IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

Statistical Text-To-Speech Synthesis based on Segment-wise Representation with a Norm Constraint

Stas Tiomkin, David Malah and Slava Shechtman

CCIT Report #754
January 2010

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL

CCIT REPORT #754 January 2010

Statistical Text-To-Speech Synthesis based on Segment-wise Representation with a Norm Constraint

Stas Tiomkin[†‡], David Malah[†], and Slava Shechtman[‡]

[†] Department of Electrical Engineering, Technion-I.I.T, Israel Institute of Technology, Haifa 32000, Israel.

[‡] Speech Technologies group, Haifa Research Lab, IBM

Email: {stast@tx, malah@ee}.technion.ac.il, slava@il.ibm.com

Abstract

In statistical HMM-based TTS systems (STTS), speech feature dynamics is modelled by first- and second-order feature frame differences, which, typically, do not satisfactorily represent frame to frame feature dynamics present in natural speech. The reduced dynamics results in over-smoothing of speech features, often sounding as muffled and buzzy synthesized speech. In this work we propose a method to enhance a baseline STTS system by introducing a segment-wise model representation with a norm constraint. The segment-wise representation provides additional degrees of freedom in speech feature determination. We exploit these degrees of freedom for increasing the speech feature vector norm to match a norm constraint. As a result, statistically generated speech features are not over-smoothed, resulting in more natural sounding speech, as judged by listening tests. The proposed method consumes less real-time memory during synthesis, and the applied iterative algorithm has faster convergence than a Global-Variance (GV) approach, reported earlier, with comparable quality.

Index Terms

Text to speech (TTS) synthesis, statistical TTS, speech feature dynamics, segment-wise model representation.

I. INTRODUCTION

There is great interest in improving the quality of Text-To-Speech (TTS) systems as the number of applications using TTS increases rapidly. For example: 1) Any activity that occupies both hands, and at the same time needs to get response from a machine, may be facilitated by using TTS as an output device. 2) Vision-impaired people may receive written information through a TTS output device. 3) Different computer applications may use TTS as an output device. In addition for the convenience of using TTS, it enables a user-friendly interface, because the TTS system output may be adjusted to produce speech with desired and predefined features. E.g., a user may choose to get his machine response to sound as the voice of a particular person. TTS is thus greatly applicable to both industrial and entertainment applications. All these facts increase the importance of high-quality TTS devices.

There are two main approaches for solving the TTS paradigm. The first one uses basic units, which are either recorded speech segments or parameters representing these segments. These units may correspond to words, phonemes or even sub phonemes, as used in this research. This speech generation method is called concatenative TTS (CTTS). In this approach, speech is generated by concatenating the best compatible segments according to certain concatenation rules. By this approach, generated speech inherently possesses natural quality. However its quality depends on the size of the recorded database, as high-quality CTTS needs an extensive database. The main disadvantage of CTTS is possible discontinuities at segment boundaries due to concatenation. The smaller is the size of the stored database, the larger the number of discontinuities that typically appear in the generated speech. Thus, in applications where storage and computational resources are limited, such as mobile devices, a small footprint system is necessary, resulting in reduced quality of CTTS generated speech.

The other TTS approach employs statistical models for speech production and is called statistical TTS (STTS). STTS does not use natural speech segments but rather generates speech from previously learned statistical models, requiring much less storage than natural segments used by CTTS. Being generated from statistical models, speech generated by STTS is smoother. However, generally, STTS-generated speech is often over-smoothed, resulting in degraded speech quality in the form of muffled and buzzy speech [4], [5], [7]. Efforts invested in handling the over-smoothing problem are reported in [4], [5].

A Global Variance based approach, detailed in [6], has been suggested to alleviate the over-smoothing problem by applying a penalty for a reduction in the GV. However, it needs additional statistics to model the global variance, and is computationally complex, [6].

In this work we improve the baseline HMM-based STTS system by introducing new concepts into the current STTS methodology and provide a systematic approach for the integration of these concepts. The new-introduced concepts are: *a)* A robust model representation, based on a segment-wise representation, instead of the conventional frame-wise representation; *b)* Norm-regulated statistical speech feature vector meeting a norm constraint. These concepts are utilized in an iterative algorithm, proposed in this work. This algorithm generates speech features with enhanced dynamics, resulting in improved generated speech naturalness, as compared to the conventional generating scheme, and verified by listening tests.

The paper is organized as follows. In Section II we provide the essentials of the baseline STTS methodology used in this research. In Section III we present the segment-wise model representation. In Section IV we present the norm-regulated constraint, applied to the synthesized speech feature vector, and an iterative algorithm that generates speech features having enhanced dynamics. In Section V we examine the performance of the enhanced statistical TTS system. And, in Section VI we summarize and suggest a future work that can be pursued in continuation of the current research.

II. HMM-BASED TEXT-TO-SPEECH SYNTHESIS

In this section we briefly describe the conventional approach for deriving the entire utterance speech feature vector in statistical TTS.

A. Speech Feature Representation

In this research speech spectrum log-amplitude, $\mathbf{A}(f')$, of every frame is modeled by a linear combination of triangular basis functions, $\mathbf{B}_n(f'), n = 1, 2, \ldots, M$, as follows:

$$log(\mathbf{A}(f')) = \sum_{n=1}^{M} c_n \cdot \mathbf{B}_n(f'), \tag{1}$$

where f' denotes a mel-scale frequency[1]. This representation is successfully used in IBM's state-of-the-art CTTS system, detailed in [8], and, a corresponding speech reconstruction unit is

[1]The mel-scale mapping is $f' = 2595 log_{10}(1 + \frac{f}{700})$.

detailed in [19]. However, it was not previously used in HMM-based speech synthesis systems. Examination of the suitability of this representation for HMM-based speech synthesis is one of the goals of this research. Other common speech representations for HMM-based speech synthesis are $MFCC$ and LPC, as detailed in [4] and [12], respectively.

A speech feature vector over an entire utterance, having *N* frames, is represented in this paper by:

$$\mathbf{c} = [\mathbf{c}_1^T, \mathbf{c}_2^T, \ldots, \mathbf{c}_N^T]^T, \tag{2}$$

where $\mathbf{c}_i = (c_i(1), c_i(2), \ldots, c_i(M))^T$ are the expansion coefficients, introduced in (1). $\mathbf{c}_i$ denotes the static feature vector of dimension $M \times 1$ of the *i-th* frame, where $M = 32$. In this research we used frames of the length of 20ms with a frame overlap of 10ms. The prosody, (pitch, energy and duration), is modeled by a context-depended regression tree, detailed in [9], [10], and [14].

The general HMM architecture assumes statistical independence between visible states, while hidden states are statistically dependent via a hidden states transition matrix, as detailed in [1]. However, this assumption is not realistic for speech modeling because temporal events in natural speech are actually not independent. To handle this discrepancy, which exists in HMM speech modeling methodology, the static speech features are augmented by the dynamic speech features, as considered in [2], [3], [4] and [5]. The static speech features along with the dynamic ones constitute an augmented speech feature space, which is the conventional space for speech modeling. The static and dynamic features are combined into a vector

$$\mathbf{o} = [\mathbf{o}_1^T, \mathbf{o}_2^T, \ldots, \mathbf{o}_N^T]^T, \tag{3a}$$

where,

$$\mathbf{o}_i = (\mathbf{c}_i^T, \Delta^1\mathbf{c}_i^T, \Delta^2\mathbf{c}_i^T)^T. \tag{3b}$$

The dynamic features $\Delta^m\mathbf{c}_i$, for the *i-th* frame, approximate the *m-th* order difference in time of the static features $\mathbf{c}_i$, as detailed in [11]. When using only the two-sided first and second order differences, the dynamic features are computed as:

$$\Delta^{1,2}\mathbf{c}_i^T = \sum_{\tau=L_-^{(1,2)}}^{L_+^{(1,2)}} \omega^{1,2}(\tau)\mathbf{c}_{i+\tau}^T, \tag{4}$$

where $\omega^{1,2}$ are the weighting coefficients of the two-sided approximated first and second order derivatives expansions, respectively, and, $L^{(1,2)}_{(+,-)}$ are the left ('-') and right ('+') expansions limits for the first and second order expansions, respectively. Consequently, the vector $\mathbf{o}$, over an entire utterance, can be obtained from $\mathbf{c}$ by a linear transformation:

$$\mathbf{o}_{3M\cdot N\times 1}=\mathbf{W}_{3M\cdot N\times M\cdot N}\mathbf{c}_{M\cdot N\times 1}, \tag{5}$$

where the matrix W is constructed according to the first and 2^{nd} difference vectors $\Delta^1\mathbf{c}_i$ and $\Delta^2\mathbf{c}_i$, respectively.

B. Statistical Model

Given a continuous mixture HMM, λ, the optimal observation vector $\mathbf{o}$ over an entire utterance is derived by:

$$\mathbf{o}^{opt} = \underset{\mathbf{o}}{argmax}\ \ P(\mathbf{o} \mid \lambda) \tag{6}$$

and

$$P(\mathbf{o}|\lambda) = \sum_{\forall \mathbf{q}} P(\mathbf{o}, \mathbf{q}|\lambda), \tag{7}$$

where $\mathbf{q} = (q_1, q_2, \ldots, q_N)$ is the state sequence.

We use 'left-to-right', without skips, context-dependent HMM models with three emitting states per phoneme for speech spectrum modeling [1]. So, every phoneme p consists of three states p_1, p_2 and p_3. The emitting probability densities are each modeled by a Gaussian mixture model.

In order to represent statistically an entire utterance we compose a statistical model over this utterance by concatenation of corresponding context-dependent HMMs, where contexts are derived from phonetic analysis of synthesized text [13].

As mentioned in Section II-A, the prosody is modeled by context-dependent regression tree, which provide the phonetic identities of states and their durations. Hence, we can reduce the general problem of solving equation (6) to the following problem, which assumes that the state sequence, $\mathbf{q}$, is given:

$$\mathbf{o}^{opt} = \underset{\mathbf{o}}{argmax}\ \ P(\mathbf{o} \mid \mathbf{q}, \lambda), \tag{8}$$

Methods for full HMM-based speech feature synthesis appear at [4] and [5].

Without loss of generality the emitting probability distributions are modeled here by a single Gaussian model, because mixture components can be considered as a sequence of sub states, where states transitions are mixture weights. Under such assumptions, the logarithm of $P(\mathbf{o} \mid \mathbf{q}, \lambda)$ can be written as:

$$log(P(\mathbf{o} \mid \mathbf{q}, \lambda)) \;\; = \;\; \frac{1}{2}(\mathbf{o}-\mathbf{m})^T\mathbf{U}^{-1}(\mathbf{o}-\mathbf{m}), \tag{9}$$

with

$$\mathbf{m} \;\; = \;\; [\mathbf{m}_{q_1}^T, \mathbf{m}_{q_2}^T, \ldots, \mathbf{m}_{q_N}^T]^T \tag{10}$$

and

$$\mathbf{U}^{-1} \;\; = \;\; diag[\mathbf{U}_{q_1}^{-1}, \mathbf{U}_{q_2}^{-1}, \ldots, \mathbf{U}_{q_N}^{-1}], \tag{11}$$

where $\mathbf{m}_{q_t}^T$ and $\mathbf{U}_{q_t}^{-1}$ are the mean vector and the inverse covariance matrix of the state q_t. The dimensions of $\mathbf{m}$ and $\mathbf{U}^{-1}$ are $3MN \times 1$ and $3MN \times 3MN$, respectively. If the emitting probability densities are each modeled by a single Gaussian model, the mixture indices can be omitted. When the state q_t has duration d_t frames, its mean vector, $\mathbf{m}_{\mathbf{q_t}}$ and its inverse covariance matrix $\mathbf{U}_{q_t}^{-1}$ are replicated d_t times within $\mathbf{m}_{3MN\times 1}$ and $\mathbf{U}_{3MN\times 3MN}^{-1}$, respectively. This aspect of the conventional representation will be considered in Section III.

Clearly, given equation (9), expression (8) is optimized for $\mathbf{o} = \mathbf{m}$, which causes the augmented speech feature vector, $\mathbf{o}$, to become a sequence of the model means. However, we are interesting in finding the optimal speech feature vector, $\mathbf{c}^{opt}$, which incorporates the speech features dynamics, $\Delta^{1,2}\mathbf{c}$. This is achieved by solving the optimization problem in (8), taking into consideration the relation between the static and dynamic features, defined by equation (5) (note that $\mathbf{W}$ is not invertable):

$$\mathbf{o}^{opt} = \underset{\mathbf{o}}{argmax} \quad P(\mathbf{o} \mid \mathbf{q}, \lambda)|_{\mathbf{o}=\mathbf{W}\mathbf{c}}\,.$$

Consequently, the cost function over an entire utterance is:

$$\begin{aligned} J(\mathbf{W}\mathbf{c}) &= -lnP(\mathbf{W}\mathbf{c} \mid \mathbf{q}, \lambda) \\ &= \frac{1}{2}(\mathbf{W}\mathbf{c}-\mathbf{m})^T\mathbf{U}^{-1}(\mathbf{W}\mathbf{c}-\mathbf{m}) \\ &= \frac{1}{2}\|\mathbf{U}^{-\frac{1}{2}}(\mathbf{W}\mathbf{c}-\mathbf{m})\|_2^2. \end{aligned} \tag{12}$$

To find the optimal solution $\mathbf{c}^{opt}$ over an entire utterance, we set the first derivative of $J(\mathbf{Wc})$ with respect to $\mathbf{c}$ to 0:

$$\begin{aligned}\frac{\partial J(\mathbf{Wc})}{\partial \mathbf{c}} &= -\mathbf{W}^T\mathbf{U}^{-1}\mathbf{Wc} + \mathbf{W}^T\mathbf{U}^{-1}\mathbf{m} \\ &= 0 \end{aligned} \tag{13}$$

consequently,

$$\mathbf{W}^T\mathbf{U}^{-1}\mathbf{Wc} = \mathbf{W}^T\mathbf{U}^{-1}\mathbf{m}. \tag{14}$$

Assuming that the matrix $\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W}$ is invertible, the optimal solution $\mathbf{c}^{opt}$ is given by:

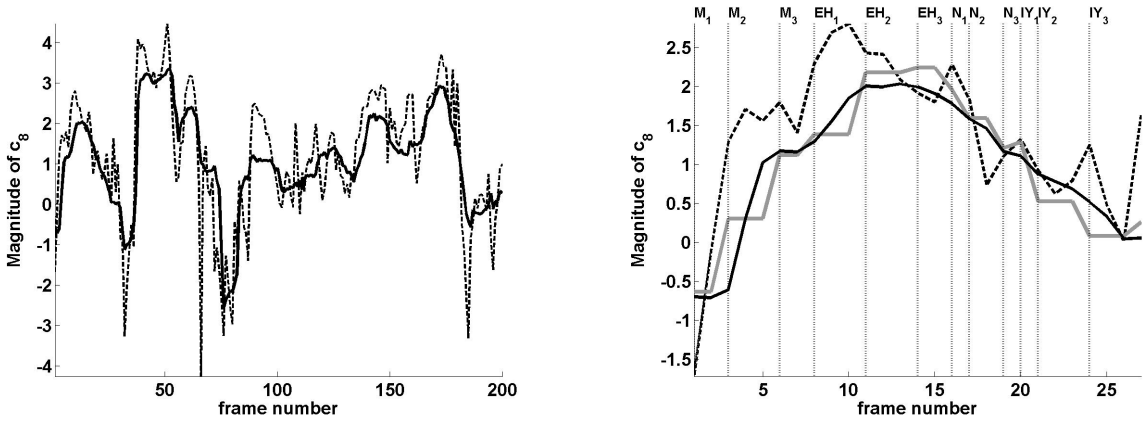
$$\mathbf{c}^{opt} = (\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{U}^{-1}\mathbf{m}. \tag{15}$$

To solve (15) directly requires $O(N^3M^3)$ computations. However, utilizing the special structure of $\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W}$, (15) can be solved by the Cholesky decomposition or the QR decomposition with $O(NM^3L^3)$, where $L = (\max L^1_+, L^1_-, L^2_+, L^2_-)$.

We can see in Fig. 1(a) that, typically, the optimal solution (15) is over-smoothed and has much less dynamics (inter-frame variations), as compared to the corresponding natural speech features. The natural 8-*th* expansion coefficient, $c_8^{natural}$ is provided as a reference, showing the range of expected variation. Perceptually, the reduced variance in speech features is associated with muffled and buzzy sound, as indicated by listening, and as also reported in [4], [5], [6], [7].

Fig. 1(b) provides zooming into the word 'Many', partitioned into the marked HMM-states, $M_1, M_2, \ldots, IY_3$, having duration in frames of $d_{M_1} = 2$, $d_{M_2} = 3$, $d_{M_3} = 2$, $d_{EH_1} = 3$, $d_{EH_2} = 3$, $d_{EH_3} = 2$, $d_{N_1} = 1$, $d_{N_2} = 2$, $d_{N_3} = 1$, $d_{IY_1} = 1$, $d_{IY_2} = 3$, and $d_{IY_3} = 3$, respectively. The state means (solid gray line) are replicated according to the state durations, e.g., the state 'M_3' lasts two frames. This zoomed part makes it clear that conventional statistically generated speech features (dashed line) pass smoothly from state to state. The statistical speech feature trajectory is a smoothed path, lacking the significant variations, in the reference natural speech feature trajectory about state means.

Thus, $\mathbf{\Delta}^{1,2}\mathbf{c}_i$ do not appear to fully capture the features dynamics, as also indicated by listening. We conclude from Fig. 1(a) and Fig. 1(b) that generated speech features should approximate the model means but, at the same time, they should fluctuate about the model



(a) Variation in time of the 8-*th* expansion coefficient, c_8, in the utterance '**Many problems in reading and writing are due to old habits**': c_8^{opt} in solid line; $c_8^{natural}$ in dashed line.

(b) Zooming in at the word '**Many**': c_8^{opt} in solid black line, $c_8^{natural}$ in dashed line. The vertical dashed lines depict the HMM states alignment, marked above the plot. The state means are shown in solid gray line.

Fig. 1. Demonstrating conventional statistically generated speech feature over-smoothing in time, compared to a reference natural speech feature.

means in order to have similar behavior to that of natural speech features. This may be achieved by a less restrictive model, which enables generating speech features with a controlled amount of fluctuations around the model means but sufficiently approximate the models. In the following sections we introduce a new concept of segment-wise model representation, which is found to improve the naturalness of generated speech.

III. Segment-Wise Model Representation

As discussed earlier, the insufficient speech feature dynamics in conventional frame-wise representation STTS systems causes over-smoothing of statistically generated speech features, resulting in muffled and buzzy speech.

In order to understand the drawbacks of the conventional frame-wise representation, consider two contiguous states, q_t and q_{t+1}, having durations d_t and d_{t+1}. In the conventional approach the augmented space speech feature frames $\mathbf{o}_t, ..., \mathbf{o}_{t+d_t-1}$ and $\mathbf{o}_{t+d_t}, ..., \mathbf{o}_{t+d_t+d_{t+1}-1}$ approximate the corresponding model means $\mathbf{m}_{q_t}$ and $\mathbf{m}_{q_{t+1}}$, replicated d_t and d_{t+1} times, respectively.

Consequently, the static features, $\mathbf{c}_t, ..., \mathbf{c}_{t+d_t-1}$, approximate the same static feature model mean, and at the same time, the corresponding dynamic features, $\mathbf{\Delta}_t^{1,2}\mathbf{c}_t, ..., \mathbf{\Delta}_{t+d_t-1}^{1,2}\mathbf{c}_{t+d_t-1}$,

approximate the same dynamic feature model mean. The covariance matrix is replicated d_t times within a segment as well, providing the same static and dynamic weight to every generated frame and inter-frames dynamics, respectively. In addition, averaging over speech features often results in a mean value of the dynamic features that is of very low magnitude. As a result, statistically generated speech features lack speech feature dynamics and do not achieve the natural variances, represented by model covariance matrices, as seen in Fig 1(b). The conventional model just connects smoothly adjacent models, involving a computationally complex matrix inversion, and redundant data storage required to store the statistics of $\Delta^{1,2}\mathbf{c}_t$, which do not have a sufficient effect, as depicted in this figure.

The above mentioned conventional representation drawbacks often cause speech feature over-smoothing. To handle the over-smoothing problem we propose to apply a segment-wise construction of the augmented space vector $\mathbf{o}$ over an entire utterance, implemented by a modified linear segment-wise transformation, denoted $\widetilde{\mathbf{W}}$.

We propose not to replicate the model mean $\mathbf{m}_{q_t}$ d_t times, but rather approximate on average d_t augmented space vectors, $\mathbf{o}_t, ..., \mathbf{o}_{t+d_t-1}$, by the model mean of state q_t, as follows:

$$\bar{\mathbf{o}}_t = \frac{1}{d_t}\sum_{k=1}^{d_t}\mathbf{o}_k, \tag{16}$$

and

$$J(\bar{\mathbf{o}}_t) = \frac{1}{2}\|\mathbf{U}_{q_t}^{-\frac{1}{2}}(\bar{\mathbf{o}}_t - \mathbf{m}_{q_t})\|_2^2, \tag{17}$$

where $\bar{\mathbf{o}}_t$, $\mathbf{m}_{q_t}$ and $\mathbf{U}_{q_t}$ are the average augmented feature vector, the model mean and the model covariance matrix of state q_t, respectively. And, $J(\bar{\mathbf{o}_t})$ is the corresponding cost function, constructed without replication of the model of the state q_t.

Consequently, using the proposed segment-wise representation, the model is less restricted and enables more dynamics in generated speech features, which is decreased in the conventional model.

The segment-wise transformation for speech feature frames pertaining to a particular state q_t with duration d_t, is:

$$\widetilde{\mathbf{W}}_{q_t} \triangleq \frac{1}{d_t}\begin{pmatrix} \mathbf{0} & \mathbf{1} & \overset{d_t-2}{\cdots\mathbf{1}\cdots} & \mathbf{1} & \mathbf{0} \\ -\frac{\mathbf{1}}{\mathbf{2}} & -\frac{\mathbf{1}}{\mathbf{2}} & \overset{d_t-2}{\cdots\mathbf{0}\cdots} & \frac{\mathbf{1}}{\mathbf{2}} & \frac{\mathbf{1}}{\mathbf{2}} \\ -\mathbf{1} & \mathbf{1} & \overset{d_t-2}{\cdots\mathbf{0}\cdots} & \mathbf{1} & -\mathbf{1} \end{pmatrix}_{3M\times M(d_t+2)} . \tag{18}$$

All the matrix elements in (18) are diagonal block matrices of dimension $M \times M$, each. A part of the segment-wise transformation for two contiguous states, q_t and q_{t+1}, having $d_t = 3$ and $d_{t+1} = 2$, is shown in (19):

$$\widetilde{\mathbf{W}}_{MK\times MN} = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \cdots \\ \cdots & -\frac{1}{6} & -\frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{6} & 0 & 0 \cdots \\ \cdots & -\frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & -\frac{1}{3} & 0 & 0 \cdots \\ \cdots & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \cdots \\ \cdots & 0 & 0 & 0 & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} \cdots \\ \cdots & 0 & 0 & 0 & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}_{MK\times MN} . \tag{19}$$

Here, K is the total number of states in a synthesized utterance. Consequently, the argument, $(\widetilde{\mathbf{W}}\mathbf{c} - \widetilde{\mathbf{m}})$, of the segment-wise cost function, denoted as $J^{sw}(\widetilde{\mathbf{W}}\mathbf{c})$, is rearranged as:

$$\widetilde{\mathbf{W}}\mathbf{c} - \widetilde{\mathbf{m}} = [\, \mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_t}, \ldots, \mathbf{w_K} \,]^T, \tag{20}$$

where $\widetilde{(\cdot)}$ denotes non replication of state models, but rather approximation on average of state models, and $\mathbf{w_t}$ is:

$$\mathbf{w_t} \triangleq \frac{1}{d_t} \sum_{i=t-\lfloor \frac{d_t}{2} \rfloor}^{t+\lfloor \frac{d_t}{2} \rfloor} \mathbf{o}_i^T - \mathbf{m}_{q_t}^T. \tag{21}$$

The segment-wise cost function, $J^{sw}(\widetilde{\mathbf{W}}\mathbf{c})$, over an entire utterance is:

$$J^{sw}(\widetilde{\mathbf{W}}\mathbf{c}) = \frac{1}{2}\|\widetilde{\mathbf{U}}^{-\frac{1}{2}}(\widetilde{\mathbf{W}}\mathbf{c} - \widetilde{\mathbf{m}})\|_2^2, \tag{22}$$

where

$$\widetilde{\mathbf{m}}_{3MK\times 1} = [\mathbf{m}_{q_1}^T, \mathbf{m}_{q_2}^T, \ldots, \mathbf{m}_{q_K}^T]^T, \tag{23}$$

and

$$\widetilde{\mathbf{U}}^{-1}_{3MK\times 3MK} = diag[\mathbf{U}_{q_1}^{-1}, \mathbf{U}_{q_2}^{-1}, \ldots, \mathbf{U}_{q_K}^{-1}] \tag{24}$$

are the non-replicated model mean vector and the covariance matrix, respectively, where $\widetilde{\mathbf{m}}_{3MK\times 1}$ and $\widetilde{\mathbf{U}}^{-1}_{3MK\times 3MK}$ consist of K state means, $\mathbf{m}^T_{q_t}$, and state covariance matrices, $\mathbf{U}^{-1}_{q_t}$, respectively, This is in contrast to the frame-wise model mean vector, $\mathbf{m}_{3MN\times 1}$, and the frame-wise model covariance matrix, $\mathbf{U}^{-1}_{3MN\times 3MN}$, defined in (10) and (11), respectively, which contain replicated terms (note the different dimensions). This defines the segment-wise representation, where all the static and dynamic features are approximated on average by the static and dynamic feature model means, respectively. As a result, statistically segment-wise generated speech features can possess enhanced speech dynamics and follow the model means in the mean, as opposed to the frame-wise synthesis, where every particular frame follows a smooth trajectory, approximating the model means.

Consequently, the conventional frame-wise cost function in (12) should be denoted as J^{fw} in order to distinguish between the two different cost functions. Here and forth, the segment-wise cost function and the frame-wise cost function will be marked with the corresponding superscripts 'sw' or 'fw', respectively.

The optimal solution for the segment-wise cost function, (22), is derived by the same steps as in (13), (14) and (15):

$$\begin{aligned}\frac{\partial J^{sw}(\widetilde{\mathbf{W}}\mathbf{c})}{\partial \mathbf{c}} &= -\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{W}}\mathbf{c} + \widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{m}} \\ &= 0 \end{aligned} \tag{25}$$

consequently,

$$\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{W}}\mathbf{c} = \widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{m}}. \tag{26}$$

Assuming the matrix $\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{W}}$ in (26) is invertible, the optimal segment-wise solution $\mathbf{c}^{opt,sw}$ is derived by:

$$\mathbf{c}^{opt,sw} = (\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{W}})^{-1}\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{m}}. \tag{27}$$

Reiterating, in the segment-wise representation we require that all the frames of state q_t approximate the model of q_t on average, (instead of frame-wise approximation used in the conventional model, where every frame approximates a corresponding model). This results in an infinite number of solutions, $\mathbf{c}^{opt,sw}$, for states having duration more than one frame. In such a case, the matrix $\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{W}}$ is non-invertible and, consequently, it requires a special treatment,

subject to the requirement on the generated speech feature norm. A solution to this problem is proposed in Section IV.

IV. Norm Constraint

We have observed[2] that the squared-norm of statistically generated speech feature vectors of entire utterances, $\|\mathbf{c}^{stt}\|_2^2$, is often quite lower than the squared-norm of natural speech feature vectors of entire utterances, $\|\mathbf{c}^{nat}\|_2^2$, because, firstly, the conventional solution, shown in (15), is the minimal norm least squares solution, and, secondly, due to the insufficient speech feature dynamics, a statistically generated speech feature vector norm is quite close to the model means norm $\|\mathbf{c}^{mdl}\|_2^2$:

$$\|\mathbf{c}^{stt}\|_2^2 \approx \|\mathbf{c}^{mdl}\|_2^2 \tag{28}$$

In Fig. 1(a) and 1(b) we saw that, typically, statistically generated frames are much smoother than the corresponding natural frames. In order to improve generated speech quality, we can enhance speech feature dynamics by applying appropriate constraints to the feature vector.

We propose to enhance speech feature dynamics by enforcing a constraint on the speech feature vector norm. In addition to the regular terms of the common statistical model cost function (12), we add a norm-dependent auxiliary term, constraining the speech feature vector norm, thus avoiding the norm reduction. The proposed approach relies on different concepts than those of the GV [6] approach, as our approach exploits the principles of regularization theory, described below. Also, our approach requires just two additional scalar parameters per speaker database, introduced in this section, while GV applies a statistical penalty for variance reduction and needs additional statistics to model global variance.

Comparing statistically generated speech features to corresponding natural speech features, we found that the norm of statistically generated speech feature vector $\|\mathbf{c}^{stt}\|_2^2$, is systematically reduced, in comparison to the norm of natural speech feature vectors, $\|\mathbf{c}^{nat}\|_2^2$, by a factor γ_0:

$$\gamma_0 = \frac{\widetilde{\|\mathbf{c}^{nat}\|_2^2}}{\widetilde{\|\mathbf{c}^{stt}\|_2^2}}, \tag{29}$$

[2]A set of 40 arbitrary sentences was generated, whose speech feature vector norms were examined, and compared to a) corresponding speech feature model mean vector norms, and b) corresponding natural speech feature vector norms. The expressions (28) and (29) represent the averaged results of the comparisons.

denoted as the enhancement factor, where $\widetilde{\|\cdot\|}$ is an averaged norm over a set of utterances generated from a particular voice.

Consequently, using (28), a constraint on the norm of speech features, $\|\mathbf{c}^{stt}\|_2^2$, should be equal to

$$\Gamma = \gamma_0 \cdot \|\mathbf{c}^{mdl}\|_2^2, \tag{30}$$

in order to compensate the norm reduction, achieving in our case:

$$\|\mathbf{c}^{stt,sw}\|_2^2 \approx \|\mathbf{c}^{nat}\|_2^2 \tag{31}$$

In the following section we provide a systematic approach for speech feature dynamics enhancement by applying such a constraint.

A. Norm-constrained cost function

We seek an optimal solution $\mathbf{c}^{opt}$ that obeys a norm constraint:

$$\begin{aligned} &\mathbf{c}^{opt} = \underset{\mathbf{c}}{argmin}\ \ J(\mathbf{W}\mathbf{c}), \\ &\text{s.t.} \quad \|\mathbf{c}\|_2^2 = \Gamma, \end{aligned} \tag{32}$$

where Γ is given in (30), $J(\mathbf{W}\mathbf{c})$ is given in (12), and $\mathbf{W}$ stands here for either the segment-wise transformation or the frame-wise one.

We solve this constrained optimization problem by means of the Lagrangian function $L(\mathbf{c}, \eta)$ with a scalar Lagrangian multiplier η:

$$\begin{aligned} L(\mathbf{c}, \eta) = &\frac{1}{2}\|\mathbf{U}^{-\frac{1}{2}}\left(\mathbf{W}\mathbf{c} - \mathbf{m}\right)\|_2^2 \\ &+\frac{\eta}{2}(\|\mathbf{c}\|_2^2 - \Gamma). \end{aligned} \tag{33a}$$

Consequently, we set the derivatives of $L(\mathbf{c}, \eta)$ with respect to $\mathbf{c}$ and η to zero:

$$\begin{aligned} \frac{\partial L(\mathbf{c}, \eta)}{\partial \mathbf{c}} = &\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W}\mathbf{c} - \mathbf{W}^T\mathbf{U}^{-1}\mathbf{m} \\ &+\eta\mathbf{c} = 0, \end{aligned} \tag{33b}$$

and

$$\frac{\partial L(\mathbf{c}, \eta)}{\partial \eta} = \left(\|\mathbf{c}\|_2^2 - \Gamma\right) = 0. \tag{33c}$$

Using (33b) we get:

$$\mathbf{c}^{opt} = (\mathbf{A} + \eta\mathbf{I})^{-1}\,\mathbf{b}, \tag{34}$$

where, $\mathbf{I}$, here and forth, stands for $\mathbf{I}_{MN\times MN}$, $\mathbf{A}$ stands for $\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W}$, $\mathbf{b}$ stands for $\mathbf{W}^T\mathbf{U}^{-1}\mathbf{m}$, and η, using (33c), is a root of the following polynomial $p(\eta)$:

$$p(\eta) = (\|(\mathbf{A} + \eta\mathbf{I})^{-1}\mathbf{b}\|_2^2 - \Gamma). \tag{35}$$

However, $\mathbf{c}^{opt}$ can not be derived analytically due to the complicated expression in (35) for extracting η. In the next section we propose a different approach for generating speech feature vectors approximating a desired norm.

B. Iterative Algorithm

Our goal is to find an optimal norm-constrained feature vector, $\mathbf{c}^{opt}$, over an entire utterance, which minimizes the model error and possesses sufficient features dynamics, without finding explicitly η of (35).

For that end, we propose to regulate the solution by adding a squared-norm term of the feature vector to the model-error term of the cost function of (22), using a factor λ to balance the contribution of the two terms.

Thus, the cost function of (22) is replaced by:

$$\begin{aligned} J_c^{sw}(\widetilde{\mathbf{W}}\mathbf{c}) \triangleq \frac{1}{2}\|\mathbf{U}^{-\frac{1}{2}}(\widetilde{\mathbf{W}}\mathbf{c} - \mathbf{m})\|_2^2 \\ +\frac{\lambda}{2}\|\mathbf{c}\|_2^2, \end{aligned} \tag{36}$$

In the proposed method, the norm term provides a solution with enhanced dynamics, by using prior information on λ.

We propose an iterative algorithm that minimizes the model cost function value, while assuring sufficient dynamics in the resulting solution. The minimization is done by means of a gradient descent algorithm as follows:

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \alpha_n\nabla(\mathbf{c}_n), \tag{37}$$

where $\nabla(\mathbf{c}_n)$ is the gradient of $J_c(\mathbf{c})$ with respect to $\mathbf{c}$, computed at iteration n, and, α_n is the step size, being updated in our experiments according to:

$$\alpha_n = \frac{1}{\|\nabla(\mathbf{c}_n)\|_2^2}, \tag{38}$$

and from (36),

$$\nabla(\mathbf{c}_n) = \widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\widetilde{\mathbf{W}}\mathbf{c}_n - \widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\mathbf{m} + \lambda\mathbf{c}_n. \tag{39}$$

A final feature vector should approximate well the models, and have a norm value that is compatible with the enhancement factor, defined in (29). We propose to apply a balancing factor λ that decreases in its absolute value with the gradient descent algorithm iterations, rather than to use a fixed λ. This way the model error term becomes more significant with the number of iterations, while the norm factor effect decreases with the number of iterations. Consequently, (39) is replaced by :

$$\nabla(\mathbf{c}_n) = \widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\widetilde{\mathbf{W}}\mathbf{c}_n - \widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\mathbf{m} + \lambda_n\mathbf{c}_n, \tag{40}$$

where λ_n is updated according to:

$$\lambda_{n+1} = \theta\lambda_n,\ 0 \leq \theta \leq 1, \tag{41}$$

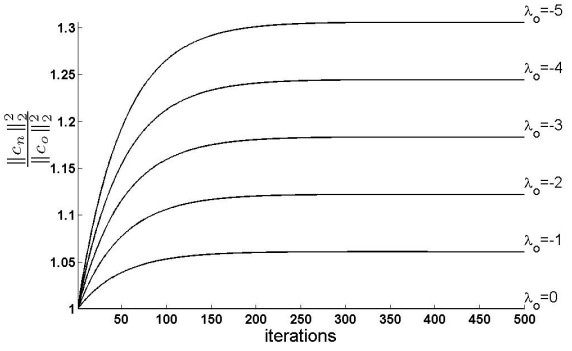
where the parameter θ is experimentally determined to enable a slow decrease of λ that is consistent with a required norm increase, as elaborated below. In our experiments we used $\theta = 0.95$, where an acceptable range of values for θ may reach 0.98.

Taking into consideration the cost function form in (36), we conclude that a negative λ value increases the feature vector norm, while a positive λ value decreases it.

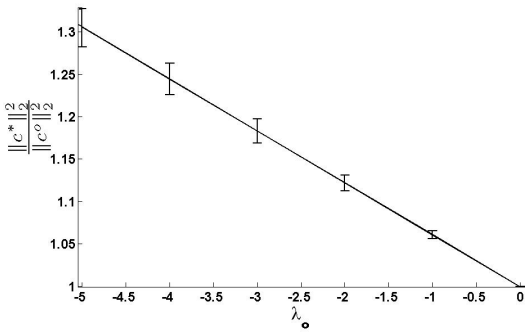
We found an empiric relation between λ_0, the initial value of λ, and the final norm of the feature vectors, allowing a norm increase that is consistent with the enhancement factor. In Fig. 2(a), we see that an increase in the negative value of λ_0 results in an increase in the final vector norm.

The desired increase in speech feature vector norm is achieved around 150 iterations, each of which consists of one multiplication of the n-*th* speech feature vector $\mathbf{c}_n$, having dimension $MN \times 1$, by the constant sparse matrix $\widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\widetilde{\mathbf{W}}$, having dimension $MN \times MN$, and one summation of two vectors of dimension $MN \times 1$.

The relation between λ_0 and, the attained maximal value of the feature vector norm $\|c^*\|_2^2$,



(a) An increase in a feature vector norm $\|c_n\|_2^2$ as a function of an initial value for λ_o, where $\|c_o\|_2^2$ is the norm of an initial vector.



(b) Relation between λ_o and the final feature vector norm $\|c^*\|_2^2$. The error bars depict the standard deviations in $\|c^*\|_2^2$ for given values of λ_o.

Fig. 2. Evolution of $\|c_n\|_2^2$ as function of λ_o.

is represented in Fig. 2(b) that is derived from Fig. 2(a) by plotting $\|c^*\|_2^2$ via λ_o. This relation was obtained by averaging λ_0 over a large set of iteratively generated utterances. The standard deviations of the final speech feature vector norm, for given values of λ_0, are represented by the error bars in Fig. 2(b). For λ_0 equal to -5, which is consistent with the enhancement factor, the standard deviation is 0.023.

Initially, as long as λ_n sufficiently effects $\nabla(\mathbf{c}_n)$, two updates affect $\mathbf{c}_n$ simultaneously: an increase in the norm of $\mathbf{c}_n$, occurring due to negative value of λ_n, and an attempt to keep $\mathbf{c}_n$ close to the model means. λ_n balances between these two updates, but its effect decreases with the number of iterations, as λ_n approaches 0.

When the effect of λ_n becomes negligible, the gradient descent algorithm steps towards the minimal model error. However, the static feature vector norm $\|\mathbf{c}_n\|_2^2$ does not decrease along with a decrease of the model error term but rather stays almost unchangeable at $\|c^*\|_2^2$. This occurs because the dynamic features, rather than the static ones exert the primary and the most significant effect on the model cost function, as described in the Appendix.

Setting λ_0 according to the above mentioned empiric relation enables an increase in the norm of $\mathbf{c}_n$ that is consistent with the norm enhancement factor introduced in (29), resulting in enhanced dynamics in generated speech, as confirmed by listening tests described in Section (V-B).

In our experiments the model means were used for the initial vector, $\mathbf{c}_o$, in the gradient descent algorithm.

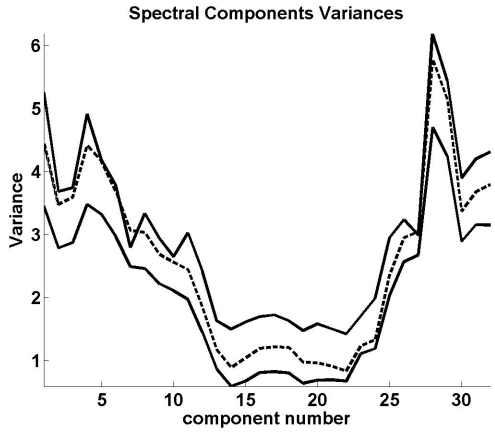
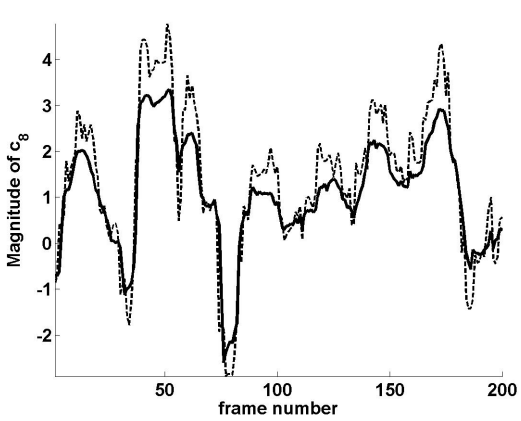


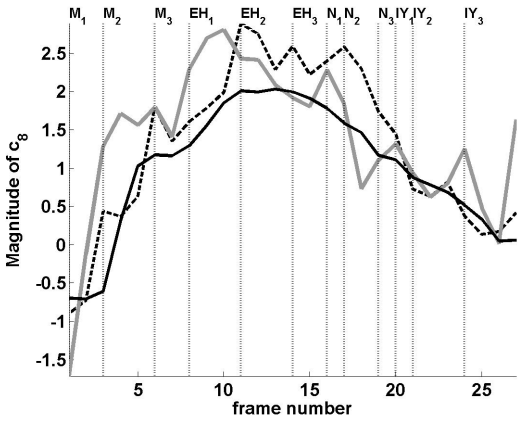
Fig. 3. Frequency components variances of natural utterance (top solid line): conventional STTS generated (bottom solid line); and proposed STTS generated utterance (middle dashed line).

In Fig. 3, we see that there is a systematic increase in speech feature dynamics, represented by the spectral components variances, computed over a set of utterances. The natural utterance speech feature variances (in the upper solid line) provide a reference for the expected speech feature variances. On the other hand, speech features generated by the conventional STTS system are often over-smoothed, and have lower variances, as described previously. Consequently, we expect that speech features generated by the proposed method, will have more variance than speech features generated by conventional method but less variance than speech features generated by a CTTS system. We see in Fig. 3 that, indeed, this is the case. Moreover, in the proposed method, in almost all bands, speech feature dynamics is closer to that of the natural speech features than to those generated by the conventional method. The last statement is confirmed by listening tests, indicating that the proposed method generates speech that sounds more natural.

The norm-regulated constraint is useful only with the segment-wise model. Applying the norm-regulated approach to the frame-wise model is not useful because the frame-wise model has its unique least-squares solution, $\mathbf{c}^{opt,fw}$, derived by (15). Clearly, the iterative solution via (37) with the frame-wise model must converge to $\mathbf{c}^{opt,fw}$, having a reduced speech feature norm, when the effect of λ decreases. On the other hand, the iterative solution via (37) with the segment-wise model, converges to a vector that approximates the required norm, as shown in Fig. 2(a).



(a) Variation in time of the 8-*th* expansion coefficient, c_8, in the utterance '**Many problems in reading and writing are due to old habits': 'c_8^{opt} - frame-wise conventional**' in solid line; 'c_8^{opt} - segment-wise' in dashed line.



(b) Zooming in at the word '**Many**': 'c_8^{opt} - segment-wise' in dashed line; 'c_8^{opt} - frame-wise conventional' in solid black line, and reference $c_8^{natural}$ in light gray line

Fig. 4. Comparison of speech feature dynamics in a conventional frame-wise model to that of the proposed segment-wise representation.

V. Experimental Results

A. Objective Evaluation

In Fig 4(a) we see that the segment-wise model enables more dynamics in generated speech feature trajectory (dashed line), compared to the more smooth trajectory by the conventional frame-wise model (solid line). Zooming in at the word 'Many' in Fig. 4(b), with marked HMM-states, we see that the frame-wise generated trajectory is much smoother (solid line) than the segment-wise generated trajectory (dashed line). The natural speech feature trajectory, in light gray line, which appears in dashed line in Fig. 1(b), is provided as a reference for expected speech feature dynamics. In the more detailed view of Fig. 4(b), we see that the segment-wise trajectory from state 'M_3' to state 'EH_3' has dynamics compared to the dynamics of the natural trajectory, while the frame-wise trajectory follows smoothly over these state model means, and, even, coincides with the mean of state 'M_3', as clearly shown in Fig. 1(b). The last fact emphasizes our assumption regarding insufficient dynamics in the conventional frame-wise models.

In our experiments, we compared the empirical data fitting by the segment-wise cost function,

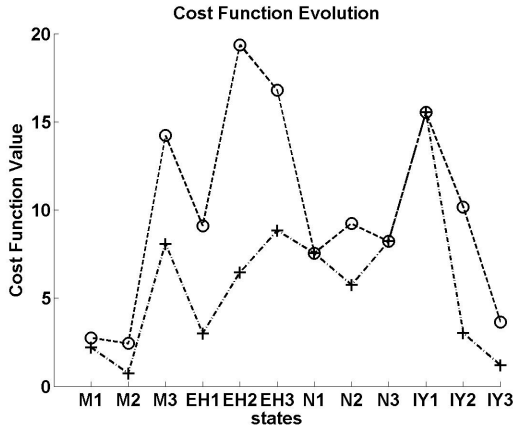


Fig. 5. Evolution of cost function over states of the word '**Many**' in a real speech sample, $\mathbf{c}^{natural}$: the segment-wise cost function, $J^{sw}(\widetilde{\mathbf{W}}\mathbf{c}^{natural})$, in pluses; the frame-wise cost function, $J^{fw}(\mathbf{W}\mathbf{c}^{natural})$, in circles. x-axis - corresponding states, y-axis - cost function value.

$J^{sw}(\widetilde{\mathbf{W}}\mathbf{c})$, to the empirical data fitting by the conventional frame-wise cost function, $J^{fw}(\mathbf{W}\mathbf{c})$. We performed this comparison by computing the cost functions values on real speech examples, $\mathbf{c}^{natural}$. In Fig. 5 we see the typical evolution of $J^{sw}(\widetilde{\mathbf{W}}\mathbf{c}^{natural})$ and $J^{fw}(\mathbf{W}\mathbf{c}^{natural})$ for a real utterance, $\mathbf{c}^{natural}$, where the x-axis depicts the states alignment of '**Many**'. Obviously, states with duration of one frame gives the same value in both cost functions, as seen for N_1, N_3 and IY_1. However, all other states have lower value in the segment-wise model, due to its more flexible construction, providing more degrees of freedom. The longer the state duration, the bigger is the difference is between the values of $J^{fw}(\mathbf{W}\mathbf{c})$ and $J^{sw}(\widetilde{\mathbf{W}}\mathbf{c})$. This demonstrates the better fit of the segment-wise model to real speech data.

B. Subjective Evaluation

We have performed three different listening tests to evaluate the naturalness of speech generated by the proposed method:

Test I:. In this test we have computed the Mean Opinion Score (MOS), according to [15], of a set of 9 arbitrary sentences, where each sentence was generated in three versions: *(i)* by the conventional statistical speech generation algorithm, mentioned in Section II, (group A), *(ii)* by the proposed speech generation scheme (group B), and *(iii)* by IBM's CTTS system, detailed in [10], [8], (group C). Thus, 27 samples were included in the test, each of which was evaluated

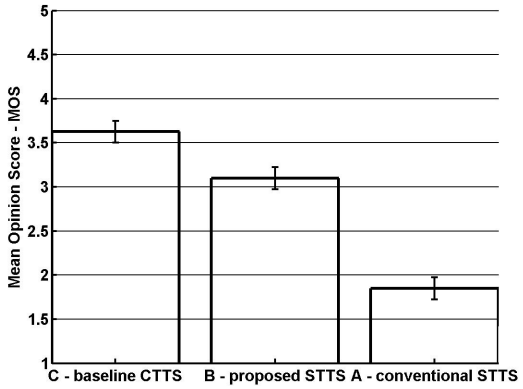


Fig. 6. Mean Opinion Score (MOS) test, comparing the CTTS system, the proposed STTS method, and the base-line STTS. The error bars indicate 95% confidence interval, computed using the 't-test'.

by 20 listeners. The same target prosody was used in the synthesis of all the tested versions of a particular sentence, so as to not be affected by different prosody targets in different systems

Fig. 6 shows the results of the MOS test for the three groups. We see that the proposed method improved the naturalness of generated speech by more than one MOS unit, in comparison to conventional STTS.

Test II:. This MOS test consisted of two sessions. In one session a group of 15 listeners evaluated samples generated by the proposed method only. In another session, another group of 15 listeners evaluated samples generated by the GV approach [6], which were downloaded from [20]. Consequently, the listeners in each group evaluated the quality of the respective approach, without being affected by the other technique results. Additionally, in contrast to the first MOS test that included an arbitrary set of sentences, the second MOS test included 25 sentences in 5 groups of 5 sentences each, selected from several different domains, having different lengths (from short simple sentences of 2-3 words, to compound sentences of 25 words) and distinctive phonetic contexts. This set of sentences is a standard set, used for evaluation of different TTS systems, as detailed in [18]:

- Gutenberg novels - 'Guten'.
- Standard news text - 'News'.
- Conversational/dialog sentences - 'Conv'.
- Phonetically confusable words - 'MRT', detailed in [16].

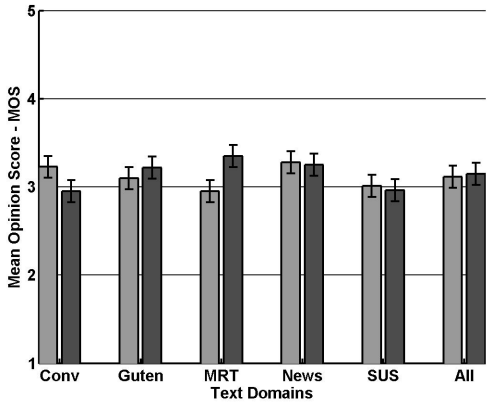


Fig. 7. Mean Opinion Score - MOS test for different text domains: 'Conv', 'Guten', 'MRT', 'News', 'SUS'. 'All' - average score for all text domains. The samples by the proposed method and by the GV method are in light gray and dark gray, respectively'. The error bars indicate 95% confidence interval, computed using the 't-test'.

- Semantically unpredictable sentences - 'SUS', detailed in [17].

Fig. 7 shows the results of this test. We see that both methods (GV and the proposed approach) achieve similar overall MOS score, as summarized by the columns 'All'. The MOS score of analysis-synthesized speech, (just analyzing the speech and synthesizing it back from its features), is 4.23, as reported in [19]. This high score for analysis-synthesized speech means that an 'encoder/decoder' introduces only a small speech quality degradation.

Test III: In this test, a set of 11 arbitrary sentences was used. Each of the sentences was generated in two versions: *(i)* by the conventional statistical speech generation algorithm, mentioned in Section II, (group A), and *(ii)* by the proposed speech generation scheme (group B). The two versions of each sentence were compared using an 'A vs B' comparison test, to provide a further indication on the improvement of speech quality generated by the proposed STTS method in comparison to the conventional statistical approach. The same 20 listeners, that participated in Test I, had three options to evaluate the relative quality of groups A and B: 'A is preferred', 'B is preferred' and 'A is the same as B'. Thus, 11 pairs of sentences were compared in the test, each of which was evaluated by 20 listeners. Fig. 8 shows the results of this test. We see that group B was preferred over group A in 91.6% of the cases, on average, 7.4% got the same preference, and group A was preferred over group B only in 1% of the cases. The same target prosody was used in the synthesis of all the tested versions of a particular sentence, so as to not

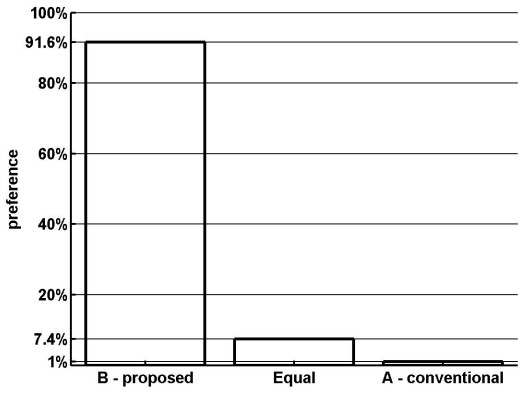


Fig. 8. 'A vs B comparison test' **(20 listeners, 11 pairs of sentences)**: the proposed STTS (group B) was preferred in 91.6% of the cases, on average; 7.4% were judged as having the same quality; and the conventional STTS (group A) was preferred only in 1% of the cases

be affected by different prosody targets in different systems

All the tests were performed with a headphone set. The only information about the samples that the listener were provided with, was that the test aims to compare different speech synthesis methods. All the listeners were graduate and undergraduate students, having no experience with TTS systems.

We conclude that the proposed segment-wise and norm constrained method significantly improves the synthesized speech quality, as compared to the baseline frame-wise conventional statistical speech synthesis method and is comparable in quality to the GV approach.

VI. Summary

In this work we propose a different approach than GV, [6], to enhance speech feature dynamics in a conventional base-line STTS system. The proposed system is based on a segment-wise augmented space representation and a norm-constrained iterative algorithm. Unlike the GV-based system, it does not involve additional data modelling and thus avoids an increase in memory footprint by about 30%, (each model requires $6M$ numbers for acoustic features, and the GV requires additional $2M$ numbers for global variance). Our proposed system just incorporates prior information on the norm reduction factor in conventional STTS feature vectors, relative to natural speech. In addition, it requires less real-time memory during synthesis than do conventional STTS

systems and than does GV. In addition, the convergence rate of the proposed iterative method is higher than in GV [6].

Both approaches achieve similar improvement in generated speech quality, relative to the conventional baseline STTS, as judged by the reported listening tests results.

Future Work

The segment-wise representation provides additional degrees of freedom in the determination of the speech feature vector. In this research we employed it for regulating the generated speech feature vector norm. However, these degrees of freedom can be employed for regulating other speech features, by properly choosing an additional term in the cost function, like we did for regulating the norm.

In current research, we are embedding the proposed STTS system in a hybrid TTS system, in which STTS and CTTS are combined, aiming to improve CTTS when it is operated at a reduced footprint. Preliminary results show that the hybrid TTS system, achieves a much better speech quality when conventional STTS is replaced by the STTS system proposed in this work.

ACKNOWLEDGMENT

This research is part of a joint research project conducted at the Signal and Image Processing Lab (SIPL), Technion–I.I.T, and IBM's Haifa Research Lab (HRL). The authors would like to thank Ron Hoory, Zvi Kons, Ariel Sagi, and Alex Sorin, for useful discussions in the course of the work and are indebted to Zvi Kons for his valuable comments to the manuscript.

The help of SIPL staff, Nimrod Peleg, Ziva Avni and Avi Rosen, is gratefully acknowledged in administrating the listening tests and in providing technical support.

The authors are thankful to the anonymous reviewers for their comments and suggestions that helped much to improve the paper.

APPENDIX A

In this appendix we elaborate on the behavior of the iterative algorithm, detailed in Section IV-B, by considering relations between the static and dynamic features to corresponding terms of the cost function, $J(\widetilde{\mathbf{W}}\mathbf{c})$.

The cost function, $J^{sw}(\widetilde{\mathbf{W}}\mathbf{c}) = \|\mathbf{U}^{-0.5}(\widetilde{\mathbf{W}}\mathbf{c} - \mathbf{M})\|_2^2$, consists of three terms: the static feature term $J_1(\mathbf{c})$, the first dynamic feature term $J_2(\Delta^1\mathbf{c})$ and the second dynamic feature term $J_3(\Delta^2\mathbf{c})$.

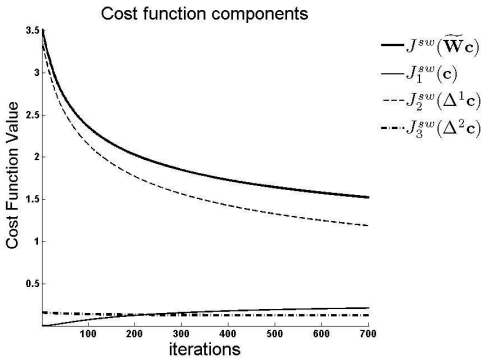


Fig. A-1. The components of the unconstrained segment-wise cost function, $J^{sw}(\widetilde{\mathbf{W}}\mathbf{c})$, where $J^{sw}(\widetilde{\mathbf{W}}\mathbf{c}) = J_1^{sw}(\mathbf{c}) + J_2^{sw}(\mathbf{\Delta^1 c}) + J_3^{sw}(\mathbf{\Delta^2 c})$.

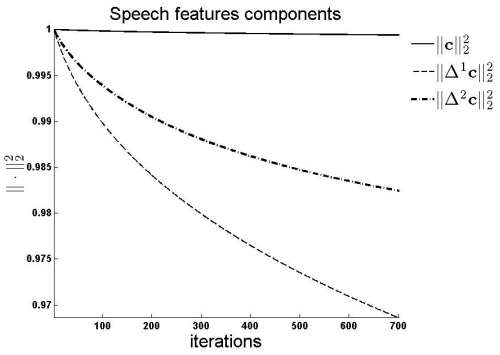


Fig. A-2. The normalized static feature norm and the normalized dynamic feature norm, obtained by the minimization of $J^{sw}(\widetilde{\mathbf{W}}\mathbf{c})$.

With a diagonal covariance matrix these terms are independent. Consequently,

$$J^{sw}(\widetilde{\mathbf{W}}\mathbf{c}) = J_1(\mathbf{c}) + J_2(\Delta^1\mathbf{c}) + J_3(\Delta^2\mathbf{c}), \tag{A-1}$$

where the terms contributions to the cost function $J(\widetilde{\mathbf{W}}\mathbf{c})$ are weighted according to corresponding variances of the static and the dynamic features (appeared on the main diagonal of the covariance matrix), which are related as follows. Denoting the variance of the static features as ρ^2, then, the variances of the $\Delta^1\mathbf{c}$ and the $\Delta^2\mathbf{c}$ are $\frac{1}{2}\rho^2$ and $6\rho^2$, respectively, according to the construction of $\Delta^1\mathbf{c}$ and $\Delta^2\mathbf{c}$, defined in (4), and the independence assumption of the model. Consequently, the most influential term is $J(\Delta^1\mathbf{c})$, because it has the smallest variance.

In Fig. A-1 we see the decomposition of the unconstrained cost function, $J^{sw}(\widetilde{\mathbf{W}}\mathbf{c})$, in solid bold line, into $J_1^{sw}(\mathbf{c})$, in solid thin line, $J_2^{sw}(\Delta^1\mathbf{c})$, in dashed line, and $J_3^{sw}(\Delta^2\mathbf{c})$, in dot dashed line, where equation (A-1) is satisfied in each iteration in the iterative algorithm demonstrated in Section IV-B.

In Fig. A-2 we see the evolution of $\|\mathbf{c}\|_2^2$, $\|\Delta^1\mathbf{c}\|_2^2$, and $\|\Delta^2\mathbf{c}\|_2^2$, corresponding to $J_1^{sw}(\mathbf{c})$, $J_2^{sw}(\Delta^1\mathbf{c})$, and $J_3^{sw}(\Delta^2\mathbf{c})$ in Fig. A-1, respectively. The norm of the static features is almost unchanged, (the change is about 0.01 $\%$ over 10000 iterations). However, the norm of the dynamic features indeed change at a rate comparable to the change in the cost function. As described above, the cost function is mostly changed due to the change in the dynamic feature norm rather than due to the change in the static feature norm. Consequently, the dynamic feature model error decreases without reducing essentially the norm of the static features. Indeed, we

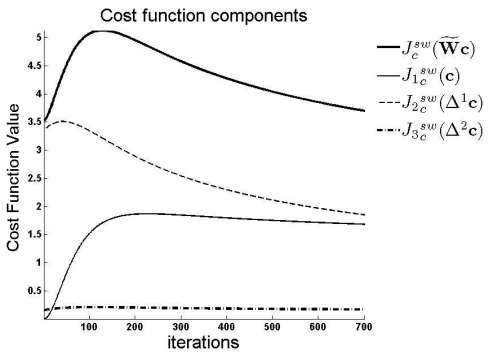


Fig. A-3. The components of the constrained segment-wise cost function, $J_c^{sw}(\widetilde{\mathbf{W}}\mathbf{c})$, where $J_c^{sw}(\widetilde{\mathbf{W}}\mathbf{c}) = J_{1c}^{sw}(\mathbf{c}) + J_{2c}^{sw}(\boldsymbol{\Delta^1}\mathbf{c}) + J_{3c}^{sw}(\boldsymbol{\Delta^2}\mathbf{c})$.

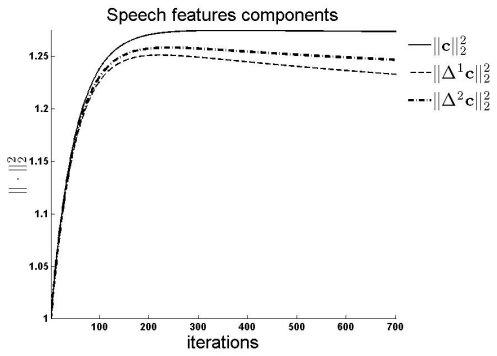


Fig. A-4. The normalized static feature norm and the normalized dynamic feature norm, obtained by the minimization of $J_c^{sw}(\widetilde{\mathbf{W}}\mathbf{c})$.

see in Fig. A-2 that the decrease in $\|\mathbf{c}\|_2^2$ is negligible compared to the decrease in $\|\Delta^1\mathbf{c}\|_2^2$ and $\|\Delta^2\mathbf{c}\|_2^2$.

The initial vector $\mathbf{c}_0$ in the iterative algorithm is constructed by replication of the model means, so, $\mathbf{c}_0$ lacks any dynamics in the intra-phoneme frames, but includes discontinuities at the inter-phonemes frames (phonemes boundaries). In Fig. A-1 we see that the initial value of $J_1(\mathbf{c})$ is zero but $J_2(\Delta^1\mathbf{c})$ and $J_3(\Delta^2\mathbf{c})$ are not zero.

From Fig. A-2 we conclude that the transitions between adjacent states are smoothed as the number of iterations increases, because $\|\Delta^1\mathbf{c}\|_2^2$ and $\|\Delta^2\mathbf{c}\|_2^2$ get smaller.

The above discussion is related to the unconstrained optimization problem. When the cost function $J_c^{sw}(\mathbf{W}\mathbf{c})$ includes an additional term, $\frac{\lambda}{2}\|\mathbf{c}\|_2^2$, regulating the norm of the static feature vector, there is an increase in the static feature vector norm, accompanied with a cost function error increase. The cost function error increase occurs as long as the balancing factor λ does not decrease sufficiently. When λ does decrease sufficiently the cost function components start competing to reduce their errors according to their significance (inverse variance), where $J_1(\Delta^1\mathbf{c}) + J_2(\Delta^2\mathbf{c})$ is more significant compared to $J(\mathbf{c})$. In Fig. A-3 we see the cost function components dynamics in the norm constrained case. The corresponding feature vector dynamics is shown in Fig. A-4. We see that $\|\mathbf{c}\|_2^2$ converges in about 150 iterations.

REFERENCES

[1] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, Issue 2, Feb. 1989, pp. 257-286.

[2] S.Furui, "Speaker independent isolated word recognition based on dynamics emphesized cepstrum," Trans. IECE of Japan, vol. 69, no.12, Dec. 1986, pp. 1310-1317.

[3] F.K.Soong, A.E.Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," Proc. ICASSP 1986 (Tokyo, Japan), Apr. 1986, pp. 877-880.

[4] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis from HMMs using dynamic features," Proc. ICASSP 1996, vol.1, 1996, pp. 389-392.

[5] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," Proc. EUROSPEECH 1995, pp. 757-760.

[6] T. Toda, K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," INTERSPEECH 2005, Lisbon, Portugal, Sept. 4-8, 2005. pp. 2801-2804.

[7] H. Zen, K. Tokuda, T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," Computer Speech and Language, vol.21, no.1, Jan. 2007, pp. 153-173.

[8] D. Chazan, R. Hoory, Z. Kons, A. Sagi, S. Shechtman and A. Sorin, "Small footprint concatenative text-to-speech synthesis using complex envelop modeling", INTERSPEECH 2005, pp. 2569-2572.

[9] R.E. Donovan, "Trainable speech synthesis", Phd thesis, Cambrige, June 1996.

[10] R.E. Donovan, and E.M.Eide, "The IBM Trainable Speech Synthesis System", Proc. ICSLP 1998, Sydney, Australia, vol.5, pp. 1703-1706.

[11] Francis B. Hildebrand, *Finite-Difference Equations and Simulations*, Section 2.2, Prentice-Hall, Englewood Cliffs, New Jersey, 1968.

[12] K. Schnel and A. Lacroix, "Combination of LSF and Pole Based Parameter Interpolation for Model Based Diphone Concatenation", INTERSPEECH 2007, pp. 2897-2900.

[13] R. E. Donovan, "Topics in decision tree based speech synthesis", Computer Speech & Language vol. 17, Issue 1, Jan. 2003, pp. 43-67.

[14] R. E. Donovan, "Text-to-speech using clustered context-dependent phoneme-based units ", US Patent 6163769, issued on Dec. 19, 2000.

[15] "Mean Opinion Score (MOS)", Recommendation P.800, Telecommunication Standardization Sector, International Telecommunication Union (ITU-T), Geneva, Switzerland.

[16] A.S. House, C.E. Williams, M.H.L. Hecker, and K.D. Kryter, Psychoacoustic speech tests: A modified rhyme test, Tech. Rep. ESDTDR- 63-403, U.S. Air Force Systems Command, Hanscom Field, Electronics Systems Division, 1963.

[17] C. Benot, M. Grice, and V. Hazan, The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences, Speech Commun., vol.18, pp.381 392, 1996.

[18] "$http://www.synsig.org/index.php/Main_Page$", SynSIG, ISCA, the International Speech Communication Association

[19] D.Chazan, R.Hoory, A.Sagi, S. Shechtman, A. Sorin, Z. Shuang and R. Bakis, "High quality sinusoidal modeling of wideband speech for the purpose of speech synthesis and modification", ICASSSP 2006, Toulouse, May 2006.

[20] "$http://hts.sp.nitech.ac.jp/nitech-hts\_blizzard2005$", Nitech-HTS samples for Blizzard Challenge 2005.