# Global Variance in Speech Synthesis with Linear Dynamical Models

Vassilis Tsiaras[1], Ranniery Maia[2], Vassilis Diakoloukas[1], Yannis Stylianou[2] and Vassilis Digalakis[1]

[1]Technical University of Crete, School of Electronic and Computer Engineering, Chania, Greece

[2]Toshiba Research Europe Limited, Cambridge Research Laboratory, Cambridge, UK

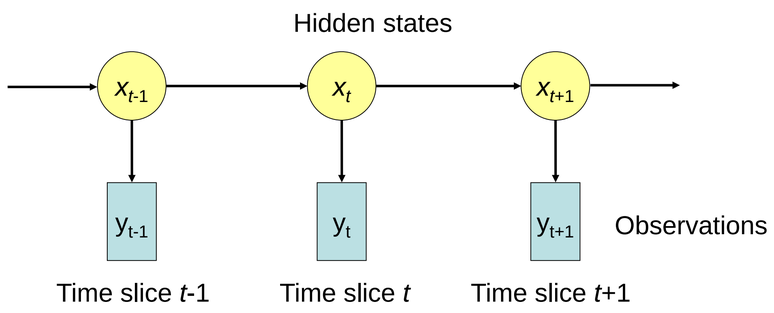
## Introduction

Hidden Markov models (HMMs) are the dominant approach for text-to-speech synthesis (TTS). However, HMMs are limited by several assumptions which do not apply to the properties of speech. In order to improve the quality of the synthesized speech many alternative statistical models have been proposed. Among these models, the Switching Linear Dynamical Models are particularly useful to represent the succession of homogeneous segments of speech. They are multi-level hybrid acoustic models where the transition between segments is described by an HMM, whereas the dynamics within each segment is described by a Linear Dynamical Model (LDM). When used as generative models, LDMs have low computational requirements, low-latency and produce speech of similar quality to HMMs using fewer parameters.

As in the case of HMMs, the trajectories of speech parameters generated from LDMs are over-smoothed due to statistical averaging of multiple trajectories during model training. This causes the degradation of perceptual quality and makes synthetic speech sound muffled. Inspired by the improvement in naturalness when the global variance (GV) is compensated in HMM-based speech synthesis, this work proposes a speech parameter generation algorithm that considers GV in LDM-based speech synthesis.

## LDMs

$$\begin{cases} x_1 \sim N(g_1, Q_1) \\ x_t = Fx_{t-1} + g + w^{(x)}, & w^{(x)} \sim N(0, Q) \\ y_t = Hx_t + \mu + w^{(y)}, & w^{(y)} \sim N(0, R) \end{cases}$$




## Generating a Segment of Speech Parameters with an LDM

Given an LDM with parameters $\theta$, we seek speech parameters $\hat{Y}$ that maximize the likelihood

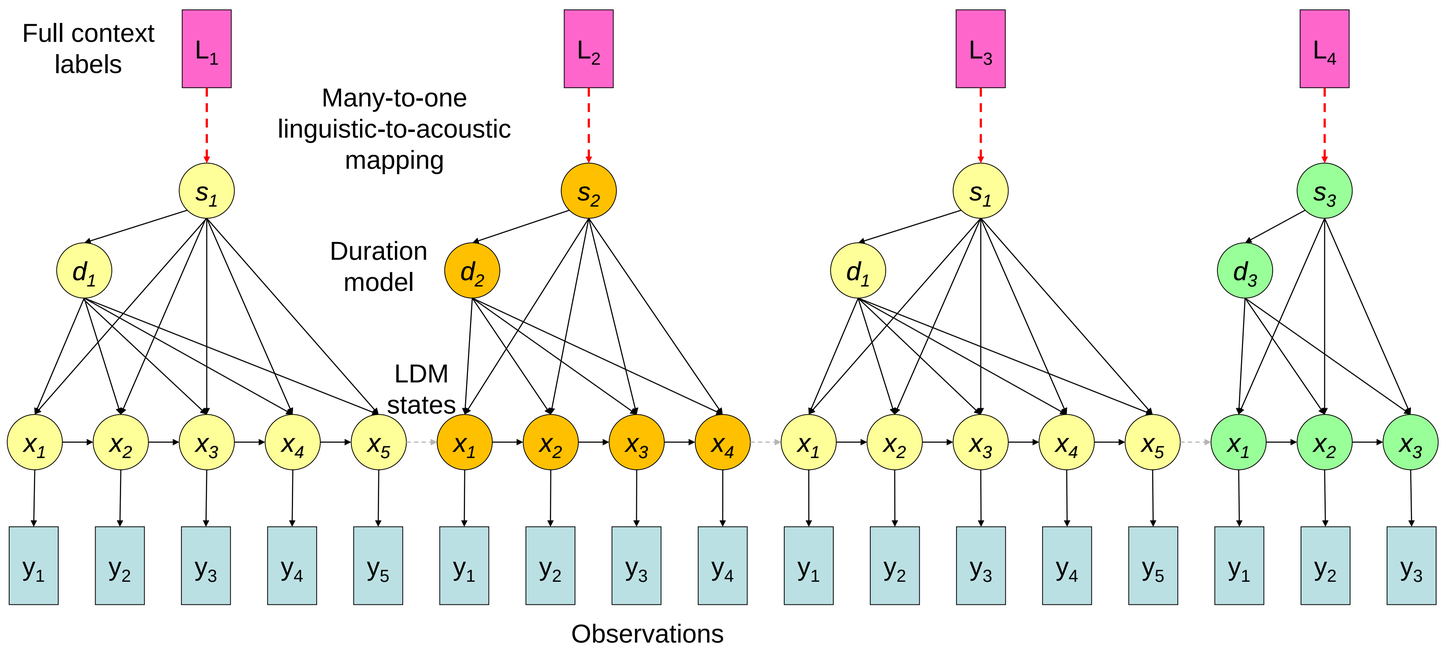
$$p(Y|\theta) = \int_X p(Y, X|\theta) dX$$

Since $p(Y, X|\theta)$ is Gaussian, the maximum of the likelihood is attained when the hidden state and the speech parameters are equal to their mean values.

$$\begin{aligned} \hat{x}_1 &= g_1 \\ \hat{x}_t &= F\hat{x}_{t-1} + g, \quad t = 2, \dots, T \\ \hat{y}_t &= H\hat{x}_t + \mu, \quad t = 1,2, \dots, T \end{aligned}$$

## Generating Speech Parameters of an Utterance with LDMs

- The front-end module of a TTS system produces linguistic and phonetic transcription of the input text.
- The linguistic labels are then associated with a sequence of LDM models through a linguistic-to-acoustic mapping.
- The duration of each segment is determined by an external model.
- A trajectory of speech parameters for an utterance is then produced as the concatenation of the trajectories generated from each one of the LDMs in the sequence.



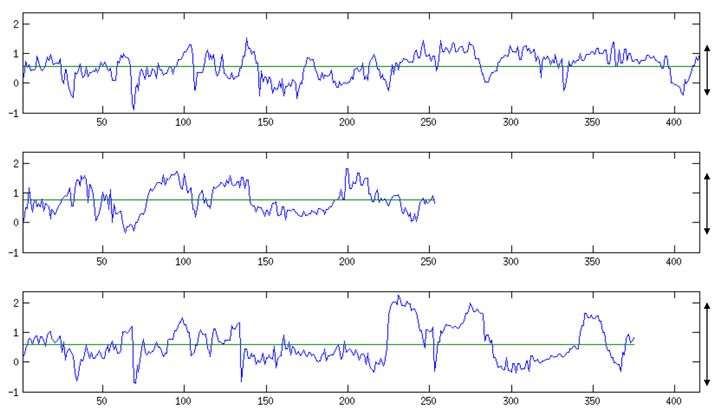

## Global Variance

For a given utterance trajectory $Y = [y_1, \dots, y_{T_u}]^\top$ of $m$ dimensional natural speech parameter vectors, the GV is defined on each dimension $k \in \{1, \dots, k, \dots, m\}$ independently as the intra-utterance variance of the $k$-th trajectory.

$$v = [v(1), \dots, v(k), \dots, v(m)]^\top \text{ where } v(k) = \frac{1}{T_u}\sum_{t=1}^{T_u}(y_t(k) - \bar{y}(k))^2 \text{ with } \bar{y}(k) = \frac{1}{T_u}\sum_{t=1}^{T_u} y_t(k)$$

The distribution of GV is modelled as a single Gaussian distribution $N(v; \mu_v, \Sigma_v)$, which is estimated from the GV vectors of the training sentences. The covariance matrix $\Sigma_v$ is diagonal, since the GV of each dimension is calculated independently of other dimensions.

**Example**




Trajectories of the 3rd cepstral coefficient

$$\mu_v(3) = mean([0.1848, 0.225, 0.3533]) = 0.2543$$
$$\Sigma_v(3,3) = var([0.1848, 0.225, 0.3533]) = 0.007745$$

## Global Variance Constrained LDM Synthesis

Given the sequence of labels of an utterance $u$, we seek speech parameters $\hat{Y}$ that jointly maximize the LDM likelihood and the likelihood of the GV.

$$p(Y|\Theta_u, \theta_v) = \int_X p(Y|X, \Theta_u, \theta_v) p(X|\Theta_u) dX$$

where the parameters $\Theta_u$ of the LDMs and the parameters $\theta_v$ of the GV Gaussian distribution are independently trained from the speech corpus.

It is assumed that:

a. the distribution of $X$ is independent of the parameter $\theta_v$ and

b. the probability density function $p(Y|X, \Theta_u, \theta_v)$ is written as a product of experts, where $Z$ is a normalizing constant and the weight $\omega$ is equal to 1.

$$p(Y|X, \Theta_u, \theta_v) = \frac{1}{Z} p(Y|X, \Theta_u) p(v(Y)|\theta_v)^{\omega T_u}$$

To reduce the computational cost, the trajectories of states $\hat{X}$, are chosen to maximize $p(X|\Theta_u)$.

Then the following log-scaled likelihood is maximized with respect to $Y$

$$L = -\frac{1}{2}\sum_{\varsigma \in labels(u)} \left( \sum_{t=1}^{T_\varsigma} (y_{\varsigma t} - \hat{y}_{\varsigma t})^\top R_q^{-1} (y_{\varsigma t} - \hat{y}_{\varsigma t}) - \frac{\omega T_u}{2} (v - \mu_v)^\top \Sigma_v^{-1} (v - \mu_v) \right)$$

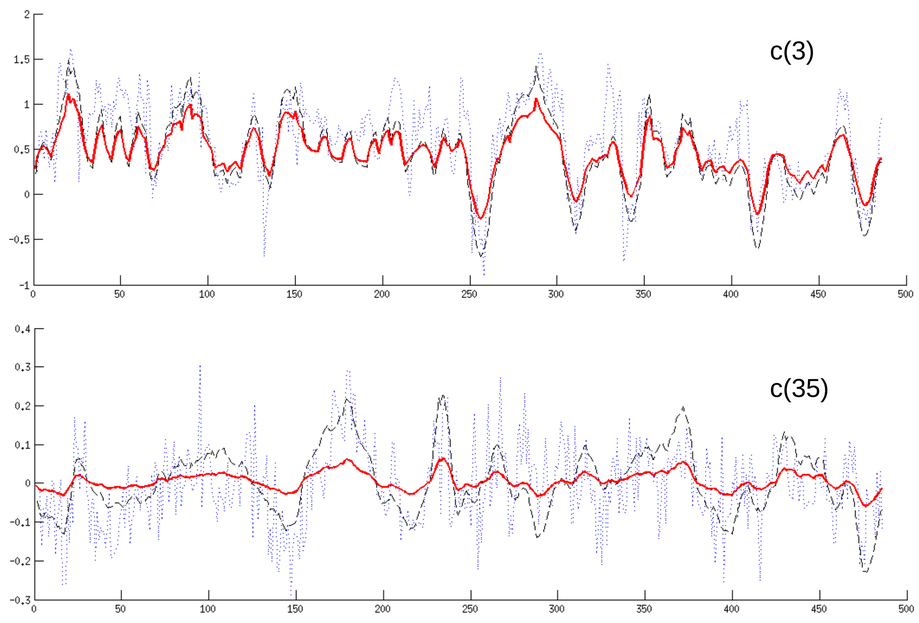
where $labels(u)$ is the sequence of labels of utterance $u$, $T_\varsigma$ is the duration associated to label $\varsigma$ and $T_u = \sum_{\varsigma \in labels(u)} T_\varsigma$ is the duration of utterance $u$. Also, $q$ is the index of the LDM model in which label $\varsigma$ is mapped. $R_q$ is the noise covariance of LDM model $q$ and $\hat{y}_{\varsigma t} = H_q \hat{x}_{\varsigma t} + \mu_q$. The index $t$ of $y_{\varsigma t}$ refers to the position of vector $y$ within the segment $\varsigma$.

To determine $Y$ we iteratively update $Y$ with the gradient method

$$Y^{(i+1)th} = Y^{(i)th} + \alpha \frac{\partial L}{\partial Y}$$

## Experiments

- A database containing 4417 sentences (approximately 5 h of speech) of an American English female speaker was used to verify the effectiveness of the proposed GV-based speech parameter method.
- Full context labels were created by using a proprietary front-end.
- From the training utterances, 40 mel-cepstral coefficients, 39 phase features, F0 and 20 mel-band-aperiodicity parameters were extracted at every 5 ms [3].
- The LDMs were trained as described in [2].
- In our experiments, GV is applied for mel-cepstral coefficients only, since the intention is to remove the muffled speech quality.
- A forced A-B test was conducted using 24 test sentences, with durations varying between 1.3–9.2 s (mean duration 3.7 s).
- Fifty four listeners took part in the test, where 11 of them were speech processing specialists.
- Each test sentence was synthesized in two versions:
  1. LDM synthesis without GV
  2. LDM synthesis with GV.
- The speech processing specialists had **100% preference**, while
- the non-specialists had **96.12% preference** for the utterances produced by the GV-based speech parameter algorithm.




Trajectories of the 3rd and 35th mel-cepstral coefficients. Blue dot–dot line: original. Red continuous thick line: LDM generated. Black dash–dash line: GV applied.

- From the above figures, it can be noticed that the effect of GV is more prominent in higher order coefficients and therefore the increase in speech quality comes mostly from improvements on the generated trajectories for higher order quefrencies of the generated cepstrum.

## Conclusions

- According to subjective preference tests, the proposed algorithm greatly improves the naturalness of the synthesized speech at an additional computational cost.

## References

1. T. Toda and K. Tokuda, “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis,” IEICE Trans. Inform. Syst., vol. E90-D, no. 5, pp. 816–824, May 2007.
2. V. Tsiaras, R. Maia, V. Diakoloukas, Y. Stylianou, and V. Digalakis, “Towards a linear dynamical model based speech synthesizer,” in Proc. Annu. Conf. Int. Speech Commun. Assoc. Interspeech, 2015, pp. 1221–1225.
3. R. Maia and Y. Stylianou, “Iterative estimation of phase using complex cepstrum representation,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2016, pp. 4990–4994.
4. V. Tsiaras, R. Maia, V. Diakoloukas, Y. Stylianou, V. Digalakis: Global Variance in Speech Synthesis with Linear Dynamical Models. IEEE Signal Processing Letters, vol. 23, no. 8, pp. 1057-1061, 2016.