Statistical Parametric Speech Synthesis with Joint Estimation of Acoustic and Excitation Model Parameters

Ranniery Maia, Heiga Zen, M. J. F. Gales

Toshiba Research Europe Ltd., Cambridge Research Laboratory, Cambridge, UK
{ranniery.maia,heiga.zen,mjfg}@crl.toshiba.co.uk

Abstract
This paper describes a novel framework for statistical parametric speech synthesis in which statistical modeling of the speech waveform is performed through the joint estimation of acoustic and excitation model parameters. The proposed method combines extraction of spectral parameters, considered as hidden variables, and excitation signal modeling in a fashion similar to factor analyzed trajectory hidden Markov model. The resulting joint model can be interpreted as a waveform level closed-loop training, where the distance between natural and synthesized speech is minimized. An algorithm based on the maximum likelihood criterion is introduced to train the proposed joint model and some experiments are presented to show its effectiveness.

Index terms: statistical parametric speech synthesis, trajectory hidden Markov model, excitation modeling, factor analysis.

1. Introduction
In typical statistical parametric speech synthesis [1], speech parameters are extracted from speech waveforms and their trajectories are modeled by a statistical model, such as a hidden Markov models (HMMs). The parameters of the models are estimated so as to maximize their likelihood given the training data. At the synthesis stage, a sentence-level sequence of states of the trained statistical models is composed according to an input text, and then speech parameters are generated so as to maximize the output probabilities of such states [1]. Finally, a speech waveform is re-constructed from the generated speech parameters by assuming the source-filter production model [2].

In this paper a novel approach to statistical parametric speech synthesis is proposed, in which acoustic model parameters are jointly estimated with parameters of a stand-alone excitation model in a way to maximize a likelihood function including both models given the speech waveform. This is done by assuming that speech can be represented by the convolution of a slowly varying vocal tract impulse response filter derived from spectral parameters, and an excitation source. In the proposed approach spectral parameter extraction is integrated in the joint training of acoustic and excitation models, in a similar way to the spectral analysis based on factor analyzed trajectory HMM of [3]. However, in the present method the speech waveform is the observed term, and its relationship with the spectral features, considered as a hidden variable, is obtained via the excitation model parameters. The proposed joint estimation of acoustic and excitation models based on the maximum likelihood (ML) criterion can be viewed as a closed-loop training method for statistical parametric synthesis, such as the one introduced in [4]. Indeed, one similarity between the current method with the one described in [4] is the integration of spectral extraction during the training. Nevertheless, the philosophy of the approach described in [4] is the explicit minimization of a distortion of amplitude spectra of natural and synthesized speech, whereas for the present case the distance between natural and synthesized speech waveforms is minimized in the time domain.

This paper is organized as follows. In Section 2 the basic idea is introduced; in Section 3 the joint modeling of acoustic and excitation models is defined; Section 4 presents a training algorithm based on the ML criterion; Section 5 shows some experiments, and concluding remarks are in Section 6.

2. Basic idea
2.1. Conventional framework
In a typical statistical parametric speech synthesizer [1], firstly a spectral parameter vector \( \mathbf{c} = [c_0 \cdots c_{T-1}]^\top \) is extracted from the speech waveform, where \( c_t = [c_t(0) \cdots c_t(C)]^\top \) is a \( C \)-th order spectral feature vector at frame \( t \), and \( T \) is the total number of frames. Estimation of acoustic model parameters is usually done through the ML criterion, i.e.,

\[
\hat{\lambda}_c = \arg \max_{\lambda_c} p(\mathbf{c} \mid \ell, \lambda_c),
\]

where \( \ell \) is a transcription of the speech waveform and \( \lambda_c \) denotes a set of acoustic model parameters.

At runtime synthesis, \( \mathbf{c} \) is generated for a given text to be synthesized \( \ell \) so as to maximize its output probability

\[
\hat{\mathbf{c}} = \arg \max_{\mathbf{c}} p(\mathbf{c} \mid \ell, \hat{\lambda}_c).
\]

These generated features together with an \( F_0 \)-generated excitation signal are utilized to synthesize the speech waveform by using the source-filter production approach [2].

2.2. Proposed framework
Since the intention of any speech synthesizer is to mimic the speech waveform as well as possible, a statistical model defined at the waveform level is proposed here. The parameters of this new model are estimated so as to maximize its likelihood given the speech waveform itself, i.e.,

\[
\hat{\mathbf{c}} = \arg \max_{\mathbf{c}} p(\mathbf{s} \mid \ell, \hat{\lambda}),
\]

where \( \mathbf{s} = [s(0) \cdots s(N-1)]^\top \) is a vector containing the speech waveform, with \( s(n) \) being a waveform value at sample \( n \), \( N \) the number of samples, and \( \hat{\lambda} \) denoting the set of parameters of the joint acoustic-excitation model.

By introducing two hidden variables: the state sequence \( q = \{q_0, \ldots, q_{T-1}\} \) (discrete); and spectral parameter \( \mathbf{c} = \)}
where \(q_t\) is the state at frame \(t\). The meanings of the terms \(p(s \mid c, q, \lambda)\), \(p(c \mid q, \lambda)\), and \(p(q \mid \ell, \lambda)\) in (5) are separately analyzed as follows:

- \(p(s \mid c, q, \lambda)\): this term concerns the speech waveform generation from spectral features and a given state sequence. Its maximization with respect to \(\lambda\) is closely related to the ML spectral estimation [5]. In this work this probability is related to the assumed speech signal generative model\(^1\).

- \(p(c \mid q, \lambda)\): this term is given as the product of state-output probabilities of speech parameter vectors if HMMs or hidden semi-Markov models (HSMMs) [1] are used for an acoustic model. If trajectory HMMs [6] are used, this probability is given as a state-sequence-output probability of the spectral parameter vector.

- \(p(q \mid \ell, \lambda)\): this term represents the probability of a state sequence \(q\) for a given transcription \(\ell\). If an HMM or trajectory HMM is used for acoustic modeling, this probability is given as a product of state-transition probabilities. In case HSMM or trajectory HSMM is used, \(p(q \mid \ell, \lambda)\) includes both state-transition and state-duration probabilities.

Note that it is possible to model \(p(c \mid q, \lambda)\) and \(p(q \mid \ell, \lambda)\) using existing acoustic models, such as HMM, HSMM, or trajectory HMM. Therefore, the problem is how to model \(p(s \mid c, q, \lambda)\).

3. Definition of the proposed framework

This section describes the distribution of the speech waveform given spectral parameters, \(p(s \mid c, q, \lambda)\). Initially, this is described in terms of the impulse response of a vocal tract filter. Later, in Section 3.2, the relationship between spectral parameters and vocal tract filter impulse response is then considered.

3.1. Assumed speech generative model

The speech signal is assumed to be generated by the process in Figure 1 [7], where

\[
s(n) = h_v(n) \ast [h_t(n) \ast t(n) + h_u(n) \ast w(n)],
\]

with \(*\) denoting linear convolution, and

- \(h_v(n)\): being the vocal tract filter impulse response;
- \(t(n)\): being a pulse train;
- \(w(n)\): being a Gaussian white noise sequence with mean zero and variance one;
- \(h_u(n)\): being the voiced filter impulse response;
- \(h_u(n)\): being the unvoiced filter impulse response.

The vocal tract, voiced and unvoiced filters are assumed to have the following transfer functions,

\[
H_v(z) = \sum_{p=0}^{P} h_v(p) z^{-p},
\]

\[
H_u(z) = \sum_{m=-M}^{M} h_u(m) z^{-m},
\]

\[
H_o(z) = \frac{K}{1 - \sum_{l=1}^{L} g(l) z^{-l}},
\]

where \(P, M, L\) are respectively the orders of \(H_v(z)\), \(H_u(z)\), and \(H_o(z)\). Filter \(H_o(z)\) is considered to have minimum-phase response since it represents the impulse response of the vocal tract filter [2]. Parameters of the generative model above comprise the vocal tract, voiced and unvoiced filters, \(H_v(z)\), \(H_u(z)\) and \(H_o(z)\), and the positions and amplitudes of \(t(n), \{p_0, \ldots, p_z-1\}\), and \(\{a_0, \ldots, a_z-1\}\), with \(Z\) being the number of pulses. Although there may exist several ways to estimate coefficients \(h_v(m)\) and \(g(l)\) and gain term \(K\), of filters \(H_v(z)\) and \(H_o(z)\), respectively, this paper will be based on the method described in [7].

Using matrix notation, with uppercase and lowercase capital letters respectively denoting matrices and vectors, (6) can be written as

\[
s = H_v h_v t + s_u,
\]

where

\[
s = \begin{bmatrix}
{s(0)} & \cdots & {s(N + \frac{M}{2} + P - 1)}
\end{bmatrix}^T,
\]

\[
H_v = \begin{bmatrix}
\tilde{h}_v(0) & \cdots & \tilde{h}_v(N-M-1)
\end{bmatrix},
\]

\[
\tilde{h}_v(i) = \begin{bmatrix}
0 & \cdots & h_v(0) & \cdots & h_v(P) & 0 & \cdots & 0
\end{bmatrix}^T_{N+M-n-1},
\]

\[
h_u = \begin{bmatrix}
\tilde{h}_u(0) & \cdots & \tilde{h}_u(N-1)
\end{bmatrix},
\]

\[
\tilde{h}_u(i) = \begin{bmatrix}
0 & \cdots & 0 & h_u(-\frac{M}{2}) & \cdots & h_u(\frac{M}{2}) & 0 & \cdots & 0
\end{bmatrix}^T_{N-n-1},
\]

\[
t = \begin{bmatrix}
t(0) & \cdots & t(N-1)
\end{bmatrix}^T,
\]

\[
s_u = \begin{bmatrix}
0 & \cdots & 0 & s_u(0) & \cdots & s_u(N + L - 1) & 0 & \cdots & 0
\end{bmatrix}^T_{\frac{M}{2} + P - L}.
\]

The vector \(s_u\) contains samples of

\[
s_u(n) = h_c(n) \ast h_u(n) \ast w(n),
\]

and can be interpreted as the error of the generative model for voiced regions of the speech signal with covariance matrix \(H_v \Phi H_v^T\), where

\[
\Phi = \begin{bmatrix} \begin{pmatrix} G \end{pmatrix}^T \end{pmatrix},
\]

\[
G = \begin{bmatrix}
\tilde{g}(0) & \cdots & \tilde{g}(N+M-1)
\end{bmatrix},
\]

\[
\tilde{g}^{(i)} = \begin{bmatrix}
0 & \cdots & 0 & \frac{1}{n} & \frac{\alpha(1)}{n} & \cdots & \frac{\alpha(n)}{n} & 0 & \cdots & 0
\end{bmatrix}^T_{N+M-1}.
\]
Figure 1: The assumed speech generation model. Excitation is created by state-dependent filtering of a pulse train and Gaussian white noise [7].

As \( w(n) \) is Gaussian white noise, \( u(n) = h_w(n)*w(n) \) is a normally distributed stochastic process. Using vector notation, the probability of \( u \) can be expressed as

\[
p(u | G) = N(u; 0, \Phi),
\]

where \( N(x; \mu, \Sigma) \) means a Gaussian distribution of \( x \) with mean vector \( \mu \) and covariance matrix \( \Sigma \). And since

\[
u(n) = H_c^{-1}(z) \{ s(n) - h(n) * h_w(n) * t(n) \},
\]

where \( h(n) \) is the probability of speech vector \( s \) becomes

\[
p(s | H_c, H_c, G, t) = N \left( s; H_c H_c t, H_c \Phi H_c^\top \right).
\]

If the last \( P \) rows of \( H_c \) are ignored, thus neglecting the zero-impulse response of \( H_c(z) \) which produces samples \( \{ s(N + \tfrac{M}{2}) \}, \ldots, s(N + \tfrac{M}{2} + P - 1) \}, \) then \( H_c \) becomes square with dimension \( N + M \) and (24) can be re-written as

\[
p(s | H_c, \lambda_c) = |H_c|^{-1} N \left( H_c^{-1} s; H_c t, \Phi \right),
\]

where \( \lambda_c = \{ H_c, G, t \} \) are parameters of the excitation part of the complete speech generative model. It is interesting to note that the term \( H_c^{-1}s \) corresponds to the residual sequence, extracted from the speech signal \( s(n) \) through inverse filtering by \( H_c(z) \).

By assuming that \( H_c \) and \( G \) in the speech generative model have the state-dependent parameter tying structure proposed in [7], (25) can be re-written as

\[
p(s | H_c, q, \lambda_c) = |H_c|^{-1} N \left( H_c^{-1} s; H_{v,q} t, \Phi_q \right),
\]

where \( H_{v,q} \) is the voiced filter impulse response matrix for state sequence \( q \), and \( \Phi_q = (G_q^\top G_q)^{-1} \), with \( G_q \) being the inverse unvoiced filter impulse response matrix for state sequence \( q \).

3.2. Relationship between vocal tract filter impulse response and spectral parameters

The previous section has derived the distribution \( p(s | H_c, q, \lambda) \). However, from (5) the complete waveform distribution requires \( p(s | c, q, \lambda) \). Therefore, a relationship between vocal tract impulse response represented here by the matrix \( H_c \), and the corresponding spectral parameters \( c \) has to be derived.

Depending on the relationship between \( H_c \) and \( c \), it is often difficult to compute \( H_c \) from \( c \) in a closed form for some choices of spectral parameters, such as mel-cepstral coefficients, line spectral pairs, etc. To address this problem, a stochastic approximation to model the relationship between \( H_c \) and \( c \) is one possibility. If the mapping between \( H_c \) and \( c \) can be represented as a Gaussian process with probability

\[
p(H_c | c, q, \lambda_c),
\]

where \( \lambda_c \) is the set of parameters of a Gaussian model that maps spectral parameters onto vocal tract filter impulse response, then

\[
p(s | c, q, \lambda_c) = \int p(s | H_c, q, \lambda_c) p(H_c | c, q, \lambda_b) dH_c,
\]

\[
= |H_c|^{-1} N \left( H_c^{-1} s; H_{v,q} t, \Phi_q \right) N \left( H_c; f_q(c), \Omega_q \right) dH_c,
\]

where \( f_q(c) \) is an approximation function to convert \( c \) to \( H_c \) and \( \Omega_q \) is a covariance matrix representing the noise of the conversion.

In the special case where the relationship between \( H_c \) and \( c \) is deterministic, \( f_q(c) \) becomes a simple mapping function in a closed form and

\[
p(H_c | c, q, \lambda_b) \text{ becomes a delta function.}
\]

3.2.1. Cepstral coefficients as spectral parameters

In this paper cepstral coefficients [5] were chosen as spectral parameters in order to avoid the mapping model discussed above. Therefore, as that the filter \( H_c(z) \) has minimum-phase response, the relationship between a given cepstral coefficient vector for frame \( t \), \( c_t = [c_t(0) \cdots c_t(C)]^\top \), and its corresponding vocal tract filter impulse response vector, \( h_{c,t} = [h_{c,t}(0) \cdots h_{c,t}(P)]^\top \), can be expressed as

\[
h_{c,t} = D_t^\top \text{EXP} [D_t c_t],
\]

where \( \text{EXP}[\cdot] \) means a vector which is derived by taking the exponential of the elements of \( [\cdot] \), and \( D_t \) is \( (P+1) \times (C+1) \) discrete Fourier transform (DFT) matrix,

\[
D_t = \begin{bmatrix}
1 & 1 & \cdots & 1 \\
1 & W_{P+1} & \cdots & W_{P+1}^C \\
\vdots & \vdots & \ddots & \vdots \\
1 & W_{P+1}^P & \cdots & W_{P+1}^{PC}
\end{bmatrix},
\]

with

\[
W_{P+1} = e^{-\frac{2\pi}{P+1}},
\]

and \( D_t^\top \) is a \( (P+1) \times (P+1) \) inverse DFT matrix

\[
D_t^\top = \frac{1}{P+1} \begin{bmatrix}
1 & 1 & \cdots & 1 \\
1 & W_{P+1}^{-1} & \cdots & W_{P+1}^{-P} \\
\vdots & \vdots & \ddots & \vdots \\
1 & W_{P+1}^{-P} & \cdots & W_{P+1}^{-P}
\end{bmatrix}.
\]

Note that \( D_t \) and \( D_t^\top \) are constant for all the frames.
3.2.2. Relationship between $H_c$ and $h_c$

The vocal tract filter impulse response-related term that appears in the generative model of (10) is $H_c$, not $h_c$. Relationship between $H_c$ given as (12) and (13) without the last $P$ rows (square version of $H_c$), and $h_c$ given as

$$h_c = \begin{bmatrix} h_{c,0}^T & \cdots & h_{c,T-1}^T \end{bmatrix}^T,$$

(32)

$$h_{c,t} = \begin{bmatrix} h_{c,t}^T(0) & \cdots & h_{c,t}^T(P) \end{bmatrix}^T,$$

(33)

with $h_{c,t}$ being the synthesis filter impulse response of the $t$-th frame, can be written as

$$H_c = \sum_{n=0}^{N-1} J_n B h_{c,n}^T.$$

(34)

In (34), $N$ is the number of samples, and

$$j_n = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}^T,$$

(35)

with $B$ being an $N(P+1) \times T(P+1)$ matrix to map a frame-based vector $h_c$ into its sample-based version. The $N \times N (P+1)$ matrices are constructed as follows

$$J_0 = \begin{bmatrix} I_{P+1} & 0_{P+1,N(P+1)-P-1} & 0_{N-P-1,P+1} & 0_{N-P-1,N(P+1)-P-1} \end{bmatrix},$$

(36)

$$J_{N-1} = \begin{bmatrix} 0_{N-1,N(P+1)-P-1} & 0_{N-1,P+1} & 0_{N-1,P} & 0_{1,P} \end{bmatrix},$$

(37)

where $0_{X,Y}$ means a matrix of zero elements with $X$ rows and $Y$ columns, and $I_X$ is an $X$-dimension identity matrix. For each sample increment the identity matrix $I_{P+1}$ moves one row down and $P+1$ columns to the right.

3.3. Joint acoustic and excitation modeling

If a trajectory HMM [6] is used to represent the probability $p(\epsilon \mid \ell, \lambda_c)$, then

$$p(\epsilon \mid q, \lambda_c) = N(\epsilon \mid \tilde{c}_q, P_q),$$

(38)

$$p(q \mid \ell, \lambda_c) = \pi_q \frac{\alpha_{q,T+1}^{T-1}}{\sum_i \alpha_{i,q_i+1}^{T-1}},$$

(39)

where $\pi_i$ is the initial probability of state $i$, $\alpha_{i,j}$ is the transition probability from state $i$ to state $j$, and $\tilde{c}_q$ and $P_q$ correspond to the mean vector and covariance matrix of trajectory HMM for $q$. In (38), $\tilde{c}_q$ and $P_q$ are given as

$$\tilde{c}_q = P_q r_q,$$

(40)

$$R_q = P_q^{-1} = W^T \Sigma_q^{-1} W,$$

(41)

$$r_q = W^T \Sigma_q^{-1} \mu_q,$$

(42)

where $W$ is typically a $3T(C+1) \times T(C+1)$ window matrix that appends dynamic features (velocity and acceleration features) to $c$. $\mu_q$ and $\Sigma_q^{-1}$ in (41) and (42) correspond to the $3T(C+1) \times 1$ mean parameter vector and the $3T(C+1) \times 3T(C+1)$ precision parameter matrix for the state sequence $q$, given as

$$\mu_q = \begin{bmatrix} \mu_{q,0}^T & \cdots & \mu_{q,T-1}^T \end{bmatrix}^T,$$

(43)

$$\Sigma_q^{-1} = \text{diag}\{\Sigma_{q,0}^{-1}, \cdots, \Sigma_{q,T-1}^{-1}\},$$

(44)

where $\mu_j$ and $\Sigma_j^{-1}$ correspond to the $3(C+1) \times 1$ mean-parameter vector and the $3(C+1) \times 3(C+1)$ precision-parameter matrix associated with state $j$, and $Y = \text{diag}\{X_1, X_2, \ldots\}$ means that matrices $\{X_1, X_2, \ldots\}$ are diagonal sub-matrices of $Y$. Mean parameter vectors and precision parameter matrices for each state $j$ are defined as

$$\mu_j = \begin{bmatrix} \mu_j^T & \Delta \mu_j^T & \Delta^2 \mu_j^T \end{bmatrix}^T,$$

(45)

$$\Sigma_j^{-1} = \text{diag}\{\Sigma_j^{-1}, \Delta \Sigma_j^{-1}, \Delta^2 \Sigma_j^{-1}\},$$

(46)

where $\Delta[\cdot]$ and $\Delta^2[\cdot]$ mean respectively a vector or matrix of velocity and acceleration features associated with vector or matrix $[\cdot]$.

The final model is obtained by combining the acoustic and excitation models, i.e.,

$$p(s \mid \ell, \lambda) = \sum_q p(s \mid c, q, \lambda_c) p(c \mid q, \lambda_c) p(q \mid \ell, \lambda_c) \text{dc},$$

(47)

where $p(c \mid q, \lambda_c)$ and $p(q \mid \ell, \lambda_c)$ are given by (38) and (39), respectively, whereas $p(s \mid c, q, \lambda_c)$ is represented by (26), and finally $\lambda = \{\lambda_c, \lambda_e\}$.

4. ML training

Parameters of the joint model $\lambda = \{\lambda_c, \lambda_e\}$ are estimated according to (3), where the likelihood function $p(s \mid \ell, \lambda)$ is given by (47), with $\lambda_e = \{H_c, G, T\}$ corresponding to parameters of the excitation model, and $\lambda_c = \{m, \sigma\}$ consisting of parameters of the acoustic model.

$$m = \begin{bmatrix} \mu_0^T & \cdots & \mu_{T-1}^T \end{bmatrix}^T,$$

(48)

$$\sigma = \text{vdiag}\{\Sigma_0^{-1}, \ldots, \Sigma_{T-1}^{-1}\},$$

(49)

where $S$ is the number of states, $m$ and $\sigma$ are respectively vectors formed by concatenating all the means and diagonals of the inverse covariance matrices of all states, with $\text{vdiag}\{[\cdot]\}$ meaning a vector formed by the diagonal elements of $[\cdot]$.

4.1. Likelihood function

Unfortunately, estimation of $\lambda$ through the expectation-maximization (EM) algorithm is intractable. Therefore, here an approximate recursive approach is adopted.

If the summation over all possible $q$ in (47) is approximated by a fixed state sequence, then $p(s \mid \ell, \lambda)$ becomes

$$p(s \mid \ell, \lambda) \approx \int p(s \mid c, \hat{q}, \lambda_c) p(c \mid \hat{q}, \lambda_c) p(\hat{q} \mid \ell, \lambda_c) \text{dc},$$

(50)

where $\hat{q} = \{\hat{q}_0, \ldots, \hat{q}_{T-1}\}$ is a fixed state sequence. Further, if the integration over all possible $c$ is approximated by single cepstral coefficient vector, then (50) becomes

$$p(s \mid \ell, \lambda) \approx p(s \mid \epsilon, \hat{q}, \lambda_c) p(\epsilon \mid \hat{q}, \lambda_c) p(\hat{q} \mid \ell, \lambda_c),$$

(51)

where $\epsilon = \{\epsilon_0, \ldots, \epsilon_{T-1}\}$ is a fixed spectral parameter vector. Finally, if the state sequence $\hat{q}$ is fixed through the entire training process, the term $p(\hat{q} \mid \ell, \lambda_c)$ can be ignored. By taking the logarithm of the resulting expression, the following log likelihood function to be maximized through the update of the acoustic and excitation model parameters can be obtained

$$\mathcal{L}(s; \epsilon, \hat{q}, \lambda_e, \lambda_c) = \log p(s \mid \epsilon, \hat{q}, \lambda_c) + \log p(\epsilon \mid \hat{q}, \lambda_c),$$

(52)
4.2. Training procedure
The optimization problem is broken into two parts: initialization and recursion.

4.2.1. Initialization
1. Extract from each utterance of the speech data an initial cepstral coefficient vector,
   \[ c = [c_0^T \cdots c_{T-1}^T]^T, \quad (53) \]
   \[ c_t = [c_t(0) \cdots c_t(T)]^T. \quad (54) \]
2. Train trajectory HMM parameters \( \lambda_c \) by using \( c \),
   \[ \hat{\lambda}_c = \arg \max_{\lambda_c} p(c | \lambda_c). \quad (55) \]
3. Determine the best state sequence \( \hat{q} \) as the Viterbi path from the trained models by using the algorithm shown in [6],
   \[ \hat{q} = \arg \max_{q} p(c, q | \lambda_c). \quad (56) \]
4. Estimate excitation parameters \( \lambda_e \) assuming \( \hat{q} \) and \( c \), by using the algorithm described in [7],
   \[ \hat{\lambda}_e = \arg \max_{\lambda_e} p(s | e, \hat{q}, \lambda_e). \quad (57) \]

4.2.2. Recursion
1. Estimate the best spectral parameter vector \( \hat{c} \) by using the log likelihood function of (52),
   \[ \hat{c} = \arg \max_{c} \mathcal{L}(s; c, \hat{q}, \lambda_e, \hat{\lambda}_c). \quad (58) \]
2. Update the acoustic model parameters by training trajectory HMM using \( \hat{c} \) as the observation,
   \[ \hat{\lambda}_c = \arg \max_{\lambda_c} p(\hat{c} | \hat{q}, \lambda_c). \quad (59) \]
3. Update excitation model parameters \( \hat{\lambda}_e \) through the algorithm described in [7], assuming \( \hat{q} \) and \( \hat{c} \),
   \[ \hat{\lambda}_e = \arg \max_{\lambda_e} p(s | e, \hat{q}, \hat{\lambda}_e). \quad (60) \]

The recursive steps may be repeated several times.

4.2.3. Estimation of the best cepstral vector \( \hat{c} \)
In Step 1 of the recursive process, cepstral coefficients are estimated given both trained excitation and acoustic models.

The log likelihood function of (52) can be written as
\[
\mathcal{L}(s; e, \hat{q}, \lambda_e, \lambda_c) = -\frac{1}{2} s^T H_c^{-\top} \Phi_q^{-1} H_c^{-1} s + \log |H_c| + s^T H_c^{-\top} \Phi_q^{-1} H_c^{-1} t - \frac{1}{2} \hat{c}^T R_q \hat{c} + r_q^T \hat{c} + \mathcal{K},
\]
where \( \mathcal{K} \) is a constant that does not depend on \( e \). The best cepstral coefficient vector \( \hat{c} \) can be calculated by utilizing any gradient-based optimization algorithm [8], where the gradient of \( \mathcal{L}(s; e, \hat{q}, \lambda_e, \lambda_c) \) with respect to \( e \) is
\[
\nabla_e \mathcal{L}(s; e, \hat{q}, \lambda_e, \lambda_c) = D^T \text{DIAG} \left( \text{EXP} \{ D[c] \} \right) D^T B^T
\sum_{n=0}^{N-1} J_n H_c^{-\top} \left[ \Phi_q^{-1} (e - v) e^\top - I_N \right] J_n - R_q \hat{c} + r_q,
\]
where \( D = \text{diag} \{ D_0, \ldots, D_{P-1} \} \), \( D^* = \text{diag} \{ D_0^*, \ldots, D_{P-1}^* \} \), \( r_q \) is the residual of the cepstral coefficient.\footnote{\( F_0 \) information was utilized to extract smooth periodograms.}

5. Experiment
A preliminary experiment was conducted to verify the effectiveness of the proposed joint estimation process. The database consisted of 100 hand-labeled sentences from a US English female speaker.

Initially, cepstral coefficients were calculated by performing the spectral analysis described in [5], with \( \gamma = 0 \) and \( \alpha = 0 \), on smooth periodograms\footnote{\( F_0 \) information was utilized to extract smooth periodograms.} as input, extracted from the speech data at every 5 ms. The number of cepstral coefficients calculated per frame was 40, i.e., \( C = 39 \).

Joint training of acoustic and excitation models \( \lambda = \{ \lambda_e, \lambda_c \} \) was then conducted as described in Section 4.2 and illustrated in the diagram of Figure 2. For estimation of the best cepstral coefficient vector \( \hat{c} \) a 256-th order impulse response of \( H_c(z) \) (\( P = 256 \)) was considered. The recursive process was repeated 4 times.

5.1. Cepstral analysis given acoustic and excitation models
Estimation of the best cepstral coefficients for each iteration of the recursive procedure was performed as described in the Section 4.2.3. Figure 3 shows the behavior of the e-dependent part of the likelihood \( \mathcal{L}(s; e, \hat{q}, \lambda_e, \lambda_c) \) for one sentence, when a simple steepest-ascent algorithm was utilized with convergence.
Figure 3: Behavior of the log likelihood $\mathcal{L}(\theta; c, q, \lambda_c, \lambda_e)$ during the estimation of the best cepstrum $\hat{c}$ for one sentence, using a simple steepest-ascent algorithm with update factor $\delta = 0.01$. These measures were taken from the first iteration of the recursive process illustrated in Figure 2.

Figure 4: Short-term amplitude spectra of a given sentence. Aside from natural and smooth spectra, spectra derived respectively from the baseline cepstra calculated in the initial analysis, and cepstra calculated in the 1-best estimation process of the fourth recursive iteration are also shown. The performance of 1-best cepstral coefficients $\hat{c}$ calculated in each recursive iteration was verified in terms of impact on the residual signal power, estimated as

$$P_e = \lim_{N \to \infty} \frac{1}{N+1} \sum_{n=-N/2}^{N/2} e^2(n) \approx \frac{1}{N} e^\ell e,$$

where $N$ in this case is the number of samples of the database. Figure 5 plots $P_e$ for each iteration. Residual signals were obtained through inverse filtering with the mel generalized log spectrum approximation (MGLSA) filter structure [5]. It can be seen that the residual power has a relative significant decrease in the first recursive iteration of the joint training.

5.2. Quality resulting from the joint modeling

An informal listening test using natural $F_0$ and phonetic durations was conducted between the proposed joint training framework and the method in which both excitation and acoustic models are separately trained, corresponding to the initialization process as illustrated in Figure 2. Excitation signal at run-time was constructed as depicted in Figure 1, with pulse trains being generated from $F_0$. The MGLSA filter, which can implement the vocal tract filter using cepstral coefficients directly, was used instead of the all-zero structure of $H_s(z)$.

According to the results the synthesized speech quality produced by both approaches is indistinguishable. One possible cause for this was the limited size of the training database.

6. Conclusion

A proposal of joint estimation of acoustic and excitation models for statistical parametric speech synthesis has been presented and its training procedure based on ML described. The resulting system becomes what can be interpreted as a statistical modeling of the speech waveform. The approximations made for the estimation of the parameters of the introduced joint acoustic-excitation model consisted of fixing the state sequence along the entire training process and calculation of a 1-best spectral coefficient vector at each iteration. Future work includes the utilization of other spectral parameterizations, estimation of $N$-best spectral parameters to approximate $p(s|\ell, \lambda)$, and experiments with larger training data.

7. References