This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2016.2580672, IEEE Signal Processing Letters

Global Variance in Speech Synthesis with Linear Dynamical Models

Vassilis Tsiaras, Ranniery Maia, Vassilis Diakoloukas, Yannis Stylianou, Vassilis Digalakis

***Abstract*—Linear Dynamical Models (LDMs) have been used in speech synthesis recently as an alternative to hidden Markov models (HMMs). Among the advantages of LDMs are the ability to capture the dynamics of speech and the achievement of synthesized speech quality similar to HMM-based speech systems on a smaller footprint. However, like in the HMM case, LDMs produce over-smoothed trajectories of speech parameters, resulting in *muffled* quality of synthetic speech. Inspired by a similar problem found in HMM-based speech synthesis, where the naturalness of the synthesized speech is greatly improved when the global variance (GV) is compensated, this paper proposes a novel speech parameter generation algorithm that considers GV in LDM-based speech synthesis. Experimental results show that the application of GV during parameter generation significantly improves speech quality.**

***Index Terms*—Statistical Parametric Speech Synthesis, Linear Dynamical Models, Global Variance**

I. Introduction

STATISTICAL Parametric Speech Synthesis (SPSS) uses acoustic models to represent the relationship between linguistic and acoustic features, the most popular being the Hidden Markov Models (HMMs) [1]. Due to inherent limitations of HMMs, there have been many attempts to develop more accurate acoustic models [2], [3], [4], [5]. Towards this direction, the Linear Dynamical Models (LDMs) have been shown to generate speech of similar naturalness to HMMs using fewer parameters [3]. LDMs have a continuous state space and they can generate smoothly-varying acoustic features. However, 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. A common and effective procedure to boost speech quality in HMM-based speech synthesis is the use of parameter generation by taking into account the Global Variance (GV) of the training parameters. The GV-based speech parameter method aims to retain the original utterance-level variation of a speech parameter trajectory. Subjective evaluation results have demonstrated that variance compensation using GV significantly improves the naturalness of synthetic speech compared to a parameter generation technique without GV modelling [6], [7].

V. Tsiaras, V. Diakoloukas and V. Digalakis are with the School of Electronic and Computer Engineering, Technical University of Crete, Greece.

R. Maia is with Cambridge Research Laboratory, Toshiba Research Europe Limited, Cambridge, UK.

Y. Stylianou is with the Computer Science Department, University of Crete, Heraklion, Greece and with Cambridge Research Laboratory, Toshiba Research Europe Limited, Cambridge, UK.

Manuscript received February ??, 2016; revised May ??, 2016.

In this work we propose a GV-based speech parameter generation algorithm for LDM speech synthesis by modifying the parameter generation algorithm which jointly maximizes a likelihood that is a combination of the LDM and GV likelihoods. The GV likelihood can be regarded as a penalty for the restoration of the variance of a generated parameter trajectory.

II. Preliminaries

A. Linear dynamical models

The LDMs are dynamical models with continuous state vectors. They consist of a state evolution process which is a linear first-order Gauss-Markov random process and an observation process which can be seen as a factor analyzer which maps the state vectors to the observation vectors. The output of the process follows a time-varying multivariate Gaussian distribution. An LDM can be specified by the following equations:

$$x_1 \sim \mathcal{N}(g_1, Q_1) \tag{1a}$$

$$x_k = Fx_{k-1} + g + w^{(x)}, \quad w^{(x)} \sim \mathcal{N}(0, Q) \tag{1b}$$

$$y_k = Hx_k + \mu + w^{(y)}, \quad w^{(y)} \sim \mathcal{N}(0, R) \tag{1c}$$

where F is a $n \times n$ state transition matrix, H is a $m \times n$ observation matrix, and Q and R are respectively $n \times n$ and $m \times m$ covariance matrices of the noise components $w^{(x)}$ and $w^{(y)}$. The $n \times 1$ vector x_1 is the initial state, taken from a normal distribution with mean vector g_1 and covariance matrix Q_1. The $n \times 1$ vector g can be considered as a driving constant force, while $m \times 1$ vector μ is close to the global mean of the observation vectors. The state x is an n-dimensional vector which evolves according to linear difference equation (1b), with initial condition g_1. The state cannot be observed directly. Instead, m-dimensional measurements y are available at discrete sampling times as described by (1c). The vectors $w^{(x)}$ and $w^{(y)}$ are called state evolution noise and observation noise, respectively, and are independent to each other.

The parameters and the hidden state in the above equations can be jointly estimated with the Expectation Maximization (EM) algorithm [8], [9]. The EM iteration alternates between performing an expectation E-step, which estimates the hidden state sufficient statistics given both the observations and the parameter values, and a maximization M-step, which calculates the parameters using the sufficient statistics from the E-step.

B. Modelling an utterance with LDMs

Typically, in parametric speech synthesis, an utterance is described as a concatenation of context-dependent phonetic

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2016.2580672, IEEE Signal Processing Letters

units. A five-state left-to-right HMM is usually used to model each of these units, where each HMM state corresponds to a segment of the phonetic unit called sub-phoneme unit. In this work, each phoneme is split into three equally-sized segments (left-middle-right segment) and each of these segments is modelled with an LDM. Assuming that information on sub-phoneme segments of an utterance u and on their context-dependent labelling is available, we can define the sequence of segments of u as $seg(u)$. If each segment $\varsigma \in seg(u)$ has a duration of T_ς, then the total duration (in frames) T_u of the utterance u is given by:

$$T_u = \sum_{\varsigma \in seg(u)} T_\varsigma \tag{2}$$

Let $Y = [y_1, \ldots, y_{T_u}] = y_{1:T_u}$ be a trajectory of speech parameters that has been synthesized by a sequence of LDMs corresponding to segments $\varsigma \in seg(u)$. The corresponding trajectory of LDM hidden states is $X = [x_1, \ldots, x_{T_u}] = x_{1:T_u}$.

To achieve high quality synthesized speech, it is important to robustly model the acoustic and linguistic context. We, therefore, build a different LDM-based top-down hierarchical decision tree for each of the sub-phoneme units [3]. In this work, we denote the linguistic-to-acoustic mapping as a function $q : labels \rightarrow acoustic\ features$ and the simpler notation $q(\varsigma)$ is adopted instead of $q(label(\varsigma))$, where $\varsigma \in seg(u)$ and $label(\varsigma)$ is the full context label of segment ς.

C. Maximum likelihood trajectory generation

In synthesis, given the parameters θ of an LDM, a time sequence of speech parameter vectors is determined by maximizing the LDM likelihood

$$p(Y|\theta) = \int_X p(Y, X|\theta) dX \tag{3}$$

The probability density functions $p(Y|\theta)$ and $p(Y, X|\theta)$ are multivariate Gaussians. Therefore, matrix $\hat{Y}$ that maximizes the marginal $p(Y|\theta)$ also maximizes the within integral joint distribution $p(Y, X|\theta)$.

$$[\hat{Y}, \hat{X}] = \arg\max_{Y,X} p(Y, X|\theta)$$

with

$$\hat{Y} = \arg\max_{Y} p(Y|\theta) \quad \text{and} \quad \hat{X} = \arg\max_{X} p(X|\theta)$$

In order to find $\hat{Y}$, first the optimum state sequence $\hat{X} = [\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_T]$, with T being the number of frames, is obtained, by maximizing $p(X|\theta)$ with respect to X.

$$\hat{X} = \arg\max_{X} \mathcal{N}(x_1; g_1, Q_1) \prod_{t=2}^{T} \mathcal{N}(x_t; F x_{t-1} + g, Q) \tag{4}$$

Since the maximum likelihood estimate of a Gaussian is its mean, the state sequence estimate is obtained by evolving in time the following set of equations:

$$\hat{x}_1 = g_1 \tag{5}$$

$$\hat{x}_t = F\hat{x}_{t-1} + g, \quad t \in \{2, \ldots, T\} \tag{6}$$

Then, given the optimum state sequence estimate $\hat{X}$ we have

$$p(Y|\hat{X}, \theta) = \prod_{t=1}^{T} p(y_t|\hat{x}_t, \theta) = \prod_{t=1}^{T} \mathcal{N}(y_t; H\hat{x}_t + \mu, R) \tag{7}$$

and the maximum of $p(Y|\hat{X}, \theta)$ is attained when

$$\hat{y}_t = H\hat{x}_t + \mu, \quad t \in \{1, 2, \ldots, T\} \tag{8}$$

D. Utterance-level trajectory generation

Equations 5, 6 and 8 show how a trajectory of speech parameters is generated by one LDM model. To synthesize speech, the natural language processing module of a TTS system produces a phonetic transcription of the text input, as well as information concerning the desired intonation and rhythm and encodes this information into a sequence of labels. These labels are then associated to a sequence of LDM models through a linguistic-to-acoustic mapping. A trajectory of speech parameters for an utterance u is then produced as the concatenation of the trajectories generated from each of the LDMs in the sequence. Let $\theta_q = \{g_{1q}, Q_{1q}, F_q, g_q, Q_q, H_q, \mu_q, R_q\}$ be the parameters of the model with index q. Then $\Theta = \{\theta_q, \forall q\}$ denotes the parameters of all LDMs and Θ_u is the set of the parameters of the acoustic models of all segments of a utterance u, i.e., $\Theta_u = \{\theta_{q(\varsigma)} \mid \varsigma \in seg(u)\}$.

Algorithm 1 summarizes the utterance level trajectory generation. The parameter ρ has a value in $[0, 1]$ and is used to balance the requirement between the optimal modelling of one segment and the continuity constraint between the trajectories of two adjacent segments. When ρ is equal to 1 then the trajectory of the current segment is generated according to the maximum likelihood criterion without taking into account the previous segment. When ρ is equal to 0, then the final state of the previous segment is used to initialize the state of the current segment. If $\rho < 1$ then in order to avoid discontinuities matrix H should be globally tied.

Algorithm 1 LDM trajectory generation for an utterance u

```
function Y = LDM_SYNTHESIS(Θ_u, seg(u), ρ)
    τ = 0
    for ς ∈ seg(u) do
        q = q(ς)
        if τ = 0 then
            x_t = g_1q
        else
            x_t = ρ g_1q + (1 − ρ) x_t
        end if
        for t = 1 : T_ς do
            τ = τ + 1
            Y(:, τ) = H_q x_t + μ_q
            x_t = F_q x_t + g_q
        end for
    end for
end function
```

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2016.2580672, IEEE Signal Processing Letters

III. Global Variance based Parameter Generation

A. Definition of global variance modelling

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

$$v = [v(1), \ldots, v(k), \ldots, v(m)]^\top \tag{9}$$

where

$$v(k) = \frac{1}{T_u} \sum_{t=1}^{T_u} (y_t(k) - \bar{y}(k))^2 \tag{10}$$

and

$$\bar{y}(k) = \frac{1}{T_u} \sum_{t=1}^{T_u} y_t(k) \tag{11}$$

The distribution of GV is modelled as a single Gaussian distribution

$$\mathcal{N}(v; \mu_v, \Sigma_v) = \prod_{k=1}^{m} \mathcal{N}(v(k); \mu_v(k), \Sigma_v(k,k)) \tag{12}$$

which is estimated from the GV vectors of the training sentences. The covariance matrix Σ_v is diagonal since the GV of each dimension is calculated independently of other dimensions. In the following sections a method to apply GV to LDM synthesized trajectories is described.

B. Global variance constrained LDM synthesis

The Global Variance Constrained (GVC) approach performs joint optimization of the LDM likelihood and the likelihood of the GV. Specifically, instead of maximizing the LDM likelihood (3), we maximize the following:

$$p(Y|\Theta_u, \theta_v) = \int_X p(Y|X, \Theta_u, \theta_v) p(X|\Theta_u) dX \tag{13}$$

where, the parameters Θ_u of the LDMs and the parameters θ_v of the GV Gaussian distribution are independently trained from the speech corpus. Here, it is assumed that the distribution of X is independent of the parameter θ_v and that the probability density function $p(Y|X, \Theta_u, \theta_v)$ is written as a product of experts [10].

$$P(Y|X, \Theta_u, \theta_v) = \frac{1}{Z} p(Y|X, \Theta_u) p(v(Y)|\theta_v)^{\omega T_u} \tag{14}$$

where $p(v(Y)|\theta_v)$ is defined by the Gaussian distribution of Eq. (12) and Z is a normalizing constant, which is ignored in maximization. The constant ω is a weight controlling the contribution of the LDMs and GV likelihoods. In HMM speech synthesis, ω is usually set to one. LDM speech synthesis experiments have shown that setting $\omega = 1$ is also a good choice, producing natural speech sounds without artifacts.

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.

$$L = \log \left(p(Y|\hat{X}, \Theta_u) p(v(Y)|\theta_v)^{\omega T_u} \right) \Rightarrow$$

$$L = -\frac{1}{2} \sum_{\varsigma \in seg(u)} \left(\sum_{t=1}^{T_\varsigma} (y_{\varsigma t} - \hat{y}_{\varsigma t})^\top R_q^{-1} (y_{\varsigma t} - \hat{y}_{\varsigma t}) \right) - \frac{\omega T_u}{2} (v - \mu_v)^\top \Sigma_v^{-1} (v - \mu_v) + \text{const} \tag{15}$$

where $\hat{y}_{\varsigma t} = H_q \hat{x}_{\varsigma t} + \mu_q$ is the trajectory produced by Eq. (8) and $q = q(\varsigma)$ is the index of the model that corresponds to the label of segment ς. The index t of $y_{\varsigma t}$ refers to the position of vector y within the segment ς.

In order to determine a Y that maximizes L, the derivative $\frac{\partial L}{\partial Y}$ is calculated. The calculations are simplified by the fact that the observations y_t, $t = 1, \ldots, T_u$ are independent of each other given the trajectory of hidden states $\hat{x}_t$, $t = 1, \ldots, T_u$. The derivative of the first term, L_1, of L with respect to y_τ is

$$\frac{\partial L_1}{\partial y_\tau} = -R_q^{-1} (y_\tau - \hat{y}_\tau) \tag{16}$$

The derivative of the second term of L with respect to y_τ is

$$\frac{\partial L_2}{\partial y_\tau} = -2\omega \Sigma_v^{-1} (v - \mu_v) \odot (y_\tau - \bar{y}) \tag{17}$$

where $\odot$ denotes the element-wise multiplication of two vectors. Then

$$\frac{\partial L}{\partial y_\tau} = \frac{\partial L_1}{\partial y_\tau} + \frac{\partial L_2}{\partial y_\tau} \quad \text{and} \quad \frac{\partial L}{\partial Y} = \left[\frac{\partial L}{\partial y_1}, \ldots, \frac{\partial L}{\partial y_{T_u}} \right]^\top \tag{18}$$

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

$$Y^{(i+1)\text{-th}} = Y^{(i)\text{-th}} + \alpha \frac{\partial L}{\partial Y} \tag{19}$$

where α is a step size parameter. Algorithm 2 implements the gradient ascent method. There are two settings of the initial trajectory $Y^{(0)\text{-th}}$. One is to use the trajectory calculated by (8). The other is to use the trajectory Y' linearly converted from the conventional one using variance scaling [11]. Both choices of the initial trajectories converge to the same final trajectory provided that the parameters α_0, ϵ_1 and ϵ_2 have been chosen so that Algorithm 2 produces trajectories of the speech parameters with as few artifacts as possible (see for example, the work of Shannon et al. [12] for a discussion on the artifacts that the GV generation algorithm may create). A drawback of Algorithm 2 is that it processes in batch mode the whole utterance while one of the advantages of LDMs is that they have on-line parameter generation algorithm (Alg.1).

C. Computational complexity

The operation $S_{inv} = (\Sigma_v + \varepsilon_1 I)^{-1}$ requires $2m$ floating point operations since matrix Σ_v is diagonal. For each loop of the iterations the number of floating point operations are:

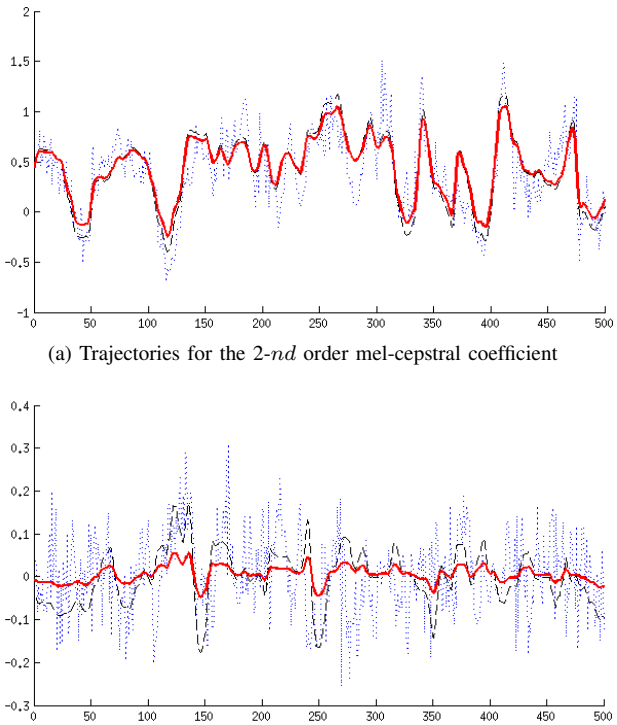
- $3mT$ operations for the $mean$ and var of Y
- $2m$ operations for the calculations of $B = S_{inv}(v - \mu_v)$
- $2m \times numSegments$ operations for $B = S_{inv}(v - \mu_v)$
- $5mT$ operations for the calculation of dL

Therefore, in total the GV Constrained LDM synthesis algorithm requires $\leq 10mT \times numIterations$ floating point operations.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2016.2580672, IEEE Signal Processing Letters

Algorithm 2 GV Constrained LDM synthesis

function Y = GV_IN_SYNTHESIS($\hat{Y}, \Theta_u, \theta_v, seg(u)$)
  $\alpha_0 = 0.001$
  $\omega = 1$
  $S_{inv} = (\Sigma_v + \varepsilon_1 I)^{-1}$ ▷ e.g., $\varepsilon_1 = 0.00002$
  $Y = \hat{Y}$
  for $i = 1 : numIterations$ **do**
    $\alpha = \alpha_0/\sqrt{i}$
    $v = var(Y, 2)$
    $\bar{y} = mean(Y, 2)$
    $B = S_{inv}(v - \mu_v)$
    $\tau = 0$
    for $\varsigma \in seg(u)$ **do**
      $q = q(\varsigma)$
      $R_{inv} = (R_q + \varepsilon_2 I)^{-1}$ ▷ e.g., $\varepsilon_2 = 0.0001$
      for $t = 1 : T_\varsigma$ **do**
        $\tau = \tau + 1$
        $y_\tau = Y(:, \tau)$; $\hat{y}_\tau = \hat{Y}(:, \tau)$
        $dL(:, \tau) = -R_{inv}(y_\tau - \hat{y}_\tau) - 2\omega B \odot (y_\tau - \bar{y})$
      end for
    end for
    $Y = Y + \alpha dL$;
  end for
end function



(a) Trajectories for the 2-nd order mel-cepstral coefficient

(b) Trajectories for the 34-th order mel-cepstral coefficient

Fig. 1. Trajectories for the 2-nd and 34-th order mel-cepstral coefficients. Blue dot-dot line: original. Red continuous thick line: LDM generated. Black dash-dash line: GV applied.

IV. Experiments

A. Conditions

A database containing 4417 sentences (approximately 5 hours of speech) of an American English female speaker was used to verify the effectiveness of the proposed GV-based speech parameter method. The audio was recorded in studio and the sampling frequency was down-sampled to 22.05 kHz for our experimental purposes. Full context labels were created by using a proprietary front-end. From the training utterances, 40 mel-cepstral coefficients and 39 phase features were extracted at every 5 ms as follows. First, a proprietary tool was utilized to detect pitch period onsets from speech. Then, pitch-synchronous spectral analysis was conducted, followed by interpolation of amplitude and phase spectra at the frame level, where frame-based amplitude and phase spectra were then converted into complex cepstra [13]. Finally, complex cepstrum vectors were applied to the method presented in [14] in order to derive the final mel-cepstral coefficients and phase features. As for the remaining speech parameters, $\ln F0$ was extracted using the SWIPE algorithm [15], while 20 mel-band-aperiodicity parameters were extracted based on the speech decomposition method shown in [16]. In our experiments, GV is applied for mel-cepstral coefficients only, since the intention is to remove the muffled speech quality. The LDMs were trained as described in [3]. At synthesis time, LDM-generated parameters were applied to the synthesis engine shown in [13] to produce speech.

B. Results

A forced A-B preference test was conducted using 24 test sentences, with durations varying between 1.3–9.2s (mean duration 3.7s), and number of words ranging from 3 to 23 (in average 9 words per sentence). Fifty four listeners took part in the test, where 11 of them were speech processing specialists and 43 had professions that are not related to speech processing. Each test sentence was synthesized in two versions: (1) LDM synthesis without GV (Algorithm 1); (2) LDM synthesis with GV (Algorithm 2). The utterance versions were played in changing orders. 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. This shows that the proposed method is highly effective in improving synthetic speech quality.

As an example, Figures 1a and 1b show parts of original trajectories of the 2-nd and 34-th order mel-cepstral coefficients and the corresponding LDM synthesized trajectories with and without applying the GV algorithm. It can be seen that the proposed algorithm generates trajectories which are closer to that extracted from natural speech. It can also 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.

V. Conclusion

We proposed an algorithm for applying GV on LDM-based speech synthesis. The algorithm modifies the likelihood of the LDMs to take into account the GV of the parameters of the original utterances. According to subjective preference tests this method greatly improves the naturalness of the synthesized speech, at an additional computational cost.

REFERENCES

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statistical parametric speech synthesis," *IEEE Transcations on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 587–597, 2013.

[3] V. Tsiaras, R. Maia, V. Diakoloukas, Y. Stylianou, and V. Digalakis, "Towards a linear dynamical model based speech synthesizer," in *Proceedings of the Annual Conference of the International Speech Communication Assiciation, Interspeech*, 2015, pp. 1221–1225.

[4] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2013, pp. 7962–7966.

[5] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2015, pp. 4470–4474.

[6] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[7] T. Toda, "Modeling of speech parameter sequence considering global variance for HMM-based speech synthesis," in *Hidden Markov Models, Theory and Applications*, P. Dymarski, Ed. InTech, 2011, ch. 6, pp. 131–150.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, pp. 1–38, 1977. [Online]. Available: http://web.mit.edu/6.435/www/Dempster77.pdf

[9] V. Digalakis, J. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 4, pp. 431–442, 1993.

[10] H. Zen, M. Gales, Y. Nankaku, and K. Tokuda, "Product of experts for statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 794–805, 2012.

[11] H. Silén, E. Hel, J. Nurminen, and M. Gabbouj, "Ways to implement global variance in statistical speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Assiciation, Interspeech*, 2012, pp. 1436–1439.

[12] M. Shannon and W. Byrne, "Fast, low-artifact speech synthesis considering global variance," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2013, pp. 7869–7873.

[13] R. Maia, M. Akamine, and M. F. J. Gales, "Complex cepstrum as phase information for statistical parametric speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2012*, 2012, pp. 4581–4584.

[14] R. Maia and Y. Stylianou, "Iterative estimation of phase using complex cepstrum representation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2016, pp. 4990–4994.

[15] A. Camacho, "A sawtooth waveform inspired pitch estimator for speech and music," Ph.D. dissertation, University of Florida, 2007.

[16] P. J. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced turbulence-noise components in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 713–726, 2001.