Statistical Text-to-Speech Synthesis with Improved Dynamics

Stas Tiomkin[†‡] and David Malah[†]

[†]Department of Electrical Engineering, Technion - I.I.T, [‡]Speech Technologies Group, IBM, Israel.

Main Text-To-Speech (TTS) Synthesis Methods

- Concatenative TTS - CTTS
  - Concatenation of natural samples from an inventory.
  - Advantages
    - Natural quality speech.
    - Low computational complexity
  - Disadvantages
    - Essential quality degradation with a footprint reduction.
    - Speaker dependent.

- Statistical TTS - STTS
  - Speech parameters generation from statistical models.
  - Advantages
    - Low footprint.
    - Voice modification.
  - Disadvantages
    - Insufficient speech features dynamics.
    - Muffled and buzzy speech.

Conventional STTS

- Speech feature vector over an entire utterance
  - $\mathbf{c} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_T]$
    - $\mathbf{c}_i = [c_1, c_2, \ldots, c_d]$
- Augmented space vector
  - $\mathbf{o} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T]$
    - $\mathbf{o}_i = [\mathbf{c}_i, \Delta^1\mathbf{c}_i, \Delta^2\mathbf{c}_i]$
    - $\Delta^{1,2}\mathbf{c}_i = \sum_{j=-L^{1,2}}^{L^{1,2}} \omega_j^{1,2}\mathbf{c}_{i+j}$
    - $\mathbf{o} = \mathbf{W}\mathbf{c}$

- Statistical model
  - $P(\mathbf{o}) \sim \mathcal{N}(\mathbf{m}, \mathbf{U})$
- Optimal Solution

$$\mathbf{c}^{opt} = \underset{\mathbf{c}}{argmin}\, ln(P(o))_{|_{\mathbf{o}=\mathbf{W}\mathbf{c}}} = \\ = (\mathbf{W}^{\mathbf{T}}\mathbf{U}^{-\mathbf{1}}\mathbf{W})^{-1}\mathbf{W}^{\mathbf{T}}\mathbf{U}^{-\mathbf{1}}\mathbf{m}. \quad (1)$$

Generated Speech Features

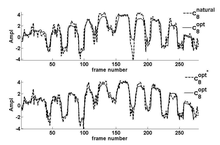


FIGURE 1: Evolution of a particular frequency component. The upper plot shows natural features (dashed line) and conventionally generated features (solid line). The bottom plot shows features generated by the proposed method (dashed line), described below, and conventionally generated features (solid line).

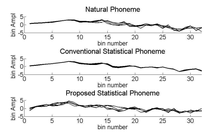


FIGURE 2: Evolution of frames within a particular phoneme in natural phoneme (top plot), in conventionally generated phoneme (middle plot), and in a phoneme generated by the proposed method (bottom plot). The conventionally generated phoneme lacks inter-frames dynamics, appeared in natural phoneme.

Drawbacks of Conventional STTS

- Statistically generated components are over-smoothed.
- Over-smoothing causes muffled and buzzy speech.
- $\Delta^{1,2}\mathbf{c}$ do not model speech dynamics sufficiently.

Previously Proposed Solution to Over-Smoothing

- Global Variance (GV) approach has been proposed[a]:
  - Applying a penalty for global variance reduction.
  - GV improves statistically generated speech quality.
  - GV is computationally complex.

[a]Tomoki Toda, Keiichi Tokuda, Speech parameter generation algorithm considering global variance for HMM-based speech synthesis, In INTERSPEECH-2005, 2801-2804

Arranging Data as Quasi-Periodic Sequences

A phoneme of T_i frames is represented as a one-dimensional coefficients sequence of length dT_i with a period d.

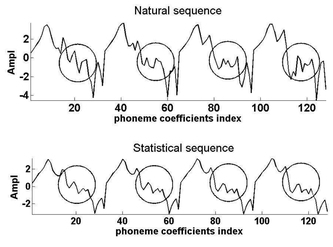
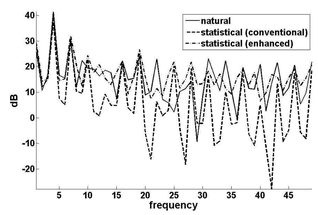


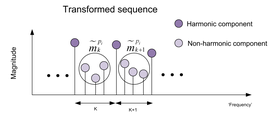
FIGURE 3: Features frames of a natural sequence (4 frames on the same plot) at the top; statistical sequence at the bottom. The frame-to-frame variations between circled regions demonstrate the low dynamics in the statistical sequence as compared to the natural sequence..

Modeling Speech Features Dynamics in the Transform Domain

- Apply a DFT of length dT_i to the quasi-periodic sequence:
  - Harmonic frequencies: $k = lT_i, l = 1, 2, \ldots, d$
  - Non-harmonic frequencies: $k+1, k+2, \ldots, k+T_i-1$, where $k = 1, T_i, 2T_i, \ldots, (d-1)T_i$
- Non-harmonic content (NHC) in statistically generated phonemes is much lower, compared to NHC in natural phonemes.



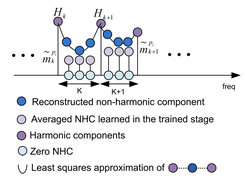
Improving Speech Features Dynamics - Learning Non-Harmonic Components Statistics



$$\widetilde{\mathbf{m}}^{p_i} = \left[\widetilde{m}_1^{p_i}, \widetilde{m}_2^{p_i}, \ldots, \widetilde{m}_d^{p_i}\right]$$

1. Apply DFT to quasi-periodic sequences of natural segments from the phoneme p_i.
2. Extract NHC between every two harmonic components.
3. Compute $\tilde{m}_k^{p_i}$ as an average value of NHC, pertaining to the k-th non-harmonic interval.
4. Compose $\tilde{\mathbf{m}}^{p_i}$ for the phoneme p_i.

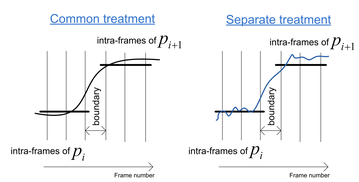
Improving Speech Features Dynamics - Enhancing Non-Harmonic Component



1. Replicate the model mean $\mathbf{m}_i$ T_i-times.
2. Apply a DFT of length T_i to the periodic sequence.
   - NHC in the transformed sequence are exactly zero.
3. Find a least squares approximation by a second order polynomial of the points: $H_k, \underbrace{\cdots \tilde{m}_k^{p_i} \cdots}_{\times T_i - 1}, H_{k+1}$.

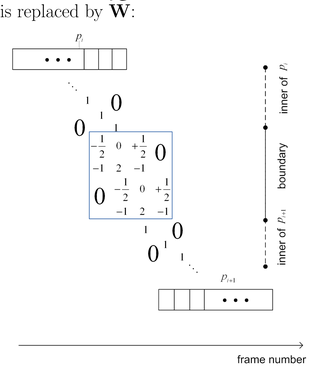
Utterance-Level Synthesis

- Problem Setting
  - The inter phoneme boundaries and intra phoneme frames are treated equally.
  - Frames of both types are over-smoothed.
    - Smooth inter phoneme transitions are desired.
    - Intra phoneme smoothing reduces features dynamics.
  - We propose a separate treatment for inter phoneme boundaries and intra phoneme frames.



Modified Linear Transformation

- Connect smoothly adjacent phonemes, using $\Delta^{1,2}\mathbf{c}$ at boundaries.
- Enable enhanced features dynamics within phonemes.
- $\mathbf{W}$ in (1) is replaced by $\widetilde{\mathbf{W}}$:



Synthesis scheme

Given an utterance, composed of K phonemes, $p_1, p_2, \ldots, p_K$, having lengths $T_1, T_2, \ldots, T_k$

- Enhance the non-harmonic components in each phoneme
- Construct model mean and model covariance matrix:
  - $\widetilde{\mathbf{m}} = [\underbrace{m_{p_1}^{stt}}_{\times T_1}, m_{p_1}^{\Delta^1}, m_{p_1}^{\Delta^2}, m_{p_2}^{\Delta^1}, m_{p_2}^{\Delta^2}, \underbrace{m_{p_2}^{stt}}_{\times T_2}, \ldots, m_{p_K}^{\Delta^1}, m_{p_K}^{\Delta^2}, \underbrace{m_{p_K}^{stt}}_{\times T_K}]$
    - $m_{p_i}^{stt}$ is the static features mean of p_i
    - $m_{p_i}^{\Delta^{1,2}}$ are the dynamic features mean of p_i
  - $\widetilde{\mathbf{U}} = diag[\underbrace{U_{p_1}^{stt}}_{\times T_1}, U_{p_1}^{\Delta^1}, U_{p_1}^{\Delta^2}, U_{p_2}^{\Delta^1}, U_{p_2}^{\Delta^2}, \underbrace{U_{p_2}^{stt}}_{\times T_2}, \ldots, U_{p_K}^{\Delta^1}, U_{p_K}^{\Delta^2}, \underbrace{U_{p_K}^{stt}}_{\times T_K}]$
    - $U_{p_i}^{stt}$ is the static features covariance of p_i
    - $U_{p_i}^{\Delta^{1,2}}$ are the dynamic features covariances of p_i

- Find the optimal solution, using $\widetilde{\mathbf{W}}, \widetilde{\mathbf{m}}$ and $\widetilde{\mathbf{U}}$ in (1):
  - $\tilde{\mathbf{c}}^{opt} = (\widetilde{\mathbf{W}}^{\mathbf{T}}\widetilde{\mathbf{U}}^{-\mathbf{1}}\widetilde{\mathbf{W}})^{-1}\widetilde{\mathbf{W}}^{\mathbf{T}}\widetilde{\mathbf{U}}^{-\mathbf{1}}\widetilde{\mathbf{m}}$.

Results & Conclusions

- Subjective evaluation
  - Comparison test 'A vs B'
    - 'A' - conventional approach
    - 'B' - proposed approach
  - Ten arbitrary utterances were generated, using each approach.
  - Twelve listeners, graduate and undergraduate students, participated.
  - 81.7% of listeners preferred 'B' utterances.
- The proposed method:
  - improves speech features dynamics.
  - is not computationally complex.
  - is a data-driven approach.