A Hybrid Text-to-Speech System That Combines Concatenative and Statistical Synthesis Units

Stas Tiomkin, David Malah, *Life Fellow, IEEE*, Slava Shechtman, and Zvi Kons

***Abstract*—Concatenative synthesis and statistical synthesis are the two main approaches to text-to-speech (TTS) synthesis. Concatenative TTS (CTTS) stores natural speech features segments, selected from a recorded speech database. Consequently, CTTS systems enable speech synthesis with natural quality. However, as the footprint of the stored data is reduced, desired segments are not always available in the stored data, and audible discontinuities may result. On the other hand, statistical TTS (STTS) systems, in spite of having a smaller footprint than CTTS, synthesize speech that is free of such discontinuities. Yet, in general, STTS produces lower quality speech than CTTS, in terms of naturalness, as it is often sounding muffled. The muffling effect is due to over-smoothing of model-generated speech features. In order to gain from the advantages of each of the two approaches, we propose in this work to combine CTTS and STTS into a hybrid TTS (HTTS) system. Each utterance representation in HTTS is constructed from natural segments and model generated segments in an interweaved fashion via a hybrid dynamic path algorithm. Reported listening tests demonstrate the validity of the proposed approach.**

***Index Terms*—Concatenative text-to-speech (CTTS), dynamic path, hybrid TTS, statistical TTS, TTS synthesis.**

I. Introduction

THERE are two main approaches for solving the text-to-speech (TTS) paradigm. The first one uses recorded speech feature segments, which may be words, phonemes, or even sub-phonemes. 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. Speech generated by this approach 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 the possible appearance of discontinuities at segment boundaries due to imperfect concatenation. The smaller the size of the stored database, the larger is the number of discontinuities that typically appear in the generated speech. Thus, in applications where storage and computational resources are limited, such as in mobile devices, a small footprint system is necessary, resulting in reduced quality of CTTS generated speech. An example for such a system is IBM's CTTS system [1], which we used in this research. A general description of the concatenative speech synthesis methods exploited in [1] is detailed in [2]. Other typical concatenative TTS systems are detailed in [3] and [4]. Similarly to the baseline CTTS system used in the current work, these systems synthesize speech by selecting and concatenating natural speech units from a inventory.

The other TTS approach employs statistical models for speech production [8] and is called statistical TTS (STTS). A thorough review of statistical parametric synthesis is provided in [9]. 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 by interpolation between 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 speech.

In this paper, we propose to combine the advantageous traits of CTTS with those of STTS into another kind of TTS systems—hybrid text-to-speech systems—denoted HTTS. The proposed HTTS system optimally (with respect to a suitable cost function) interweaves natural segments with statistical model-generated segments via a hybrid dynamic-path algorithm developed in this work.

The term hybrid TTS has been used in previous works, but in different ways. One hybrid approach is described by Pollet *et al.* in [10]. Additional systems exploiting hidden Markov models (HMMs) for units selection appear at [11]–[16]. The main and critical difference between those works and ours is that they use HMMs for the selection of natural speech from an inventory, and we do not. An approach for combining a concatenative TTS system with HMM-based target prosody is described in [17]. It differs from our proposed system in that we use interchangeably natural segments of a baseline concatenative system together with statistically generated speech segments.

Another hybrid approach is described in [18]. In this paper, an STTS is the "backbone" of the described hybrid system, with CTTS being used to improve the speech trajectory generated by STTS. Two sequences are generated per utterance: a CTTS-generated sequence and a STTS-generated sequence. The HMMs of the STTS are adapted by the CTTS sequence according to a proposed weight function. An approach similar in some ways to [18] is described in [19]. Here, speech is generated by HMMs, but with natural units of CTTS being used for the HMM means. Thus, it is a kind of post-training model adaptation, which is not applied in our hybrid system.

Manuscript received February 18, 2010; revised June 05, 2010; accepted October 08, 2010. Date of publication October 25, 2010; date of current version May 13, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mark Gales.

S. Tiomkin is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel, and also with the Speech Technologies Group, IBM Research Laboratory, Haifa University Campus, Mount Carmel, Haifa 31905, Israel (e-mail: stasti@gmail.com).

D. Malah is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel, (e-mail: malah@ee.technion.ac.il).

S. Shechtman and Z. Kons are with the Speech Technologies Group, IBM Research Laboratory, Haifa University Campus, Mount Carmel, Haifa 31905, Israel (e-mail: slava@il.ibm.com; zvi@il.ibm.com).

Digital Object Identifier 10.1109/TASL.2010.2089679

Other hybrid approaches were investigated in [20] and [21]. In these works, statistically generated segments are introduced according to different data sparsity criteria, i.e., according to the availability of natural segments in certain contexts.

The main components constituting our proposed HTTS system are the mentioned hybrid dynamic-path algorithm that allocates natural segments along with statistical boundary-constrained model-generated segments, and a corresponding hybrid speech feature-vector generating algorithm. Thus, the proposed HTTS system inherits the naturalness of the corresponding baseline CTTS and the smooth transitions of STTS. Moreover, the proposed HTTS system is a generalization of both CTTS and STTS, because it can work in either a pure CTTS mode or a pure STTS mode, depending on a *hybridism ratio* parameter, which affects the ratio of the number of natural segments to the number of statistically generated segments comprising a synthesized utterance. Speech generated in an intermediate (hybrid) mode consists of both natural and statistically generated segments, interweaved within an utterance.

Obviously, the quality of the hybrid system depends on the qualities of the baseline CTTS and STTS systems. In particular, different STTS systems result in hybrid systems having different qualities. Consequently, an enhanced STTS system will provide a better HTTS. To demonstrate this aspect we compared the quality of HTTS that uses a conventional STTS to the quality of HTTS that uses a STTS with improved dynamics, developed in our earlier work [22]. It is based on a segment-wise STTS, denoted as SW-STTS. Speech features generated by SW-STTS are less smooth than in a conventional STTS, and as a result, the generated speech sounds more natural. Using SW-STTS in our hybrid system, resulted in better naturalness than with a conventional STTS, as confirmed by listening tests.

To generate a hybrid speech feature-vector over an entire utterance, we developed an iterative algorithm. The algorithm constrains certain frames to remain unchanged (CTTS frames), while affecting free statistical frames (the remaining frames). This algorithm is based on a gradient descent formulation with linear constraints. This algorithm, along with the proposed hybrid dynamic-path algorithm, are at the core of the proposed HTTS system.

This paper is organized as follows. In Section II, we provide the essentials of the baseline CTTS and STTS methodologies. In Section III, we present the proposed HTTS system. In Section IV, we provide experimental results and discuss them. Finally, in Section V we conclude this research and suggest future work that can be pursued.

II. Baseline-CTTS and STTS Systems

In this section, we briefly describe the baseline CTTS and STTS systems that are used in the proposed hybrid system.

A. Concatenative Text-to-Speech (CTTS) Synthesis

In concatenative systems, speech is synthesized by concatenating natural speech segments, or speech segment features, denoted in the literature as candidates, units, or segments. These segments are basic elements for speech synthesis. In different systems, these segments represent phonemes or sub-phonemes. The current research is based on the IBM embedded concatenative text to speech system, which uses sub-phonemes as basic segments/units, as detailed in [1] and [2]. The main functional blocks of CTTS are: 1) a language dependent text to phoneme processor; 2) acoustic context-dependent decision trees for the different sub-phonemes, holding in their leaves parametric representations of natural speech segment features; 3) target-prediction trees, holding context dependent target energy, pitch, and duration, which are combined with the values predicted by a rule-based phonetic text analyzer, detailed in [5]; 4) a dynamic search algorithm, producing the optimal sequence (in a given inventory) of natural speech-segment features that matches target values and minimizes a concatenation distance between segments; 5) a speech generator that composes speech from the speech feature sequence found by the dynamic search. Because the proposed hybrid TTS system modifies the dynamic search algorithm, we outline here the algorithm used in the reference CTTS system.

Each acoustic leaf, which represents a particular phoneme (or sub-phoneme) in a phoneme sequence comprising an input text, holds a number of candidates (segments), for generating a speech utterance. Clearly, the more candidates are in any given acoustic leaf the higher the quality of generated speech. However, there is an exponential number of possible combinations of candidates to compose an utterance. Dynamic programming with an appropriate cost function is applied to find an optimal combination of candidates. A description of the baseline IBM embedded CTTS system is provided in [6].

All possible segment concatenations are examined during a forward pass of the dynamic search. The best path is found on this fully connected trellis by back-tracing. The best path segments are sent to the speech generator, which uses overlap and add synthesis. In Section III-A, we propose a modified dynamic search, enabling the interweaving of natural segments with statistically generated segments, which is the essence of the proposed HTTS system.

B. Statistical Text-to-Speech (STTS) Synthesis

In this subsection, we briefly describe the conventional approach for deriving an entire utterance speech feature-vector in a statistical HMM-based TTS. The details of the statistical model appear in [8] and [9].

1) Speech Parameters: In this research, the log-amplitude of the speech spectrum $\mathbf{A}(f')$ of every frame is approximated by a linear combination of triangular basis functions $\mathbf{B}_k(f')$, $k = 1, 2, \ldots, d$, as follows:

$$log(\mathbf{A}(f')) \approx \sum_{k=1}^{d} c_k \cdot \mathbf{B}_k(f') \tag{1}$$

where f' denotes a mel-scale frequency.[1] $\mathbf{B}_k(f')$ are used as a basis for speech spectrum expansion, with expansion coefficients c_k rather than a filter bank as in a conventional MFCC representation.

This representation is successfully used in an IBM's CTTS system, detailed in [23], and a corresponding speech reconstruction unit is detailed in [24].

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

2) Statistical Speech Features Representation: 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 $\mathbf{c}_i = (c_i(1), c_i(2), \ldots, c_i(d))^T$ are the expansion coefficients, introduced in (1). $\mathbf{c}_i$ denotes the static feature vector of dimension $d \times 1$ of the ith frame, where $d = 32$. The prosody (pitch, energy and duration) is modeled by context-dependent regression trees, detailed in [1], [2], and [7].

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

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

where $\mathbf{o_i} = (\mathbf{c}_i^T, \Delta^1\mathbf{c}_i^T, \Delta^2\mathbf{c}_i^T)^T$.

The vector $\mathbf{o}$, over an entire utterance, can be obtained from $\mathbf{c}$ by a linear transformation

$$\mathbf{o}_{3MN\times 1} = \mathbf{W}_{3MN\times MN}\mathbf{c}_{MN\times 1} \tag{4}$$

where the matrix W is constructed according to the first and second difference vectors $\Delta^1\mathbf{c}_i$ and $\Delta^2\mathbf{c}_i$, respectively, as detailed in [8].

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

$$\mathbf{o}^{\text{opt}} = \underset{\mathbf{o}}{\text{argmax}}\ P(\mathbf{o} \mid \lambda) \tag{5}$$

where $P(\mathbf{o}|\lambda) = \sum_{\forall \mathbf{q}} P(\mathbf{o}, \mathbf{q}|\lambda)$, and $\mathbf{q} = (q_1, q_2, \ldots, q_N)$ 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 [25]. Hence, 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 concatenating corresponding context-dependent HMMs, where contexts are derived from a phonetic analysis of the synthesized text [2].

As mentioned in Section II-B2, the prosody is modeled by context-dependent regression trees, which provide the phonetic identities of states and their durations. Hence, we can reduce the general problem of solving (5) to the following problem, which assumes that the state sequence $\mathbf{q}$ is given as

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

Methods for full HMM-based speech feature synthesis appear in [8].

In this paper, as in many other works that are described in the recent review [9] on TTS systems, we use a single Gaussian model with a diagonal covariance matrix.

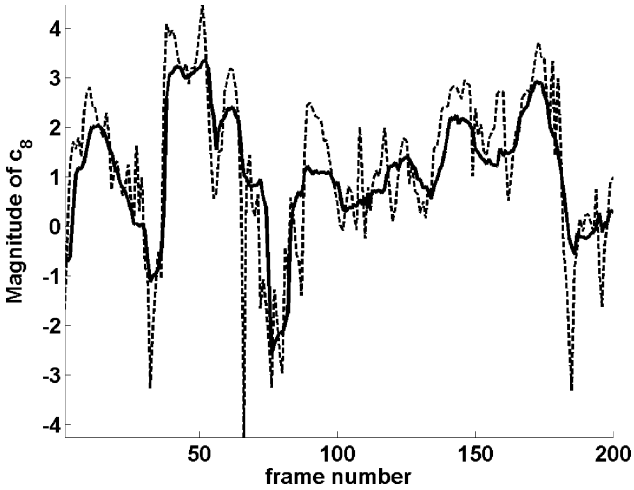


Fig. 1. Variation in time of the eighth expansion coefficient, c_8, in the utterance "Many problems in reading and writing are due to old habits": c_8^{opt} in solid line; c_8^{natural} in dashed line.

Under such assumptions, the logarithm of $P(\mathbf{o} \mid \mathbf{q}, \lambda)$ can be written as

$$ln(P(\mathbf{o} \mid \mathbf{q}, \lambda)) = -\frac{1}{2}(\mathbf{o} - \mathbf{m})^T\mathbf{U}^{-1}(\mathbf{o} - \mathbf{m}) \tag{7}$$

with $\mathbf{m} = [\mathbf{m}_{q_1}^T, \mathbf{m}_{q_2}^T, \ldots, \mathbf{m}_{q_N}^T]^T$ and $\mathbf{U}^{-1} = \text{diag}[\mathbf{U}_{q_1}^{-1}, \mathbf{U}_{q_2}^{-1}, \ldots, \mathbf{U}_{q_N}^{-1}]$, 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, respectively.

To find an optimal solution over an entire utterance, (4) is used to define the following cost function:

$$J(\mathbf{Wc}) = -\ln P(\mathbf{Wc} \mid \mathbf{q}, \lambda). \tag{8}$$

The optimal solution $\mathbf{c}^{\text{opt}}$ that minimizes $J(\mathbf{Wc})$ is given by

$$\mathbf{c}^{\text{opt}} = (\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{U}^{-1}\mathbf{m}. \tag{9}$$

We can see in Fig. 1 that, typically, the optimal solution (9) is over-smoothed and has much less dynamics (inter-frame variations), as compared to the corresponding natural speech features.

Perceptually, the reduced variance in speech features is associated with muffled sound, as was indicated by listening, and as also reported in [8].

In the following subsection, we briefly describe the concept of segment-wise (SW) model representation, detailed in [22], found to improve the naturalness of statistically generated speech.

C. Improved Quality STTS Using a Segment-Wise Representation With a Norm Constraint

As discussed earlier, the insufficient speech features dynamics in conventional STTS systems causes over-smoothing of statistically generated speech features, resulting in muffled speech.

The SW-STTS speech synthesis method [22] generates iteratively a statistical speech feature-vector, applying a constraint to the speech feature-vector norm, preventing its reduction. The generated speech sounds more natural and less muffled, as

confirmed by listening tests. The SW-STTS approach exploits the segment-wise augmented-space representation that allows speech features to fluctuate around model means, rather than to follow them tightly. In SW-STTS, the model mean $\mathbf{m}_{q_i}$ is not replicated T_i times, but rather, it is approximated by the average of T_i augmented-space vectors, $\mathbf{o}_{t_i}, \ldots, \mathbf{o}_{t_i+T_i-1}$. The average augmented-space feature vector $\bar{\mathbf{o}}_i$ is defined as

$$\bar{\mathbf{o}}_i = \frac{1}{T_i}\sum_{k=1}^{T_i} \mathbf{o}_{t_i+k-1} \tag{10}$$

where $\mathbf{o}_k$ is the conventional augmented-space feature vector, mentioned in (3) and T_i is length of ith segment. Consequently, the segment-wise transformation for the ith segment, having T_i frames, is

$$\widetilde{W}_i = \frac{1}{T_i}\begin{pmatrix} \mathbf{0} & \mathbf{1} & \underset{T_i-2}{\cdots \mathbf{1} \cdots} & \mathbf{1} & \mathbf{0} \\ -\frac{\mathbf{1}}{\mathbf{2}} & -\frac{\mathbf{1}}{\mathbf{2}} & \underset{T_i-2}{\cdots \mathbf{0} \cdots} & \frac{\mathbf{1}}{\mathbf{2}} & \frac{\mathbf{1}}{\mathbf{2}} \\ -\mathbf{1} & \mathbf{1} & \underset{T_i-2}{\cdots \mathbf{0} \cdots} & \mathbf{1} & -\mathbf{1} \end{pmatrix}_{3d\times d(T_i+2)} \tag{11}$$

where all the matrix elements in (11) are diagonal block matrices of dimension $d \times d$ each, where d is defined in (1). The corresponding cost function $J(\bar{\mathbf{o}}_i)$ constructed without replication of the model of the state q_i is

$$J(\bar{\mathbf{o}}_i) = \frac{1}{2}\|\mathbf{U}_{q_i}^{-1/2}(\bar{\mathbf{o}}_i - \mathbf{m}_{q_i})\|_2^2 \tag{12}$$

where $\bar{\mathbf{o}}_i$, $\mathbf{m}_{q_i}$, and $\mathbf{U}_{q_i}$ are the average augmented feature vector, the model mean and the model covariance matrix of state q_i, respectively.

It was observed that the squared-norm of statistically generated speech feature-vectors $\|\mathbf{c}^{stt}\|_2^2$ is often quite lower than the squared-norm of natural speech feature-vectors $\|\mathbf{c}^{\mathrm{nat}}\|_2^2$. Because the conventional solution, shown in (9), is a minimal norm least squares solution, and because of insufficient speech-features dynamics, the statistically generated speech feature-vector norm is very close to the statistical models features norm. The SW-STTS considers the following minimization problem [22]:

$$J_c^{sw}(\mathbf{c}) \doteq \frac{1}{2}\|\mathbf{U}^{-1/2}(\widetilde{\mathbf{W}}\mathbf{c} - \mathbf{m})\|_2^2 + \frac{\lambda}{2}\|\mathbf{c}\|_2^2 \tag{13}$$

where λ is considered as a parameter, balancing between the model error term and the norm term, rather than a free variable of the function. To minimize this cost function a gradient descent algorithm is applied as follows:

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \alpha_n \nabla(\mathbf{c}_n) \tag{14}$$

where $\nabla(\mathbf{c}_n)$ is the gradient of $J_c^{sw}(\mathbf{c})$ with respect to $\mathbf{c}$, computed at iteration n, and $\alpha_n = 1/\|\nabla(\mathbf{c}_n)\|_2^2$ is the step size. From (13)

$$\nabla(\mathbf{c}_n) = \widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\widetilde{\mathbf{W}}\mathbf{c}_n - \widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\mathbf{m} + \lambda\mathbf{c}_n. \tag{15}$$

Since the final feature vector should approximate well the models and also approximate a desired norm value, it is proposed in [22] to apply a negative balancing factor λ that decreases in its absolute value with n, 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. Consequently, (15) is replaced by

$$\nabla(\mathbf{c}_n) = \widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\widetilde{\mathbf{W}}\mathbf{c}_n - \widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\mathbf{m} + \lambda_n\mathbf{c}_n \tag{16}$$

where λ_n is updated according to

$$\lambda_{n+1} = \theta\lambda_n, \qquad 0 \le \theta \le 1. \tag{17}$$

In [22], an empirical relation between λ_0, the initial value of λ, and the final norm of the feature vectors is found, allowing a desired norm increase, which results in speech features with enhanced dynamics. In our experiments, we used $\theta = 0.95$, where an acceptable range of values for θ may reach 0.98.

III. Proposed Hybrid TTS System

The goal of the current research is to efficiently combine the advantages of CTTS and STTS into a hybrid TTS system (HTTS). The hybridism in the proposed system is in the interweaving of natural segments with statistically generated segments, using an appropriate hybrid dynamic-path algorithm.

The proposed system is based on the following. 1) A hybrid dynamic-path that defines positions for statistical models within an utterance. It aims to include as many as possible long natural segment sequences, and smooth out discontinuities by optimally connecting natural segments by statistically generated segments. The hybrid dynamic-path is used to determine boundaries (natural speech segments) for statistical units. 2) Boundary constrained statistical model. 3) A hybrid gradient descent algorithm with linear constraints, where statistical segments within a synthesized utterance are generated from a constrained statistical model, while natural segments stay unchanged and provide boundary conditions to the statistically generated parts.

A. Hybrid Dynamic Path Algorithm

The major disadvantage of concatenative speech synthesis systems is the existence of spectral discontinuities between some adjacent speech feature-vectors, causing unpleasant artifacts in the generated speech. These discontinuities can occur when originally contiguous natural segments are not available for concatenation from the speech database, as detailed in [2] and [7].

Theoretically, a perfect CTTS system, having an unlimited number of natural segments in any possible context, is able to concatenate natural segments in their natural order, as they appeared in the training sentences set, and hence is expected to have natural speech quality. Obviously, such a system is infeasible. Any feasible CTTS may only approximate the perfect CTTS system, trying to concatenate as many as possible originally contiguous natural segments.

Thus, in the proposed HTTS system, we aim to interweave natural segments with statistically generated segments, where the statistical segments are positioned to smooth discontinuities while enabling as long as possible natural segments sequences, as they appear in the training database. Consequently, we aim to better approximate the characteristic of the ideal CTTS.

Accordingly, we propose to determine the positions of statistical segments as follows. Assume that we have a sequence of contexts $L_1, L_2, \ldots, L_K$, representing the stages

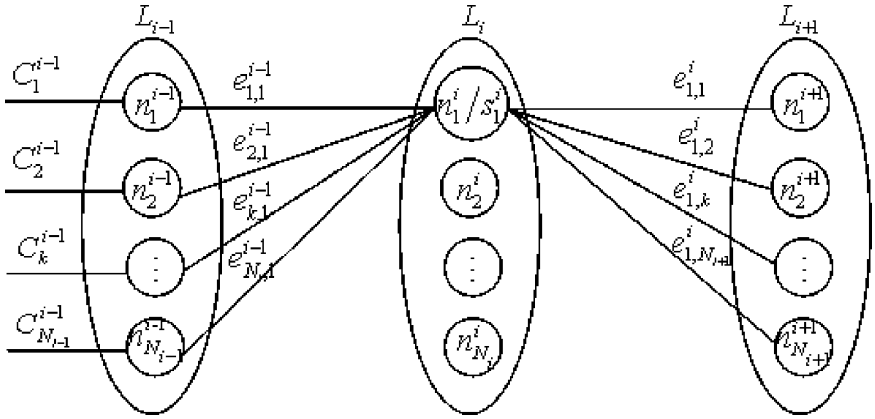


Fig. 2. Determination of statistical segments location. $e_{j,k}^{i-1}$ is the spectral distance between node n_j^{i-1} and node n_k^i, C_j^i is the best partial path at node n_j^i, L_i denotes ith stage of the dynamic search. The notation n_1^i/s_1^i means that natural segment n_1^i may be exchanged with statistical segment s_1^i.

of the dynamic search, where context L_i holds segments $n_1^i, n_2^i, \ldots, n_{N_i}^i$, representing the hybrid nodes, as shown Fig. 2, where C_k^{i-1} and $e_{k,1}^{i-1}$ denote the cumulative cost of a survivor path to node n_k^{i-1} and the transition cost between n_k^{i-1} and n_1^i, respectively. Note that since stage L_i represents a certain acoustic context, it includes corresponding natural segments and a statistical model of this acoustic context. Statistical segments are generated from this statistical model by applying proper boundary constraints. These boundary constraints are determined by the adjacent natural segments to a given statistical model. Thus, the total number of possible statistical segments considered in the dynamic search, resulting from a single statistical model in L_i and all possible boundary constraints, is $N_{i-1} \cdot N_i \cdot N_{i+1}$, where N_i is the number of natural segments in L_i.

Any node n_j^i can be replaced by a statistical segment s_j^i as described below, where s_j^i is generated by the boundary-constrained statistical model, described in Section III-B, to ensure smooth connections to adjacent natural segments. The decision about this exchange is done when back-tracing the best path. If n_j^i will be replaced by s_j^i, the left boundary of s_j^i is constrained by a segment from stage L_{i-1}. Consequently, s_j^i is generated from the boundary-constrained model, where boundaries are determined by neighbors of s_j^i in the optimal path found by the dynamic search.

So, s_j^i is adjusted exactly to its neighboring natural segments, which define the boundaries for the corresponding statistical model of s_j^i. Obviously, s_j^i depends on its neighboring natural segments, and it is generated dynamically for each pair of adjacent natural segments, which do not connect smoothly.

In a CTTS system the most appropriate segments are concatenated by means of the Viterbi algorithm, which gradually advances from the first stage L_1 to the final stage L_K, computing a survivor path to each node in each stage in order to find the optimal path by back-tracing through the best survivor path.

In Fig. 2, when computing a survivor to node n_1^i, the first node of stage L_i, the existence of the following condition is examined:

$$e_{j,1}^{i-1} \geq \epsilon, \quad \forall j, \quad j = 1, 2, \ldots, N_{i-1} \tag{18}$$

where ϵ is a permitted spectral distance (error). The spectral distance is defined in this work as the norm of a difference between the feature vector $\mathbf{c}$ [defined in (2)] of the last frame of a given natural segment and the feature vector of the first frame of the following natural segment.

If (18) holds, then any path passing through n_1^i includes a spectral discontinuity at the transition from L_{i-1} to L_i. Consequently, generated speech quality could be degraded by this spectral discontinuity. Hence in such a case, it is proposed to replace n_1^i by a boundary-constrained statistical model s_1^i. Since s_1^i is determined such that it is constrained to smoothly connect to its neighbors, the survivor path to s_1^i is determined by

$$p_{s_1^i} = \underset{j}{\operatorname{argmin}}\, C_j^{i-1} \tag{19}$$

which means that s_1^i continues smoothly the best survivor path from stage L_{i-1} to stage L_i, because no transition cost is added for this transition.

Although, the natural node n_1^i is replaced by the statistical node s_1^i, the spectral distance from the nodes at stage L_{i+1} to the first node of L_i is still computed as the spectral distance between n_1^i and n_j^{i+1}, $j = 1, \ldots, N_{i+1}$. This way we give a higher priority to dynamic paths that include natural segments with smaller spectral distance in the original pure CTTS system dynamic paths.

If node n_1^i connects smoothly to some node n_j^{i+1} of L_{i+1}, (i.e., the condition in (18) is not satisfied), and this transition is included in the optimal sequence of segments computed by the search algorithm, then the right boundary of s_1^i, which replaces n_1^i, is constrained by n_j^{i+1}. If not, then, the right neighbor of s_1^i will be a statistical model, corresponding to the acoustic model of L^{i+1} with appropriate boundary constraints, determined by dynamic search. In the former case, there is a discontinuity left to n_1^i and a smooth connection right to it. Hence, s_1^i connects smoothly segment sequences before and after n_1^i. In the latter case, there are no smooth connections both to the right of n_1^i and to the left of n_1^i in the original path. So, to remove these discontinuities, a number of statistical models are introduced, replacing CTTS segments up to the stage where a smooth connection exists to the next stage. Speech feature parameters at this location are generated by the sequence of boundary-constrained

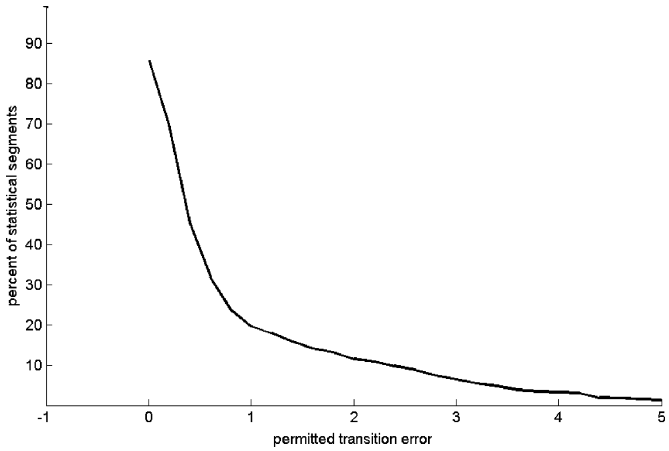


Fig. 3. Illustrative relation between the *hybridism ratio* parameter, (the ratio of the number of statistical segments to the overall number of segments in an utterance) ξ and the permitted spectral distance ϵ. Note that in the range $\epsilon \in [-1, 0]$, $\xi = 100\%$.

statistical models, which connect optimally to their adjacent natural segments. Consequently, statistical speech features are generated dynamically during synthesis. This way a possible local discontinuity is alleviated.

As a result, most of possible discontinuities disappear from the hybrid path, while contiguous sequences of natural segments are given preference.

The number of statistical segments within a generated utterance is affected by the value of the permitted spectral distance parameter ϵ. Setting the permitted distance parameter error value to any negative number results in STTS, because the spectral distance is, by definition, non-negative, and hence, any natural segment would be replaced by a statistical one. Whereas, setting the value of ϵ in a range of positive numbers, results in HTTS. Finally, setting the value of ϵ to a very large positive number results in CTTS. The special case, when the permitted spectral distance value is set to zero, is considered to be an unforced hybrid TTS mode, since statistical models are introduced any time two natural segments do not connect with zero distance.

The proposed HTTS system is a generalization of both CTTS and STTS, because it can work in either a pure CTTS mode or a pure STTS mode, depending on a hybridism ratio parameter ξ which is the ratio of the number of statistical segments to the overall number of segments in an utterance.

We established an empirical relation between ξ and the permitted value of ϵ, for the used database, shown in Fig. 3. This relation was found by synthesizing an arbitrary set of 40 sentences in English by the proposed hybrid TTS, having a footprint of 8.3 MB. This set of sentences was generated using $\epsilon = 0.2\,\mathrm{k}$, $k = 0, 1, \ldots, 25$, and one negative value of $\epsilon = -1$. For each permitted spectral distance from the used range, the value of ξ was computed as the average ratio over these sentences. This relation was found for a given female speaker voice used in the IBM CTTS, which is constructed according to [1] and [2].

This empirical relation does not define strictly the number of statistical segments within a generated utterance, but rather it illustrates the effect of the permitted distance parameter ϵ on their number. More research is needed to find which parameters (such as the permitted spectral distance, initial database size, pre-selection ratio, etc.) of a baseline CTTS system affect directly the hybridism ratio parameter ξ.

B. Boundary Constrained Model

For STTS, it is shown in Section II-B (item 3) that the optimal solution $\mathbf{c}^{\mathrm{opt}}$ is the most probable statistically-derived vector over an utterance of N frames. In a HTTS system we have an arbitrary number of natural frames along the utterance. Consequently, we would like to synthesize the optimal vector, given these natural frames.

The smooth connection of natural segments to statistical segments, within an entire speech feature vector $\mathbf{c}_{dN\times 1}$, is done by means of $\mathbf{\Delta}^{1,2}\mathbf{c}_i$ as follows.

Assume that we have to connect the natural segment $\mathbf{c}^{\mathrm{nat}} = [\mathbf{c}_1^{\mathrm{nat}}, \mathbf{c}_2^{\mathrm{nat}}, \ldots, \mathbf{c}_{T_i}^{\mathrm{nat}}]$, having T_i frames, to the left boundary of the statistically generated segment $\mathbf{c}^{stt} = [\mathbf{c}_1^{stt}, \mathbf{c}_2^{stt}, \ldots, \mathbf{c}_{T_j}^{stt}]$, having T_j frames. This connection is done by using the left boundary dynamic features $\mathbf{c}^{stt}$, (based on [8])

$$\widetilde{\mathbf{\Delta}}^1\mathbf{c}_1 = \frac{1}{2}(\mathbf{c}_2^{stt} - \mathbf{c}_{T_i}^{\mathrm{nat}}) \tag{20a}$$

$$\widetilde{\mathbf{\Delta}}^2\mathbf{c}_1 = (\mathbf{c}_{T_i}^{\mathrm{nat}} - 2\mathbf{c}_1^{stt} + \mathbf{c}_2^{stt}) \tag{20b}$$

instead of the unconstrained features

$$\mathbf{\Delta}^1\mathbf{c}_1 = \frac{1}{2}(\mathbf{c}_2^{stt} - \mathbf{c}_{T_i}^{stt}) \tag{20c}$$

$$\mathbf{\Delta}^2\mathbf{c}_1 = (\mathbf{c}_{T_i}^{stt} - 2\mathbf{c}_1^{stt} + \mathbf{c}_2^{stt}). \tag{20d}$$

This arrangement enables composition of an hybrid speech feature-vector, where some frames are constrained to belong to natural segments while others are statistically generated segments. Below we provide a general framework for the generation of an hybrid speech feature-vector over an entire utterance with an arbitrary number and positions of natural frames, including the situation in which an entire utterance is composed totally of either natural segments or statistically generated segments, as in CTTS and STTS systems, respectively.

The corresponding constrained optimization problem is now as follows:

$$\mathbf{c}^{\mathrm{opt}} = \underset{\mathbf{c}}{\operatorname{argmin}} \ \ln(P(\mathbf{W}\mathbf{c})) \tag{21a}$$

$$\text{s.t.} \quad \mathbf{A}\mathbf{c} = \mathbf{c}^*. \tag{21b}$$

$\mathbf{c}^*$ represents a given natural unit sequence and $\mathbf{A}$ is a design matrix for selecting which units should be composed from the natural units. By setting $\mathbf{A}\mathbf{c} = \mathbf{c}^*$, we can force the statistical speech feature generation algorithm to include these natural units. Let $\mathbf{c}^*_{dk\times 1} = [\mathbf{c}^{*\,T}_{i_1}, \mathbf{c}^{*\,T}_{i_2}, \ldots, \mathbf{c}^{*\,T}_{i_k}]^T$ be a vector that is composed of the k constrained natural frames, $\mathbf{c}^{*\,T}_{i_n}$, $n = 1, 2, \ldots, k$, at positions $i_1, i_2, \ldots, i_k$, respectively, and $\mathbf{A}_{dk\times dN}$ is a linear transformation from $\mathbf{c}_{dN\times 1}$ to $\mathbf{c}^*_{dk\times 1}$. The a_{i_n}th row of $\mathbf{A}$ is thus defined by

$$\mathbf{a}_{i_n} = [\mathbf{0}_{1\times(d(i_n-1)(1+i_n-1))/2}, \mathbf{1}_{1\times d}, \mathbf{0}_{1\times(d(i_n+1)(i_n+1+N))/2}]_{1\times dN}. \tag{22}$$

The optimal solution for this constrained optimization problem is derived by means of a Lagrangian function with a vectorial Lagrange multiplier $\boldsymbol{\gamma}_{dk\times 1}$

$$L(\mathbf{c},\boldsymbol{\gamma}) = \ln(P(\mathbf{Wc})) + (\mathbf{Ac}-\mathbf{c}^*)^T\boldsymbol{\gamma}. \tag{23}$$

Consequently, the speech feature vector for the boundary constrained model is derived by $\partial L(\mathbf{c},\boldsymbol{\gamma})/\partial\mathbf{c} = \mathbf{0}$, resulting in

$$\mathbf{c}^{\text{opt}} = (\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{U}^{-1}\mathbf{m} + (\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W})^{-1}\mathbf{A}^T\boldsymbol{\gamma}. \tag{24}$$

Using (24) in (21b), $\boldsymbol{\gamma}$ is given by

$$\begin{aligned}\boldsymbol{\gamma} = &(\mathbf{A}(\mathbf{W^T U^{-1} W})^{-1}\mathbf{A}^T)^{-1}\mathbf{c}^* \\ &- (\mathbf{A}(\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W})^{-1}\mathbf{A}^T)^{-1} \\ &\cdot \mathbf{A}(\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{U}^{-1}\mathbf{m}.\end{aligned} \tag{25}$$

We can see in Fig. 4 that the boundary constrained optimal solution has obviously two different types of frames: natural frames, and frames pertaining to statistically generated segments that have less variation in time. This mismatch typically results in unpleasant artifacts in the generated speech.

In the next section, we demonstrate an approach which resolves the above-mentioned mismatch by applying SW-STTS instead of the conventional STTS.

C. Hybrid Speech Generation Algorithm

We propose to combine the hybrid speech feature-vector representation, shown in Section III-B with the SW-STTS approach shown in Section II-C.

Consider the minimization of the doubly constrained cost function $J_{c,c}(\widetilde{\mathbf{W}}\mathbf{c})$ which is an extension to the norm-constrained cost function, defined in (13)

$$J_{c,c}(\widetilde{\mathbf{W}}\mathbf{c}) = \|\mathbf{U}^{-1}(\widetilde{\mathbf{W}}\mathbf{c}-\mathbf{M})\|_2^2 + (\mathbf{Ac}-\mathbf{c}^*)^T\boldsymbol{\gamma} + \lambda\|\mathbf{c}\|_2^2 \tag{26}$$

where the first term aims to approximate the statistical models, the second term constrains required frames to natural segments, and finally, the last term enhances speech features dynamics by systematically increasing speech feature-vector norm, as described in Section II-C. Using a gradient descent iterative algorithm, the hybrid speech feature vector obtained after the nth iteration is

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \alpha_n\widetilde{\nabla}(\mathbf{c}_n) \tag{27}$$

where $\widetilde{\nabla}(\mathbf{c}_n)$ is the gradient of $J_{c,c}(\widetilde{\mathbf{W}}\mathbf{c})$, which is

$$\begin{aligned}\widetilde{\nabla}(\mathbf{c}_n) &= \widetilde{\mathbf{W}}^{\mathbf{T}}\mathbf{U}^{-1}\widetilde{\mathbf{W}}\mathbf{c}_n - \widetilde{\mathbf{W}}^{\mathbf{T}}\mathbf{U}^{-1}\mathbf{m} + \mathbf{A}^T\boldsymbol{\gamma} + \lambda_n\mathbf{c}_n \\ &= \mathbf{P}\mathbf{c}_n - \mathbf{Q} + \mathbf{A}^T\boldsymbol{\gamma} + \lambda_n\mathbf{c}_n\end{aligned} \tag{28}$$

where $\widetilde{\mathbf{W}}^{\mathbf{T}}\mathbf{U}^{-1}\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{W}}^{\mathbf{T}}\mathbf{U}^{-1}\mathbf{m}$ are denoted as $\mathbf{P}$ and $\mathbf{Q}$, respectively.

It can be quite simply shown that $\mathbf{A}\widetilde{\nabla}(\mathbf{c}_n) = 0, \forall n$. Hence, we can compute the vectorial Lagrangian multiplier $\boldsymbol{\gamma}$ as follows:

$$\mathbf{APc}_n - \mathbf{A}Q + \mathbf{AA}^T\boldsymbol{\gamma} + \lambda_n\mathbf{Ac}_n = 0 \tag{29}$$

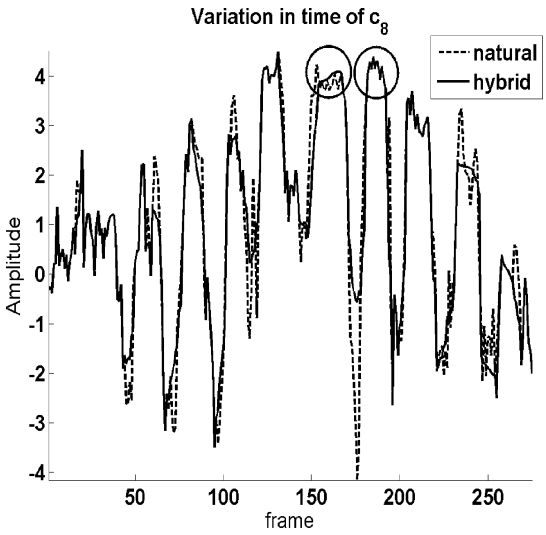


Fig. 4. Variation of c_8 in time, over an entire utterance. Dashed line: all segments are natural (CTTS). Solid line: hybrid system result, in which the statistical segments are determined by using a conventional STTS. (a) The entire utterance. (b) Zooming into the circled segments.

where $\mathbf{AA}^T = \mathbf{I}$ by the definition of $\mathbf{A}$. Consequently,

$$\begin{aligned}\boldsymbol{\gamma} &= \mathbf{AQ} - \mathbf{APc}_n - \lambda_n\mathbf{Ac}_n \underset{\mathbf{Ac}_n=\mathbf{c}^*,\forall n}{=} \\ &= \mathbf{AQ} - \mathbf{APc}_n - \lambda_n\mathbf{c}^*.\end{aligned} \tag{30}$$

The gradient in the update step (27) is

$$\widetilde{\nabla}(\mathbf{c}_n) = \mathbf{Pc}_n - \mathbf{Q} + \mathbf{A}^T\mathbf{AQ} - \mathbf{A}^T\mathbf{APc}_n - \lambda_n\mathbf{A}^T\mathbf{c}^* + \lambda_n\mathbf{c}_n \tag{31}$$

where λ_n is updated by the rule given in Section II-C by (17).

Obviously, natural frames affect statistically generated segments, while remaining unchanged due to the constraints. Statistically generated segments are connected smoothly to their neighboring natural segments. However, the hybrid speech feature-vector norm increases with the iterations of the algorithm. The proposed scheme combines the hybrid speech feature-vector representation with the iterative solution for the norm constrained speech feature-vector. As a result, the overall

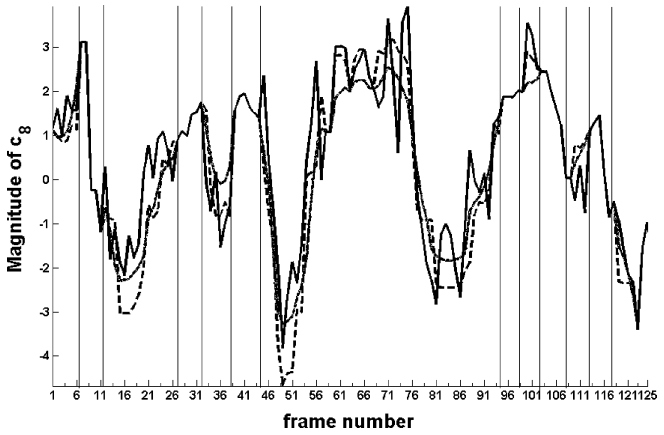


Fig. 5. Variation in time of c_8 for a hybrid-generated utterance, using the segment-wise representation statistical model (dashed line) in comparison to a hybrid-generated utterance using the conventional statistical model (solid gray line). The corresponding natural utterance (from CTTS) is shown in a solid black line. The vertical lines mark the constrained natural segments, where all the lines coincide, such as in frames: 7–12, 28–32, 39–44, 94–100, 102–108, and 113–118.

quality of the proposed hybrid generated speech, using a statistical model with improved dynamics (SW-STTS), is better than the quality of hybrid generated speech that uses a conventional STTS, as can be expected from the example shown in Fig. 5, and is confirmed by the listening tests results reported below.

IV. Experimental Results and Discussion

A. Results

1) HTTS Experimental Setup: We examined different compositions of the baseline CTTS and the baseline STTS in the proposed HTTS system. These baseline systems are built from a single female speaker voice. CTTS natural segments are aligned by three-state HMMs, as described in [2] and [1]. The baseline IBM parametric CTTS generates speech by a vocoder, described in [23]. We used frames of length of 20 ms in duration (440 samples, at a sampling rate of 22 kHz) with a frame overlap of 10 ms.

We simulated a CTTS system with different voice footprints by using different numbers of speech feature segments (candidates) in the nodes of the dynamic search trellis. We have experimented with voice footprints having memory size of 5, 7, 8.3, 12, and 22 MB. The simulated system footprints were controlled by the number of stage-to-stage candidate transitions in the dynamic search. The number of stage-to stage transitions define a number of possible speech feature segments in each node of the dynamic search. In all the experiments we used a fixed permitted spectral distance error.[2]

All the systems had the same number of context based acoustic leaves, which was set to 25 000. The memory size needed to store the statistical models is estimated to be about 1.3 MB. Thus, the sizes of the corresponding HTTS systems that were examined were: 6.3, 8.3, 13.3, and 23.3 MB.

The number of statistical segments within a hybrid utterance is different for CTTS systems having different footprint sizes. The smaller the footprint size, the more discontinuities typically appear in generated utterances, and hence more statistical segments are expected.

[2]This parameter is manually fine-tuned for a given CTTS system according to listening tests.

TABLE I
TTS Systems Compared by MOS Tests

system	size	type	MOS
A	22 MB	CTTS	3.7
B	8.3 MB	HTTS (SW-STTS)	3.5
C	8.3 MB	CTTS	3.17
D	1.3 MB	SW-STTS	3.05
E	8.3 MB	HTTS (Conv. STTS)	3.0
F	1.3 MB	Conv. STTS	1.9

Examining different compositions of hybrid utterances, we found that almost all natural segments are replaced by statistical models in a HTTS based on the 5 MB baseline CTTS. On the other hand, almost all natural segments remain in a hybrid utterances for a HTTS based on the 22 MB CTTS. We conclude that the HTTS systems are more useful when they are based on intermediate size CTTS systems.

The HTTS system, having an intermediate size (among the examined HTTS systems) of around 7 to 8 MB, interweaves a marked amount of both segment types (natural and statistical), where the ratio between natural segments to statistical models varied from 30% to 70%, for different sentences and permitted spectral errors in the range $0.2 \leq \epsilon \leq 0.7$.

2) Subjective Evaluation—Mean Opinion Score (MOS) Test: We have performed listening tests to evaluate the quality of speech generated by the proposed HTTS method.

In these tests we have evaluated the MOS, according to [28], for a set of ten English sentences. Each sentence was synthesized by the six different TTS systems listed in Table I, generated from the same U.S. English female speaker. Where A is the baseline IBM CTTS system, having a footprint of 22 MB, described in Section II-A; B is the proposed hybrid system, composed of a IBM CTTS system having a footprint of 7 MB, and system D; C is the baseline IBM CTTS system, having a footprint of 8.3 MB; D is the statistical segment-wise system, described in Section II-C; E is the proposed hybrid system, composed of a IBM CTTS system, having a footprint of 7 MB, and system F; F is the baseline (conventional) STTS system, described in Section II-B3. The tests were taken by 20 graduate and undergraduate students, having no experience with TTS system; none of them is a native English speaker, but all are fluent in English. All the tests were performed with a headphone set. The only information about the samples that the listeners were provided with was that the test aims to compare different speech synthesis methods.

As an addition to Table I, Fig. 6 shows the MOS test results with error bars indicating the 95% confidence intervals. We see that the proposed method, system B, outperforms the concatenative system C, having the same footprint of 8.3 MB. However, it is inferior to the concatenative system A, having a footprint of 22 MB.

B. Discussion

In this subsection, we discuss the advantages of the proposed HTTS over both the baseline STTS and the baseline CTTS. In Section IV-B1, we discuss the improvement of STTS by the proposed HTTS. In Section IV-B2, we discuss the improvement

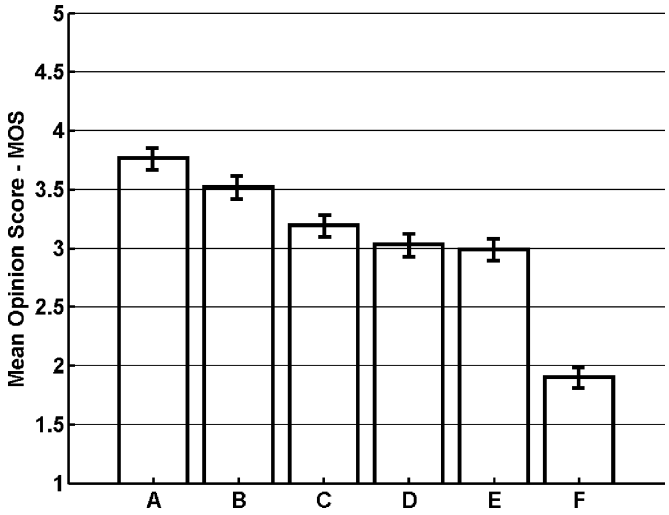


Fig. 6. Mean opinion score (MOS) test result, comparing the six TTS systems described in Table I. The error bars indicate 95% confidence intervals, computed using the "t-test,"

of CTTS by the proposed HTTS. Also, in that subsection we consider the effect of the baseline CTTS footprint size on the performance of the proposed HTTS.

1) Improvement of STTS by HTTS: As described previously, the main disadvantage of STTS synthesis is its unnatural quality. While on the other hand, its main advantages are smooth transitions in speech features between adjacent phonemes within a generated utterance, and a small footprint size.

The boundary constrained statistical speech synthesis (Section III-B) enables a smooth connection between statistically generated speech feature-vectors and natural speech feature-vectors. The positions of the natural speech feature vectors are determined by the hybrid dynamic path algorithm, presented in Section III-A.

The overall naturalness of statistically generated speech is improved by the introduction of natural speech segments, and is limited only by the allowed footprint size.

In Table I and Fig. 6, we see that introducing natural speech feature-segments into conventional statistically generated speech feature-vector, system E, results in a MOS increase of about one MOS unit, compared to the pure conventional statistical synthesis, system F.

The quality of segment-wise statistically generated speech increases by the hybrid scheme as well, system B, (increase of about half a MOS unit, compared to system D). However, the increase is lower than the increase for conventional statistically generated speech. The reason is that the initial segment-wise quality is higher than the quality of conventional statistically generated speech.

We conclude that both variants of STTS systems considered were improved by properly substituting statistical segments with segments from a CTTS system.

2) Improvement of CTTS by HTTS: One of the main drawbacks of CTTS synthesis is the possibility of encountering abrupt transitions between adjacent speech feature-segments. These abrupt transitions often cause unpleasant audible artifacts in the generated speech. The lower the CTTS system footprint, the more discontinuities appear in the generated speech. Yet, in spite of these artifacts, it possesses natural features and it sounds less muffled, as compared to statistically generated speech.

In this paper, we found that the proposed HTTS method can reduce discontinuities in speech features (generated by CTTS systems having a low footprint), by applying boundary constraints between statistical models and natural segments in the hybrid dynamic path, thus improving the overall generated speech quality.

In Fig. 6, we see that the proposed HTTS system, using the segment-wise statistical speech model, system B, outperforms the CTTS system C, with the same footprint of 8.3 MB.

The smaller the footprint of the baseline CTTS system, the larger is the improvement in the quality of speech generated by the HTTS system, in comparison to it. The bigger the footprint of the baseline CTTS is, the rarer are discontinuities in the generated speech, and as a result, the improvement obtained by the HTTS system decreases.

V. Conclusion

In this paper, we designed a hybrid TTS (HTTS) system by combining STTS with CTTS. The designed HTTS combines the advantageous characteristics of STTS, (optimal, smooth transitions between adjacent segments) with those of CTTS (naturalness of natural segments). The HTTS interweaves natural segments with statistical models, where the positions of statistical models are defined by the proposed hybrid dynamic path algorithm of Section III-A. In order to optimally connect natural segments to statistical models, boundary constrained statistical models are applied (Section III-B). A hybrid speech feature-vector over an entire utterance is generated iteratively. Natural segments are unchanged during the iterations, and constrain statistical segments, while statistical segments are updated according to the hybrid cost function gradient, as described in Section III-C. As a result, according to the performed MOS listening tests (Section IV-A2), hybrid generated speech sounds more natural than the corresponding pure statistically generated speech.

Concerning the comparison of the proposed HTTS system to CTTS systems, we conclude that the footprint of the compared CTTS system should be considered as well. CTTS systems having larger footprint (20 MB and more) are hardly improved by the combination with STTS because such systems generate speech with a very small number of audible discontinuities. On the other hand, CTTS systems having a small footprint (5 MB and less), generate speech in which almost all segments are not connected smoothly, resulting in a noticeable degradation in the generated speech quality. In this case, HTTS is very close to STTS.

In this paper, we show that the proposed HTTS is advantageous at an intermediate working point of about 7 MB (for the baseline CTTS footprint). However, determining the working point is currently not based on some optimality criterion, but rather on subjective evaluations, and this issue is considered as one of the possible continuations of this research, which are further discussed in the next subsection.

It is important to note that the improvement of the CTTS system by STTS system was successful is this work because

both systems use the same speech models and operate in the same speech features space.

A. Future Work

The proposed dynamic path algorithm is based on a cost function derived from the spectral distance between consecutive natural segments. A more sophisticated approach for interweaving statistical models with natural segment should rely on a metric reflecting a tradeoff between discontinuities in natural segments to the unnaturalness of statistically generated segments. To derive such a metric, further research is needed. In particular, the degradation in synthesized speech quality, caused by the spectral discontinuities between consecutive natural segments, should be compared to the degradation in synthesized speech quality caused by the unnaturalness of statistically generated segments. This hybrid metric should then be used in the hybrid dynamic path algorithm.

In this paper, we use the same model (a single Gaussian component per HMM state for a given phoneme) for every phoneme. Different models for different broad phonetic classes can improve the overall quality of statistically generated speech. Probably, certain phonetic classes should be excluded from statistical modeling at all, e.g., fricative and plosive phonemes that are seldom modeled properly, causing degradation in generated speech quality.

ACKNOWLEDGMENT

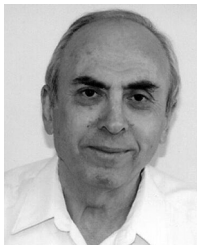
This research is part of a joint research project conducted at the Signal and Image Processing Lab (SIPL), Technion—Israel Institute of Technology, and IBM's Haifa Research Lab (HRL). The authors would like to thank R. Hoory, the head of the Speech Technologies Group in HRL, for his support, and A. Sagi (formerly at HRL) and A. Sorin (HRL) for useful discussions in the course of the work. S. Tiomkin and D. Malah would like to thank HRL-IBM for permission to use their speech databases and TTS software.

REFERENCES

[1] R. E. Donovan and E. M. Eide, "The IBM trainable speech synthesis system," in *Proc. ICSLP'98*, Sydney, Australia, vol. 5, pp. 1703–1706.
[2] R. E. Donovan, "Topics in decision tree based speech synthesis," *Comput. Speech Lang.*, vol. 17, no. 1, pp. 43–67, Jan. 2003.
[3] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. Acoust. Speech Process.*, Munchen, Germany, 1996, vol. 1, pp. 373–376.
[4] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Proc. Eurospeech'97*, pp. 601–604.
[5] R. Fernandez, Z. Kons, S. Shechtman, Z. W. Shuang, R. Hoory, B. Ramabhadran, and Y. Qin, "The IBM submission to the 2008 text-to-speech blizzard challenge," in *Proc. Blizzard Challenge'08*, Brisbane, Australia, Sep. 2008.
[6] E. Eide, R. Fernandez, R. Hoory, W. Hamza, Z. Kons, M. Picheny, A. Sagi, S. Shechtman, and Z. W. Shuang, "The IBM submission to the 2006 Blizzard text-to-speech challenge," in *Proc. Blizzard Challange'06*, Pittsburgh, Sep. 2006.
[7] R. E. Donovan, "Text-to-Speech using Clustered Context-Dependent Phoneme-based Units," U.S. Patent 6,163,769, Dec. 19, 2000.
[8] K. Tokuda, T. Yoshimura, T. Masuko, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP'00*, Istanbul, Turkey, pp. 1315–1318.
[9] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.'09*, pp. 1039–1064.
[10] V. Pollet and A. Breen, "Synthesis by generation and concatenation of multiform segments," in *Proc. Interspeech'08*, pp. 1825–1828.
[11] Z. Ling and R. Wang, "HMM-based unit selestion using frame sized speech segments," in *Proc. Interspeech'06*, Pittsburgh, PA, Sep. 17–21, 2006, pp. 2034–2037.
[12] Z. Ling and R. Wang, "HMM-based hierarchical unit selection combining Kullback–Leibler divergence with likelihood criterion," in *Proc. ICASSP'07*.
[13] Z. Ling and R. Wang, "Minimum unit selection error training for HMM-based unit selection speech synthesis system," in *Proc. ICASSP'08*, Las Vegas, NV, pp. 3949–3952.
[14] Z. Ling, L. Qin, H. Lu, Y. Gao, L. Dai, R. Wang, Y. Jiang, Z. Zhao, J. Yang, J. Chen, and G. Ping, "The USTC and iflytek speech synthesis systems for Blizzard Challenge 2007," in *Proc. BLZ3'07*, 017.
[15] Z. Ling, H. Lu, G. Hu, L. Dai, and R. Wang, *The USTC System for Blizzard Challenge 2008*. Hefei, China: Univ. of Sci. and Technol. of China, 2008.
[16] H. Lu, Z. Ling, M. Lei, C. Wang, H. Zhao, L. Chen, Y. Hu, L. Dai, and R. Wang, *The USTC System for Blizzard Challenge'09*. Hefei, China: iFlytek Speech Lab, Univ. of Sci. and Technol. of China, 2009.
[17] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "Ximera: A new TTS from ATR based on corpus-based technologies," in *Proc. SSW5'04*, 2004.
[18] X. Gonzalvo, A. Gutkin, J. Socoro, I. Iriondo, and P. Taylor, "Local minimum generation error criterion for hybrid HMM speech synthesis," in *Proc. Interspeech'09*, Brighton, U.K., pp. 416–419.
[19] M. Plumpe, A. Acero, H. W. Hon, and X. Huang, "HMM-based smoothing for concatenative speech synthesis," in *Proc ICSLP*, Sydney, Australia, 1998.
[20] T. Okubo, R. Mochizuki, and T. Kobayashi, "Hybrid voice conversion of unit selection and generation using prosody dependent HMM," *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 11, pp. 2775–2782, 2006.
[21] M. Ayelett and J. Yamagishi, "Combining statistical parametric speech synthesis and unit-selection for automatic voice cloning," in *Proc. LangTech'08*, 2008.
[22] S. Tiomkin, D. Malah, and S. Shechtman, "Statistical text-to-speech synthesis based on segment-wise representation with a norm constraint," *IEEE Trans, Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 1077–1082, Jul. 2010.
[23] D. Chazan, R. Hoory, Z. Kons, A. Sagi, S. Shechtman, and A. Sorin, "Small footprint concatenative text-to-speech synthesis using complex envelop modeling," in *Proc. Interspeech'05*, Lisbon, Portugal, pp. 2569–2572.
[24] 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," in *Proc. ICASSP'06*, Toulouse, France, pp. 877–890.
[25] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[26] S. Furui, "Speaker independent isolated word recognition based on dynamics emphesized cepstrum," *Trans. IECE Japan*, vol. 69, pp. 1310–1317, Dec. 1986.
[27] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," in *Proc. ICASSP'86*, Tokyo, Japan, pp. 877–880.
[28] "Mean Opinion Score (MOS)," Telecomm. Standardization Sector, Int. Telecomm. Union (ITU-T), Geneva, Switzerland, Rec. P.800.

Stas Tiomkin received the B.Sc. degree in computer engineering and physics from the Hebrew University of Jerusalem, Jerusalem, Israel, in 2004 and the M.Sc. degree in electrical engineering from the Technion—Israel Institute of Technology, Haifa, in 2009, with his graduate work in improving the quality of text-to-speech synthesis.

During the M.Sc. degree, he was a Teaching Assistant in the Electrical Engineering Department, the Technion, and as a Projects Supervisor in the Signal and Image Processing Lab (SIPL) at the Technion. His professional interests are in speech synthesis, speech recognition, speech enhancement, speech encoding, signal processing, and machine learning. Currently, he is a Speech Processing Engineer, developing speech recognition systems and text-to-speech synthesis systems for mobile and embedded platforms.



David Malah (S'67–M'71–SM'84–F'87–LF'09) received the B.Sc. and M.Sc. degrees from the Technion—Israel Institute of Technology, Haifa, Israel, in 1964 and 1967, respectively, and the Ph.D. degree from the University of Minnesota, Minneapolis, in 1971, all in electrical engineering.

Following two years on the staff of the Electrical Engineering Department, University of New Brunswick, Fredericton, NB, Canada, he joined the Technion in 1972 , where he is an Elron-Elbit Professor of Electrical Engineering. During the period 1979 to 2001, he spent about 6 years, cumulatively, of sabbaticals and summer leaves at AT&T Bell Laboratories, Murray Hill, NJ, and AT&T Labs, Florham Park, NJ, conducting research in the areas of speech and image communication and the summer of 2004 at the Georgia Centers for Advanced Telecommunications Technology (GCATT), working in the area of video processing. Since 1975, he has been the academic head of the Signal and Image Processing Laboratory (SIPL), the Technion, which is active in image/video and speech/audio processing research and education. From 2006 to 2010, he served as the Director of the Center for Communication and Information Technologies (CCIT), Electrical Engineering Department, the Technion. His main research interests are in image, video, speech and audio coding; speech and image enhancement; text-to-speech synthesis; hyperspectral image analysis; data embedding in signals, and digital signal processing techniques.

Prof. Malah is a recipient of the 2007 International IBM Faculty Award. He is on the Editorial Board of the *Journal of Visual Communication and Image Representation*, since 1999, and as of 2010, he is on the Senior Editorial Board of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING.

Slava Shechtman received the B.Sc. and M.Sc. degrees (*cum laude*) in electric engineering from the Technion—Israel Institute of Technology, Haifa, Israel, in 1999 and 2004, respectively. His M.Sc. thesis explored speech modeling and very low bit rate speech coding.

He joined the Speech Technology Group, IBM Haifa Research Labs, in 2004. His major areas of interest are speech modeling, synthesis, coding and transformation. Since 2010 he has been leading Speech Modeling for the TTS research project at IBM.

Zvi Kons received the B.A. and M.Sc. degrees in physics from the Technion—Israel Institute of Technology, Haifa, Israel, in 1992 and 1999, respectively.

He is a Researcher in the Speech Technologies Group, IBM Haifa Research Lab. His main interest is currently speech synthesis and speech processing. Previous experience includes research and development in areas of signal processing, image processing, and computer vision. He joined IBM in 2001 and currently he is leading the IBM embedded text-to-speech project.