e^{-\frac{1}{2}(\mathbf{o}-\mathbf{m})^T \mathbf{U}^{-1}(\mathbf{o}-\mathbf{m})}$

Statistical Utterance Composition

- Utterance Model

$$\begin{aligned} \mathbf{m} &= [\mathbf{m}_1, \mathbf{m}_1, \mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_2, \ldots, \mathbf{m}_K, \mathbf{m}_K]_{3dN\times 1} \\ \mathbf{U} &= diag[\mathbf{U}_1, \mathbf{U}_1, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_2, \ldots, \mathbf{U}_K, \mathbf{U}_K]_{3dN\times 3dN} \end{aligned}$$

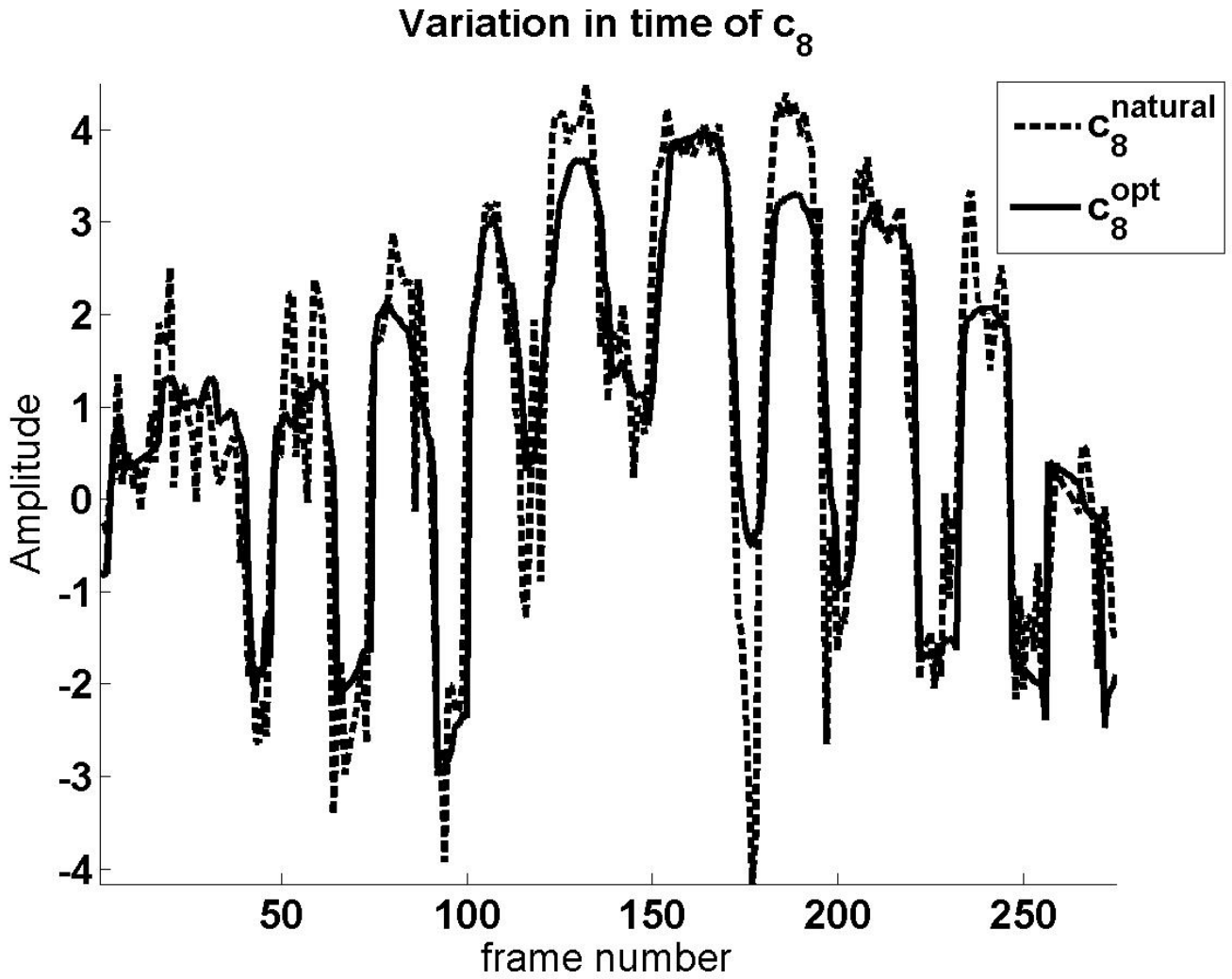
- An optimal solution determination
  - Cost function:

$$\begin{aligned} J(\mathbf{o})|_{\mathbf{o}=\mathbf{Wc}} &= -ln(P(\mathbf{o}))|_{\mathbf{o}=\mathbf{Wc}} \\ &= \tfrac{1}{2}(\mathbf{Wc}-\mathbf{m})^T\mathbf{U}^{-1}(\mathbf{Wc}-\mathbf{m}) + K \\ &= \tfrac{1}{2}\|\mathbf{U}^{-\frac{1}{2}}(\mathbf{Wc}-\mathbf{m})\|_2^2 + K \end{aligned}$$

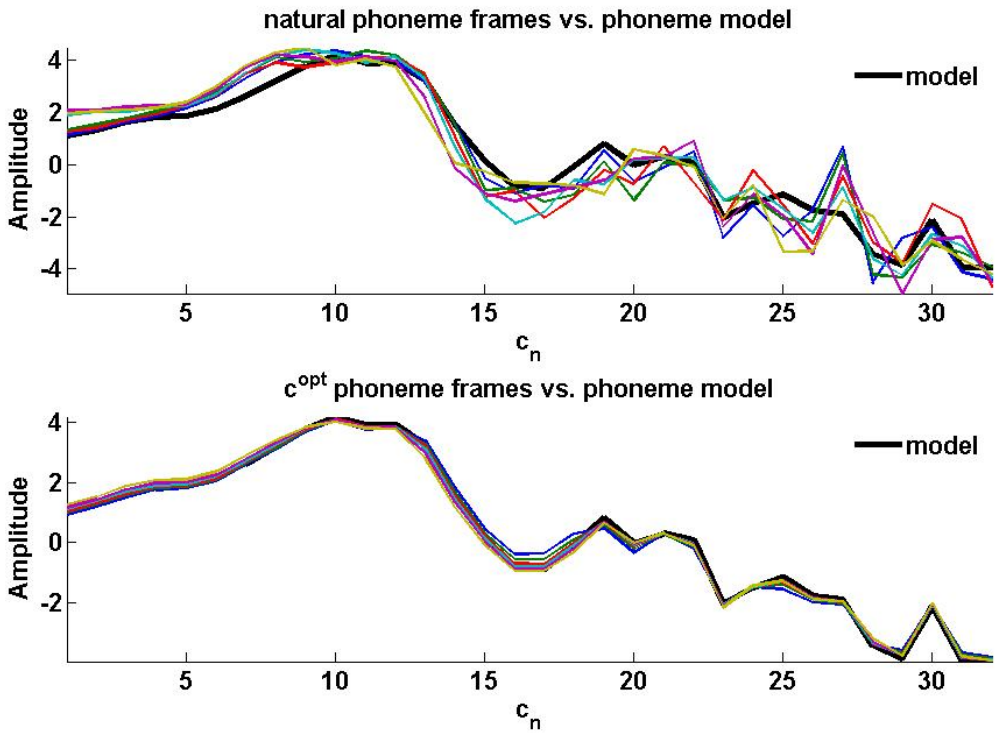
  - Optimal Speech Feature Vector:

$$\frac{\partial J(\mathbf{Wc})}{\partial c} = 0, \Rightarrow \mathbf{c}^{opt} = (\mathbf{W^TU^{-1}W})^{-1}\mathbf{W^TU^{-1}m}$$

Optimal Solution Features



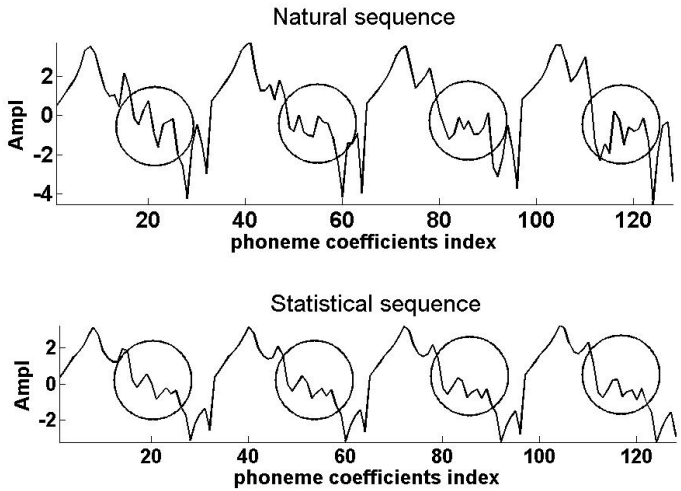
Optimal Solution Features



"Answer the following question as carefully and completely as possible." **ctts stts**

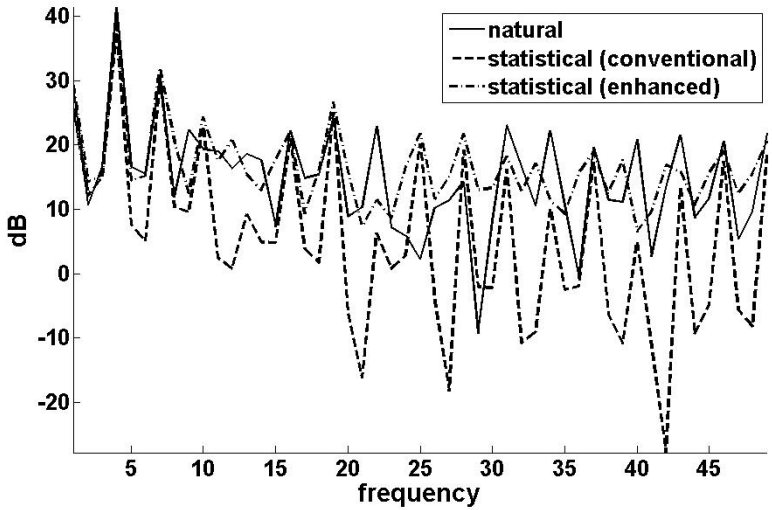
Outline

Insufficient Dynamics - Model Mean Replication



- Consider segments, $d \times T_i$, as vectors, $1 \times dT_i$.
- Transform vectors, $1 \times dT_i$, by the FFT of length $1 \times dT_i$.
- Examine *Non-Harmonic Components*.
  - In training, learn non-harmonic component statistics.
  - In synthesis, match non-harmonic component to their statistics.

Enhance the Non-Harmonic Components



- Proposed synthesis approach
  - Improve *intra-phoneme frames* in the transform domain.
  - Connect smoothly *inter-phoneme frames* by $\Delta^{1,2}$.
  - Achieved by new arrangements for W, M, U.
  - The generated speech quality is improved.

Conventional Augmented Space Construction

Frame Wise, 'fw', $\mathbf{o} = W\mathbf{c}$, $J^{fw}(\mathbf{Wc}) = \frac{1}{2}\|\mathbf{U}^{-\frac{1}{2}}(\mathbf{Wc} - \mathbf{m})\|_2^2$

$$\underbrace{\begin{pmatrix} \mathbf{0}_{d\times d} & \mathbf{1} & \mathbf{0} & \vdots \\ -\frac{\mathbf{1}}{\mathbf{2}} & \mathbf{0} & \frac{\mathbf{1}}{\mathbf{2}} & \vdots \\ -\mathbf{1} & \mathbf{2} & -\mathbf{1} & \vdots \\ \vdots & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \vdots & -\frac{\mathbf{1}}{\mathbf{2}} & \mathbf{0} & \frac{\mathbf{1}}{\mathbf{2}} \\ \vdots & -\mathbf{1} & \mathbf{2} & -\mathbf{1} \end{pmatrix}}_{\mathbf{W}_{3dN\times dN}} \underbrace{\begin{pmatrix} \vdots \\ \vdots \\ \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \\ \vdots \\ \vdots \end{pmatrix}}_{\mathbf{c}_{dN\times 1}} - \underbrace{\begin{pmatrix} \vdots \\ \vdots \\ \mathbf{m} \\ \mathbf{m} \\ \mathbf{m} \\ \vdots \\ \vdots \end{pmatrix}}_{\mathbf{m}_{3dN\times 1}}$$

- $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$ - three frames of a certain phoneme.
- $\mathbf{m} = [\mathbf{m}^c, \mathbf{m}^{\Delta^1}, \mathbf{m}^{\Delta^2}]$ - replicated model mean.

Proposed Augmented Space Construction

Segment Wise, 'sw', $\widetilde{\mathbf{o}} = \widetilde{W}\mathbf{c}$, $J^{sw}(\widetilde{\mathbf{W}}\mathbf{c}) = \frac{1}{2}\|\widetilde{\mathbf{U}}^{-\frac{1}{2}}(\widetilde{\mathbf{W}}\mathbf{c} - \widetilde{\mathbf{m}})\|_2^2$

$$\frac{1}{3}\underbrace{\begin{pmatrix} \mathbf{0}_{d\times d} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ -\frac{\mathbf{1}}{\mathbf{2}} & -\frac{\mathbf{1}}{\mathbf{2}} & \mathbf{0} & \frac{\mathbf{1}}{\mathbf{2}} & \frac{\mathbf{1}}{\mathbf{2}} \\ -\mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & -\mathbf{1} \end{pmatrix}}_{\widetilde{\mathbf{W}}_{3dK\times dN}} \underbrace{\begin{pmatrix} \vdots \\ \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \\ \vdots \end{pmatrix}}_{3N\times 1} - \underbrace{\begin{pmatrix} \vdots \\ \mathbf{m} \\ \vdots \end{pmatrix}}_{\widetilde{\mathbf{m}}_{3KN\times 1}}$$

- $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$ - three frames of a certain phoneme.
- $\mathbf{m} = [\mathbf{m}^c, \mathbf{m}^{\Delta^1}, \mathbf{m}^{\Delta^2}]$ - non replicated model mean.

Cost Function

$$J_c^{SW}(\widetilde{\mathbf{W}}\mathbf{c}) = \frac{1}{2}\|\widetilde{\mathbf{U}}^{-\frac{1}{2}}(\widetilde{\mathbf{W}}\mathbf{c} - \widetilde{\mathbf{m}})\|_2^2 + \frac{\lambda}{2}\|\mathbf{c}\|_2^2$$

- Model term
- Norm controlling term
- Balancing factor

Cost Function

$$J_c^{SW}(\widetilde{\mathbf{W}}\mathbf{c}) = \frac{1}{2}\|\widetilde{\mathbf{U}}^{-\frac{1}{2}}(\widetilde{\mathbf{W}}\mathbf{c} - \widetilde{\mathbf{m}})\|_2^2 + \frac{\lambda}{2}\|\mathbf{c}\|_2^2$$

- Model term
- Norm controlling term
- Balancing factor

Iterative Speech Feature Generation

- Minimize the norm-constrained cost function
  - $J_c^{sw}(\widetilde{\mathbf{W}}\mathbf{c}) = \frac{1}{2}\|\widetilde{\mathbf{U}}^{-\frac{1}{2}}(\widetilde{\mathbf{W}}\mathbf{c} - \widetilde{\mathbf{m}})\|_2^2 + \frac{\lambda}{2}\|\mathbf{c}\|_2^2$
- Using the gradient descent algorithm
  - $\mathbf{c}_{n+1} = \mathbf{c}_n - \alpha_n \nabla(\mathbf{c_n})$
  - $\nabla(\mathbf{c_n}) = \widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{W}}\mathbf{c}_n - \widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\mathbf{m} + \lambda\mathbf{c}_n$
- Applying a variable balancing factor, ensuring
  - A required norm increase.
  - A good approximation to the models.

Variable balancing factor

- Exponentially decreasing
  - $\lambda_{n+1} = \theta\lambda_n,\ 0 \leq \theta \leq 1$
- Corresponding gradient
  - $\nabla(\mathbf{c_n}) = \widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{W}}\mathbf{c}_n - \widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{m}} + \lambda_n\mathbf{c}_n$
- λ_0
  - Sufficient to compensate the norm reduction.
  - Experimentally derived.

Variable Balancing Factor, λ

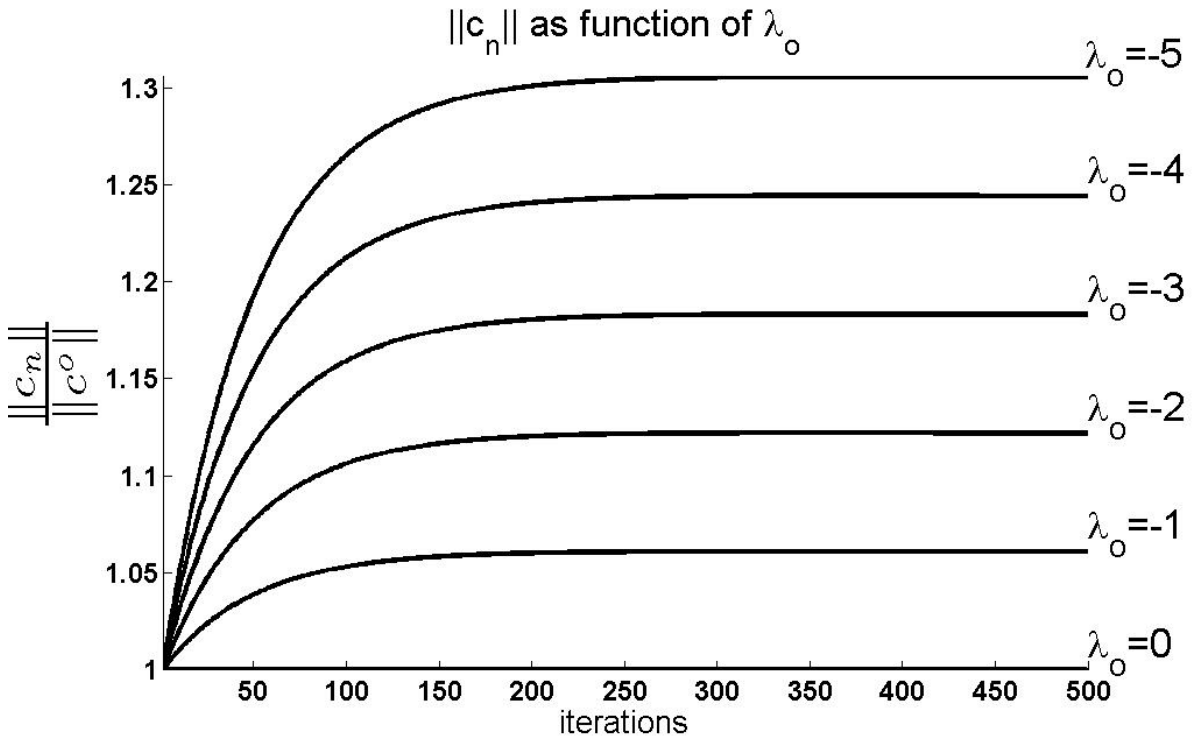


Figure: $\mathbf{c}_n$ norm increases as a function of an initial value for λ_o, where $\|c^o\|$ is a norm of an initial vector.

Cost Function Value Comparison

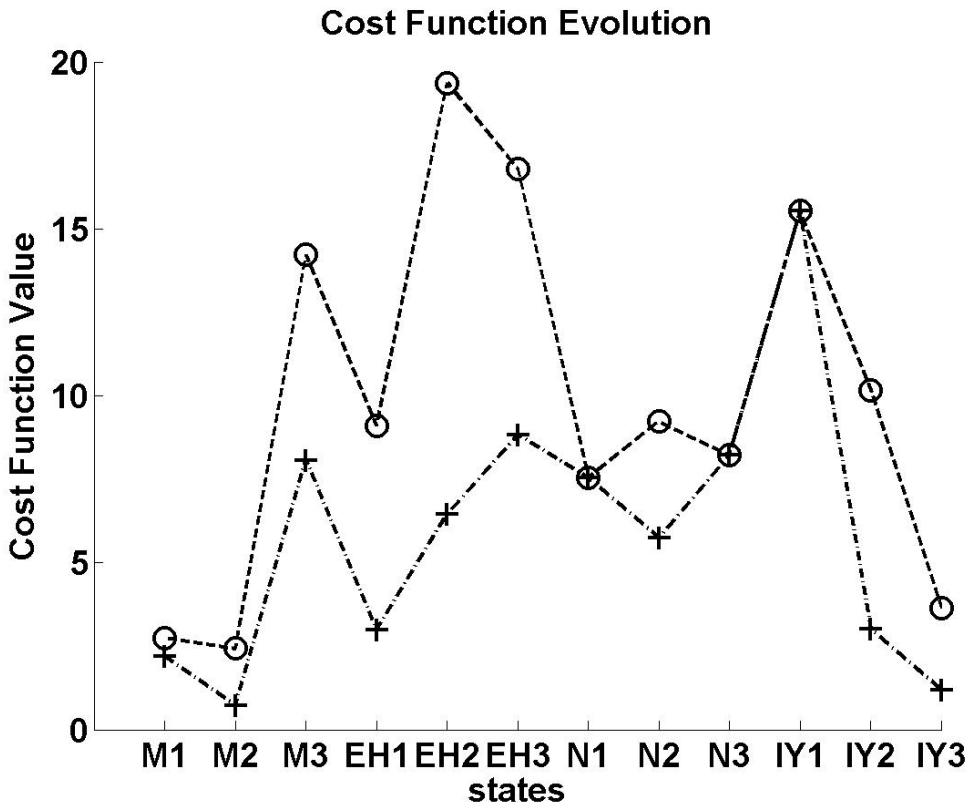


Figure: Frame-wise cost function values in circles; Segment-wise cost function values in pluses

Results - Generated Speech Feature Trajectory

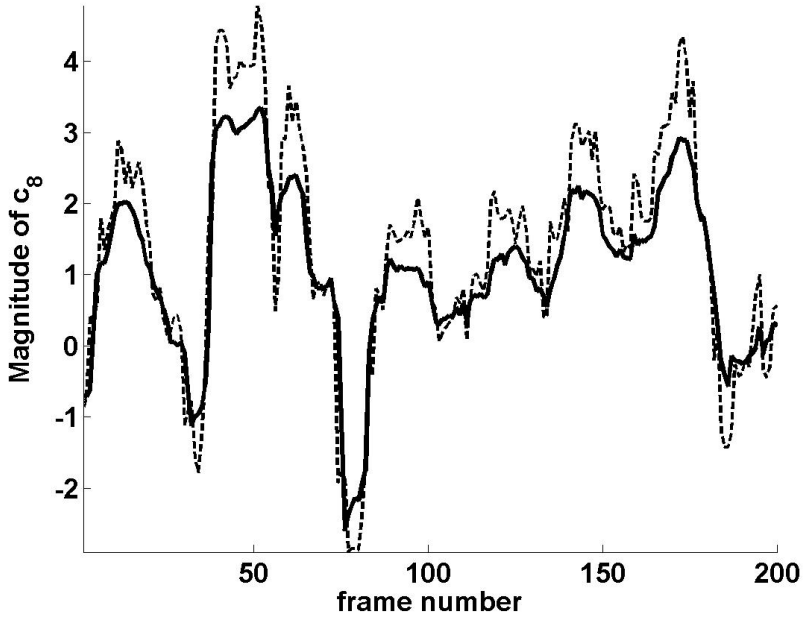


Figure: '**Many problems in reading and writing are due to old habits**', solid line - 'fw' trajectory; dashed line - 'sw' trajectory.

Generated Speech Feature Trajectory - Zooming in

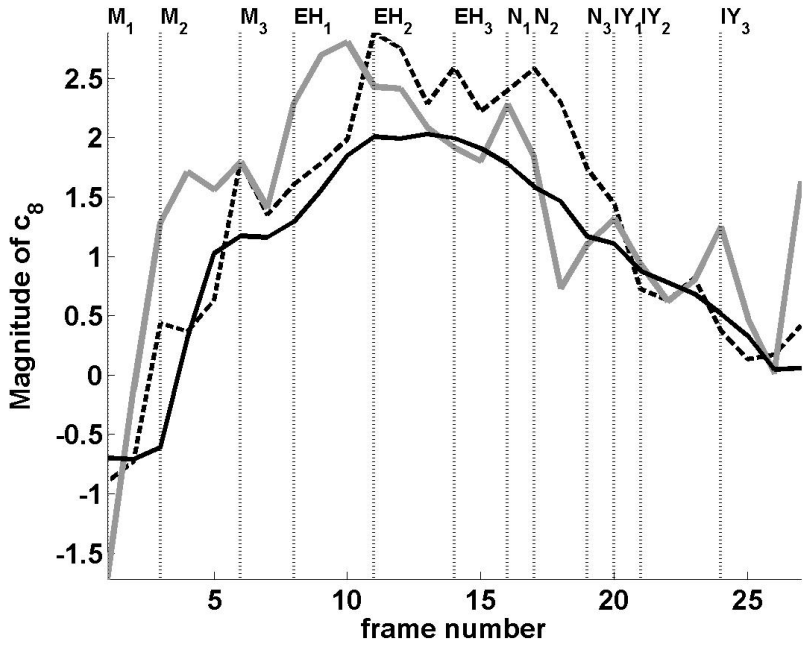
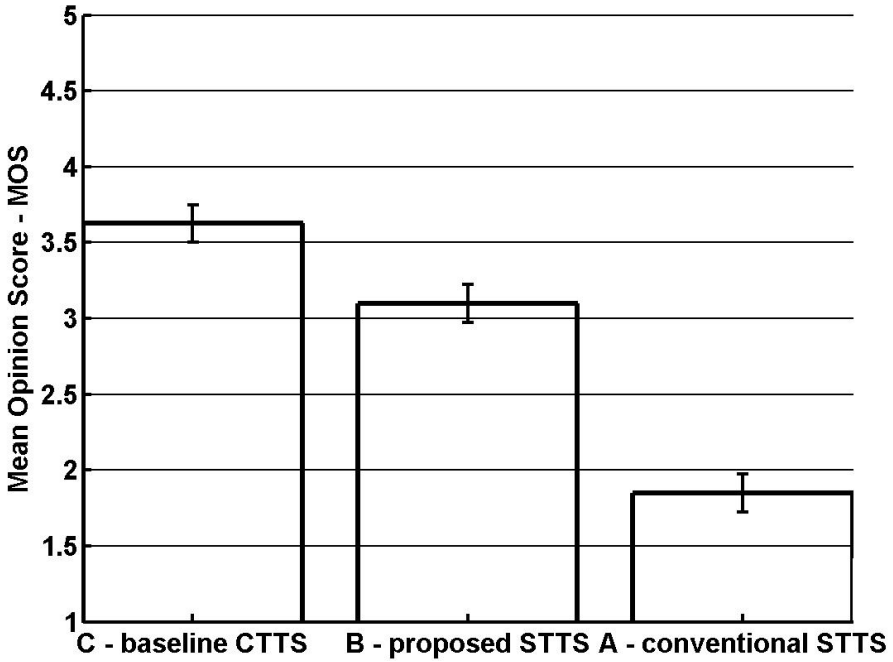


Figure: '**Many**', solid line - 'fw' trajectory; dashed line - 'sw' trajectory; light grey line - natural trajectory. HMM states marked above.

Subjective Evaluation - MOS Test



"Answer the following question as carefully and completely as possible." conventional stts proposed stts

27 samples, 20 listeners.

Outline

Proposed Hybrid TTS Concepts

- Combine natural segments with statistical models.
  - Hybrid Dynamic Path.
- Natural segments connect optimally to statistical segments.
  - Constrained Statistical Model.
- Use improved statistical models in hybrid vector.

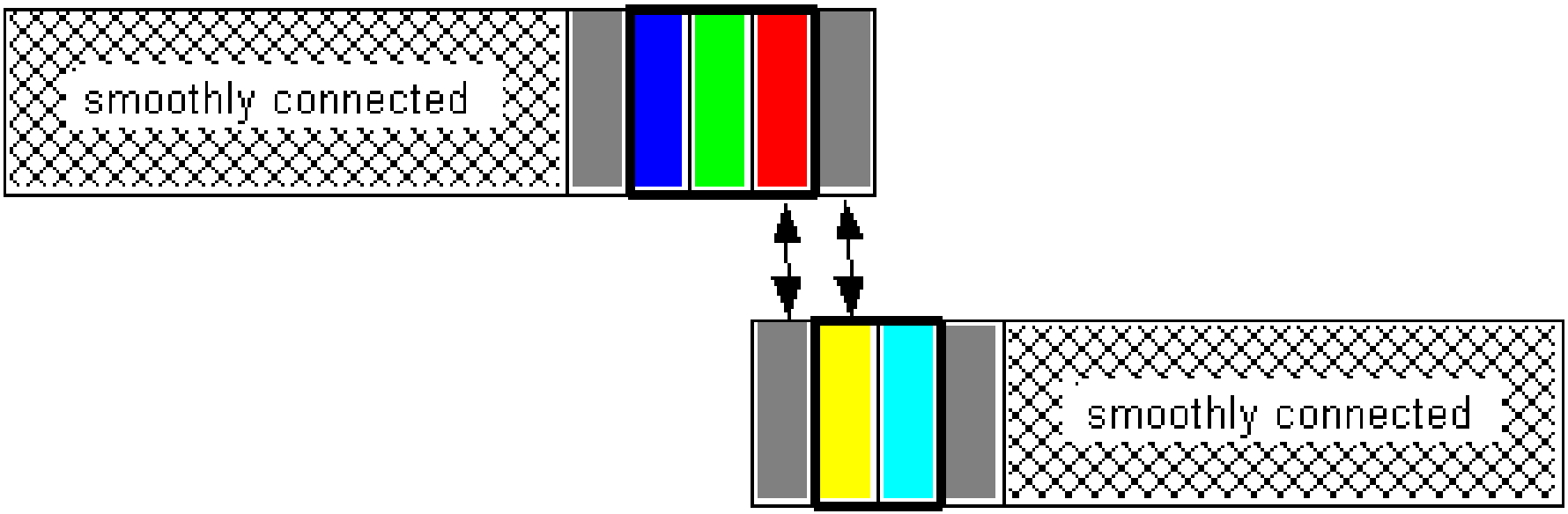
Proposed Hybrid TTS Concepts

- Combine natural segments with statistical models.
  - Hybrid Dynamic Path.
- Natural segments connect optimally to statistical segments.
  - Constrained Statistical Model.
- Use improved statistical models in hybrid vector.

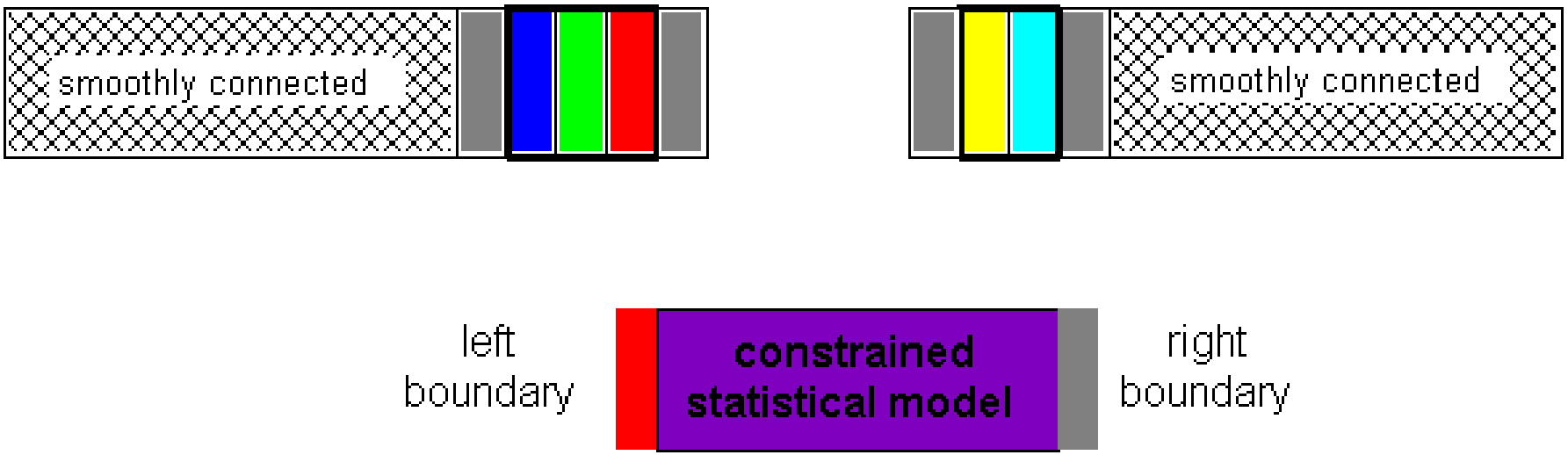
Proposed Hybrid TTS Concepts

- Combine natural segments with statistical models.
  - Hybrid Dynamic Path.
- Natural segments connect optimally to statistical segments.
  - Constrained Statistical Model.
- Use improved statistical models in hybrid vector.

Composing Hybrid Utterance

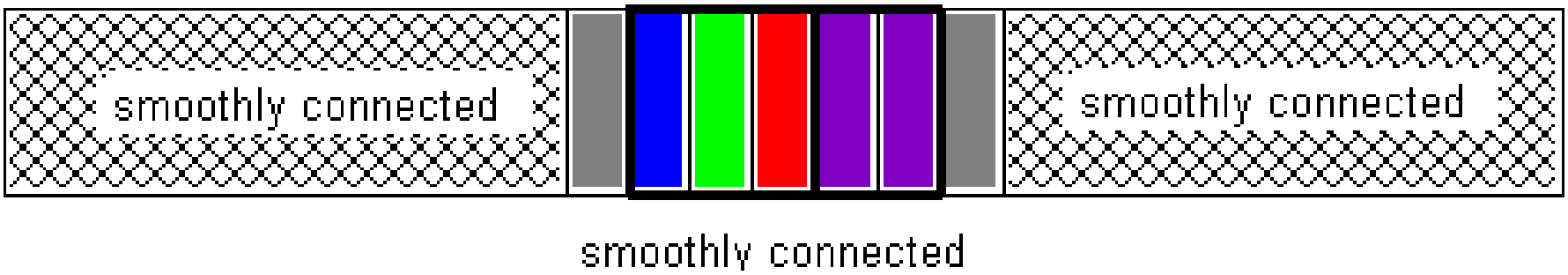


Composing Hybrid Utterance



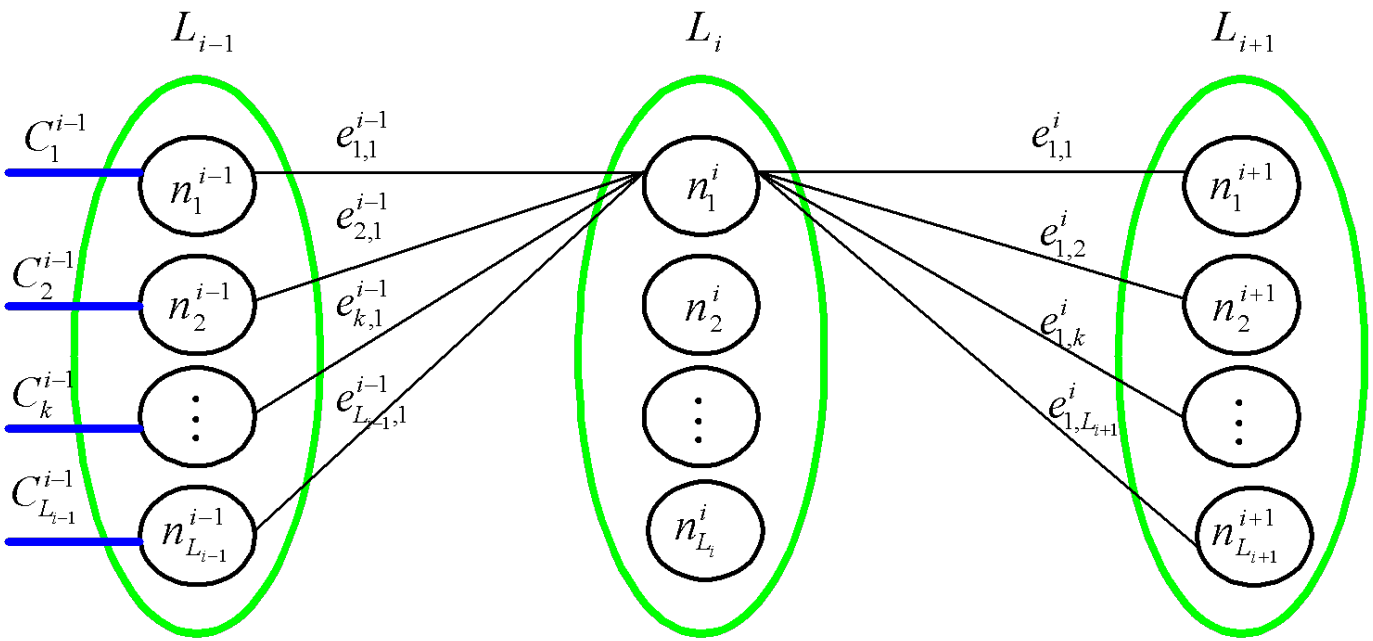
Constrained statistical model bridge these two smooth parts.

Composing Hybrid Utterance



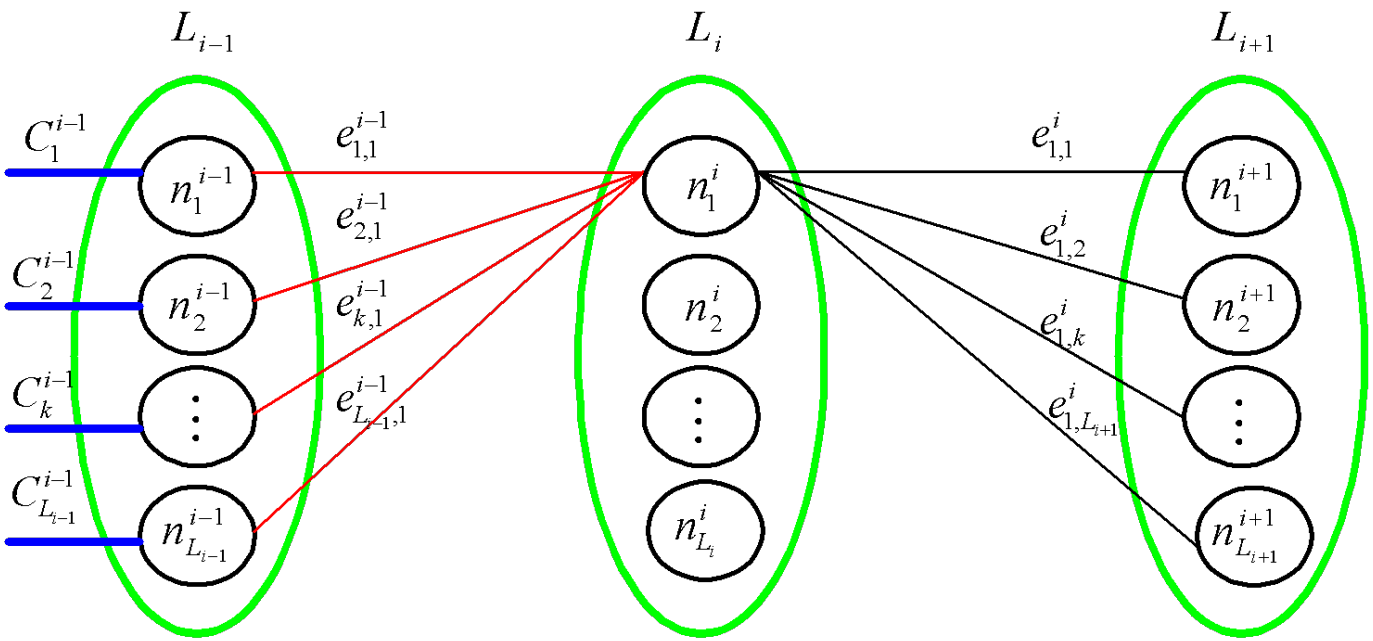
- Hybrid utterance
  - Discontinuities alleviated
  - Optimal Connections
- Features
  - Improved naturalness due to natural segments
  - Smoothed transitions due to statistical models

Hybrid Dynamic Path



- L_i - *i-th* stage of DP, representing a particular phoneme.
- n_k^i - *k-th* node (segment) of L_i.
- C_k^i - a cumulative cost at n_k^i.
- $e_{j,k}^i$ - a concatenative cost between n_j^i and n_k^{i+1}

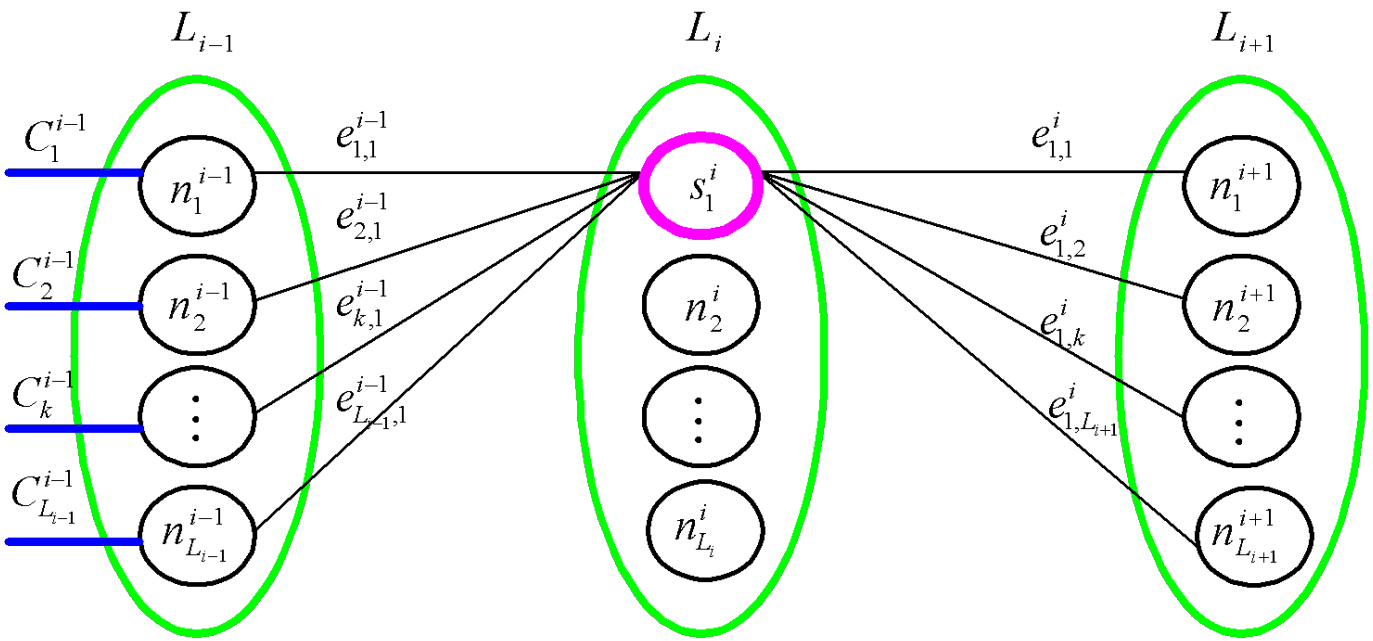
Hybrid Dynamic Path



$$\text{if } \forall j,\ e_{j,1}^{i-1} > \epsilon$$

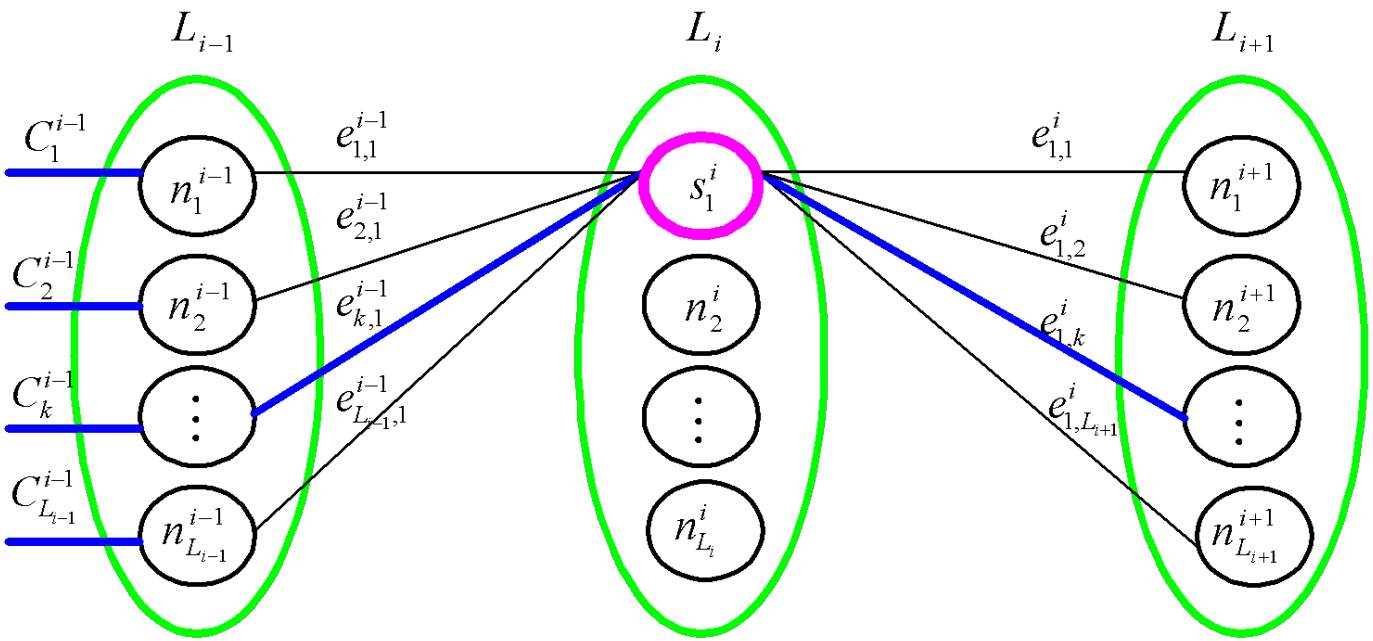
then any path, passing through n_1^i, includes a discontinuity.

Hybrid Dynamic Path



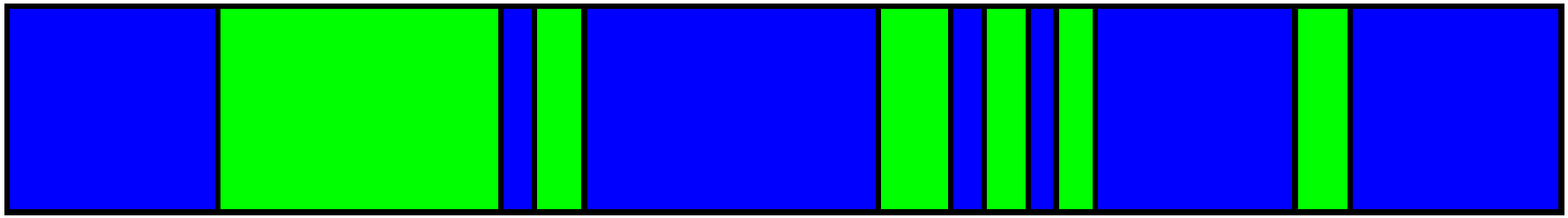
- Replace n_1^i by s_1^i.

Hybrid Dynamic Path



where $C_1^i = \min_j(C_j^{i-1})$

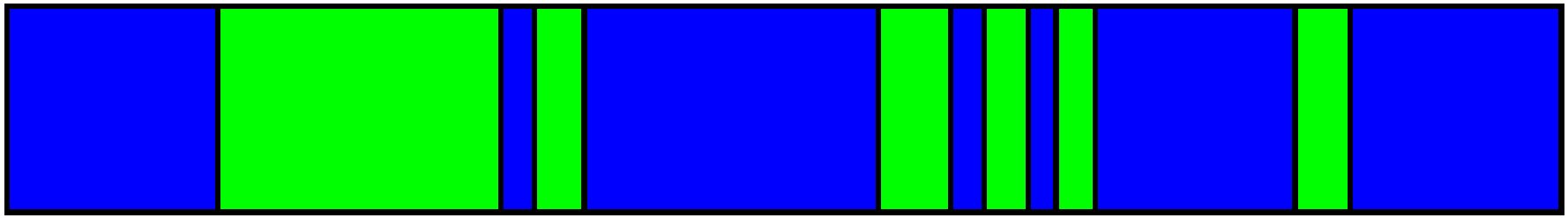
Hybrid Speech Feature Vector



"Meaning is the most essential part of all thought processing"

- Natural.
- Statistical.
- Allocations by the *hybrid dynamic path*
- Optimal Connections.

Hybrid Speech Feature Vector



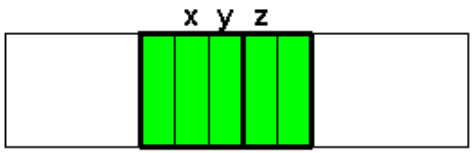
"Meaning is the most essential part of all thought processing"

- Natural.
- Statistical.
- Allocations by the *hybrid dynamic path*
- Optimal Connections.

Constrained Statistical Model

Segments connections via boundary dynamic features

- Unconstrained



$$\Delta^1 y = \frac{1}{2}(z^{stt} - x^{stt}),$$
$$\Delta^2 y = -x^{stt} + 2y^{stt} - z^{stt}.$$

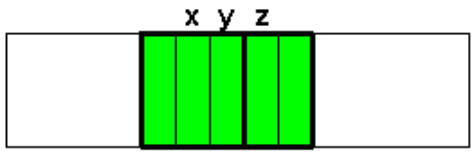
- Constrained

$$\Delta^1 y = \frac{1}{2}(z^{nat} - x^{stt}),$$
$$\Delta^2 y = -x^{stt} + 2y^{stt} - z^{nat}.$$

Constrained Statistical Model

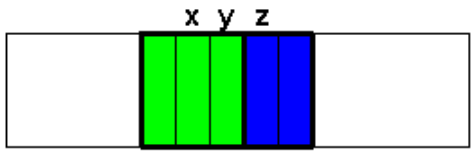
Segments connections via boundary dynamic features

- Unconstrained



$$\Delta^1 y = \frac{1}{2}(z^{stt} - x^{stt}),$$
$$\Delta^2 y = -x^{stt} + 2y^{stt} - z^{stt}.$$

- Constrained



$$\Delta^1 y = \frac{1}{2}(z^{nat} - x^{stt}),$$
$$\Delta^2 y = -x^{stt} + 2y^{stt} - z^{nat}.$$

Optimal Hybrid Speech Feature Vector

Problem Setting

$$\begin{aligned} c^{opt,hybrid} &= \underset{\mathbf{c}}{argmin}\, J(W\mathbf{c}), \\ \text{s.t. } Ac &= c^{*}. \end{aligned}$$

Definitions

$$\begin{aligned} J(W\mathbf{c}) &- \text{statistical model over an utterance,} \\ A_{dT\times dK} &- \text{constraints matrix,} \\ c^{*} &- \text{K constrained natural frames.} \end{aligned}$$

Optimal Hybrid Speech Feature Vector

Problem Setting

$$\begin{aligned} c^{opt,hybrid} &= \underset{\mathbf{c}}{argmin}\, J(W\mathbf{c}), \\ \text{s.t. } Ac &= c^{*}. \end{aligned}$$

'A' example

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \end{pmatrix}_{dT \times dK}$$

Optimal Hybrid Speech Feature Vector

Problem Setting

$$\begin{aligned} c^{opt,hybrid} &= \underset{\mathbf{c}}{argmin}\, J(W\mathbf{c}), \\ \text{s.t. } Ac &= c^*. \end{aligned}$$

Solution via Lagrangian function

$$L(c,\gamma) = \frac{1}{2}(W\mathbf{c}-\mathbf{m})^T U^{-1}(W\mathbf{c}-\mathbf{m}) + \gamma(Ac-c^*),$$

where

$\gamma_{1\times dK}$ – vectorial Lagrange multiplier.

Hybrid Cost Function

$$J(\mathbf{c}, \gamma) = \frac{1}{2}\|U^{-\frac{1}{2}}(W\mathbf{c} - \mathbf{m})\|_2^2 + \frac{\lambda}{2}\|\mathbf{c}\|_2^2 + \gamma(A\mathbf{c} - \mathbf{c}^*)$$

- Models term
- Norm controlling term
- Constraints on natural frames

Hybrid Cost Function

$$J(\mathbf{c}, \gamma) = \frac{1}{2}\|U^{-\frac{1}{2}}(W\mathbf{c} - \mathbf{m})\|_2^2 + \frac{\lambda}{2}\|\mathbf{c}\|_2^2 + \gamma(A\mathbf{c} - \mathbf{c}^*)$$

- Models term
- Norm controlling term
- Constraints on natural frames

Hybrid Speech Features Trajectory

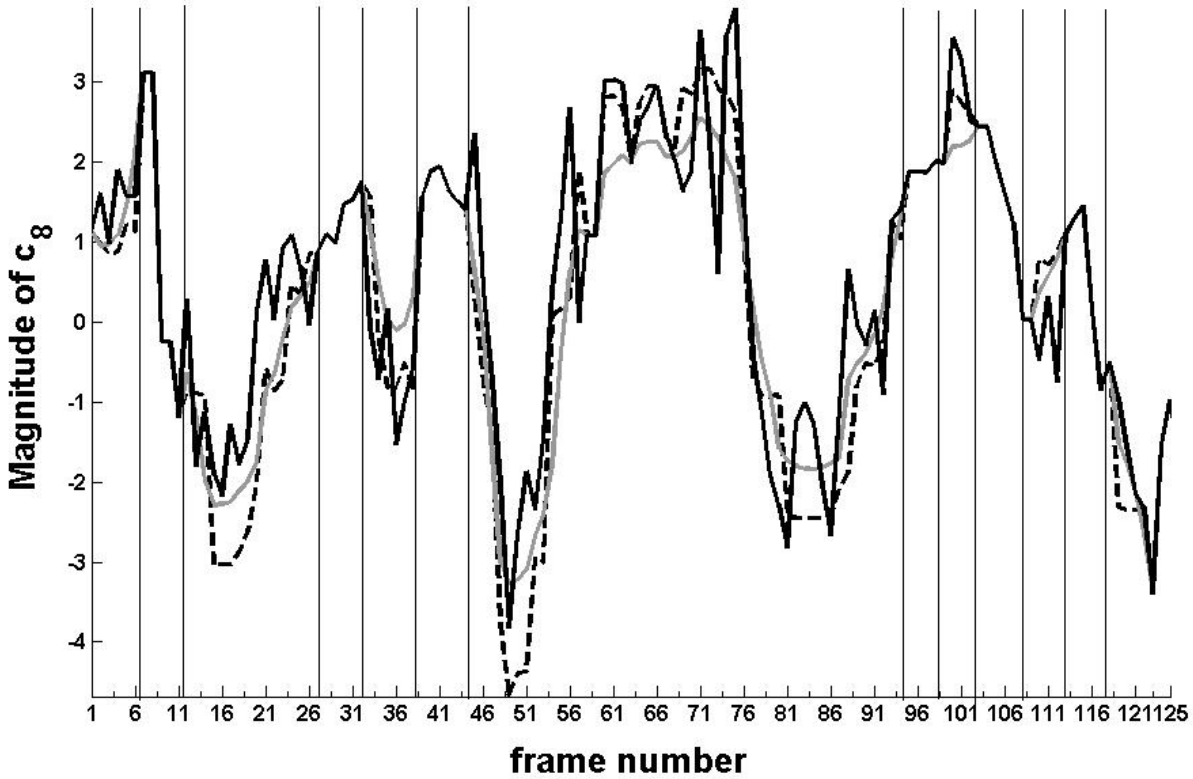
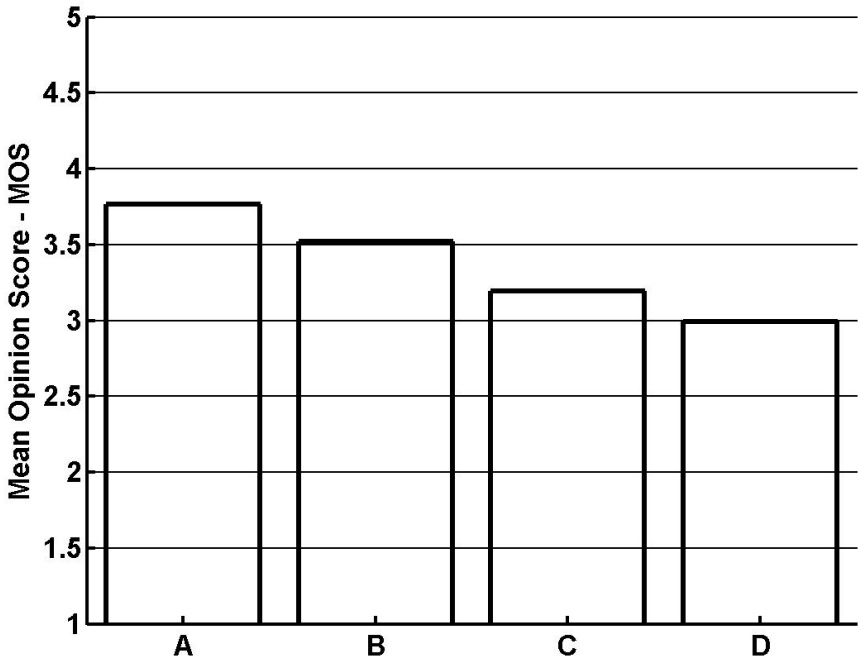


Figure: natural trajectory in solid black line; hybrid - 'sw' in dashed line; hybrid - 'fw' in light grey line

Subjective Evaluation

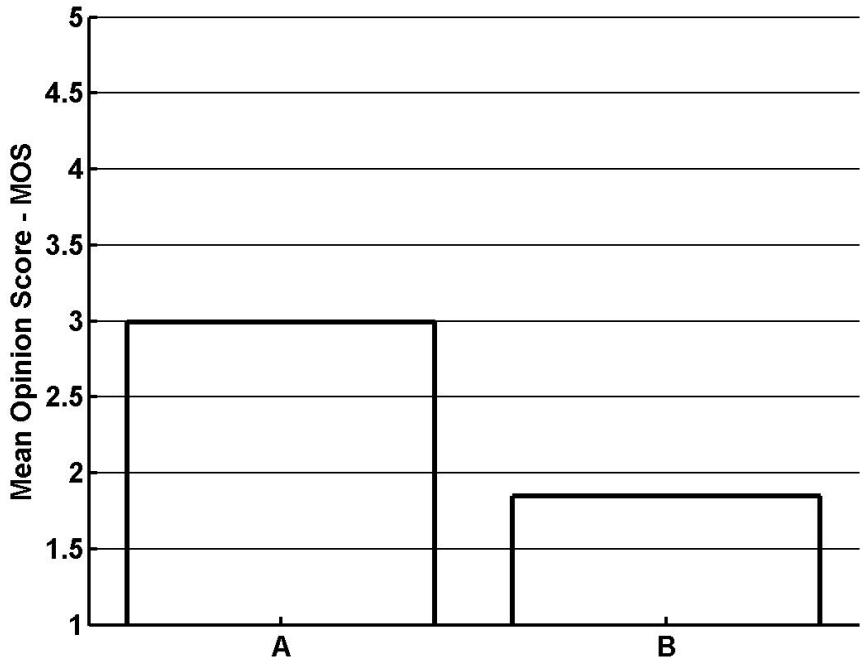


A - CTTS(22MB), B - SW-HTTS(8.3MB), C - CTTS(8.3MB), D - FW-HTTS(8.3MB)
40 samples, 10 listeners.

"Now we will say name again." ctts , htts

Discussion

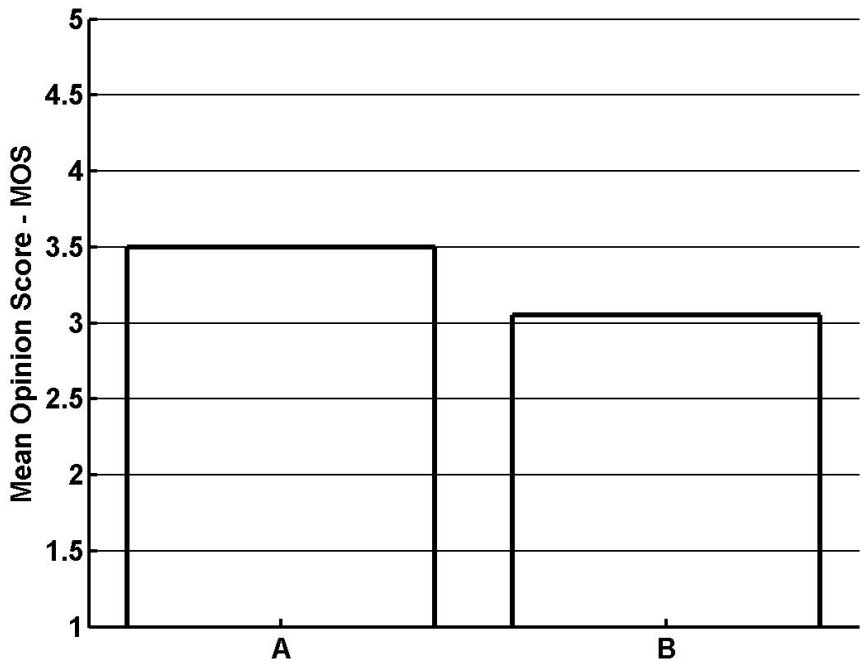
Improving conventional STTS by HTTS



A - FW-HTTS, B - baseline FW-STTS.

Discussion

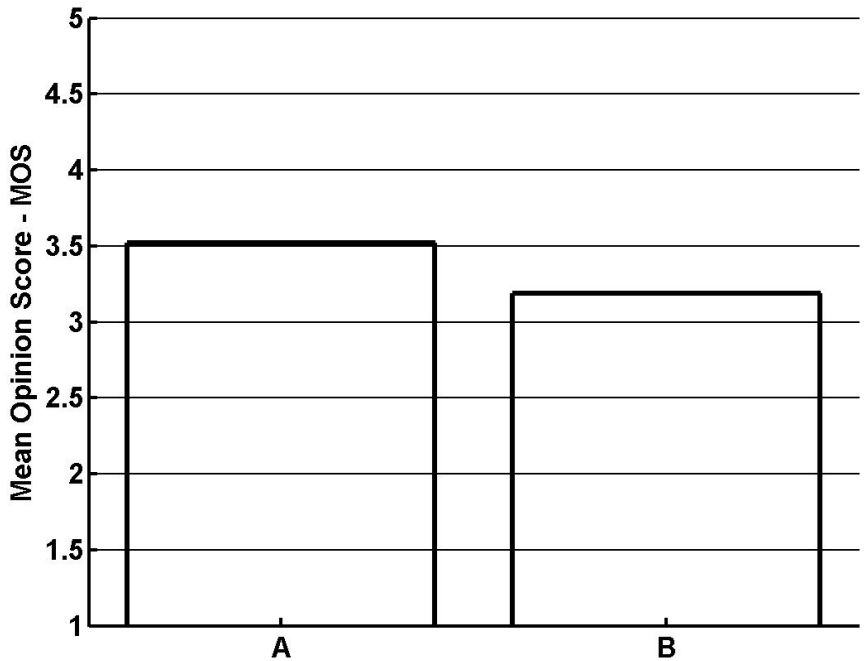
Improving 'segment-wise' STTS in HTTS



A - SW-HTTS, B - baseline SW-STTS.

Discussion

Improving CTTS by 'segment-wise' HTTS



A - SW-HTTS(8.3MB), B - CTTS(8.3MB).

Outline

Summary

- Improved Quality STTS
  - Transform Domain Enhancement
  - Segment-Wise Representation
  - Norm Regulated Speech Feature Vector
  - Iterative Solution with a variable balancing factor
- Hybrid Text-To-Speech System
  - Hybrid Dynamic Path
  - Boundary Constrained Statistical Models
  - Iterative Hybrid Speech features vector generation
- Publications
  - A paper on statistical dynamics, INTERSPEECH-2008.
  - A paper on the Wegment-Wise STTS, IEEE Transactions on Audio, Speech and Language Processing, in revision.
  - A paper on the proposed hybrid TTS, in preparation.

Future Work

- Phase modeling.
- Prosody modeling.
  - Explore a general framework for speech features/prosody modeling.
- Hybridism at broad phonetic classes level.
  - Different phoneme classes to be modeled differently.

THANK YOU!