A SEGMENT-WISE HYBRID APPROACH FOR IMPROVED QUALITY TEXT-TO-SPEECH SYNTHESIS

RESEARCH THESIS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

STAS TIOMKIN

SUBMITTED TO THE SENATE OF THE TECHNION -
ISRAEL INSTITUTE OF TECHNOLOGY

IYAR, 5769 HAIFA MAY 2009

Acknowledgements

This research thesis was done under the supervision of professor David Malah in the department of Electrical Engineering.

I would like to express my deep gratitude to my supervisor, Prof. David Malah, for his dedicated personal guidance and inspiration. I have learned a variety of research skills from him including conducting an academic research, scientific writing, determining research goals, and achieving them.

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). I would like to thank Ron Hoory, the head of the Speech Processing group at HRL, Slava Shechtman, Zvi Kons, Ariel Sagi, and Alex Sorin from HRL, for useful discussions in the course of the work.

The help of the Signal and Image Processing Laboratory (SIPL) staff, Nimrod Peleg, Ziva Avni and Avi Rosen, is gratefully acknowledged in administrating the listening tests and in providing technical support. It was a great experience to work in SIPL with such colleagues.

I owe many thanks to my mother Sofia and brother Anton for their support and patience.

The generous financial help of Technion and IBM-HRL is gratefully acknowledged.

Dedicated to Evy with Love.

Contents

List of Figures

Abstract

Concatenative Text-To-Speech (TTS) synthesis and statistical TTS synthesis are the two main approaches to text-to-speech synthesis. A concatenative TTS (CTTS) directly uses parameters of natural speech features, selected from a recorded speech database. Consequently, CTTS systems enable speech synthesis with natural quality. However, since the desired segments, having required characteristics, are not always available, other segments with the closest characteristics to the required ones are used instead. Concatenation of such segments may result therefore in audible discontinuities. Consequently, the smaller the footprint size of the TTS system is, the lower is the quality of generated speech that is achieved. On the other hand, a statistical TTS systems (STTS), while having a smaller footprint than CTTS, generate speech that is free of such discontinuities but, in general, is of lower quality than CTTS in terms of naturalness, as often it sounds muffled and buzzy. This is due to over-smoothing of model-generated speech features.

In this research we developed two approaches aimed to improve the quality of TTS generated speech. First, we develop two techniques for improving the quality of the baseline STTS system. Second, we propose a technique for combining CTTS with STTS for a new class of TTS systems,

denoted hybrid TTS (HTTS).

In STTS, speech feature dynamics is modeled 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.

To enhance a baseline STTS system we propose two methods. First, we propose to represent speech features dynamics in the transform domain and not directly in terms of frame to frame variation. In the transform domain, the insufficient dynamics is characterized explicitly by a marked attenuation in inter-harmonic components. We found that the quality of speech generated by a STTS system is improved by enhancing these attenuated components.

Second, we introduce 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. The segments-wise representation method is superior to the transform domain method in terms of generated speech quality. However, the first method is less computational complex, as compared to the second one.

In addition we propose in this work to combine the advantages of CTTS and STTS into a new type of TTS, denoted HTTS. This is an hybrid system in which, for each utterance, natural segments and model-generated segments are interweaved via a hybrid dynamic path algorithm. As a results, speech generated by the proposed HTTS includes less discontinuities than the baseline CTTS system does, and it sounds more natural than the baseline STTS.

We designed a TTS system where both developed techniques, HTTS and improved STTS are applied, and subjectively tested.

List of Symbols

α_n	Step size.
$\Delta^1\mathbf{c}_i^T$	Speech feature vector derivative approximation of 1^{st} order.
$\Delta^2\mathbf{c}_i^T$	Speech feature vector derivative approximation of 2^{st} order.
η	Scalar Lagrange multiplier.
γ	Vectorial Lagrange multiplier.
γ_0	Enchantment factor.
λ_{n+1}	Balancing factor.
$\mathbf{c}^{opt,sw}$	Optimal segment-wise speech feature vector.
$\mathbf{c}^{opt}$	Optimal solution over an entire utterance..
$\mathbf{c}_{M,T}$	Speech feature vector over an entire utterance.
$\mathbf{c}_M$	Speech feature frame.
$\mathbf{o}_{3M,1}$	Augmented space speech feature vector.
$\mathbf{o}_{3M,T}$	Augmented space speech feature vector over an entire utterance.

$\widetilde{\mathbf{m}}_{3MK\times 1}$	Non-replicated model mean vector.
$\mathbf{c}_{n+1}$	Speech feature vector after n iterations.
$\mathbf{m}_{q_t}^{T}$	Statistical model mean vector.
$\mathbf{o}^{opt}$	Optimal augmented speech feature vector.
$\mathbf{U}_{q_t}^{-1}$	Statistical model covariance mean vector.
$\mathbf{W}_{3M\cdot N\times M\cdot N}$	Linear Transformation from speech feature vectors to augmented space speech feature vectors.
$\tilde{s}_{\omega}(n)$	Sinusoidal speech model.
$\widehat{\Theta}$	The optimal sequence of speech segments.
$\widetilde{\mathbf{U}}_{3MK\times 3MK}^{-1}$	Non-replicated model covariance matrix.
A	Constraints defining matrix.
B_n	Triangular basis functions.
c_n	Speech spectrum amplitude expansion coefficient.
$d(\Theta, T)$	Dynamic programming cost function.
$d_u(\theta_j, t_j)$	Prosody component of dynamic programming cost function.
f'	Mel-Scale Frequency.
$J(\mathbf{Wc})$	The cost function over an entire utterance.

$J^{sw}(\widetilde{\mathbf{W}}\mathbf{c})$	Cost function over an entire utterance.
$log(\mathbf{A}(f')$	Speech spectrum amplitude.
M	Speech feature frame dimension.
$\nabla(\mathbf{c}_n)$	Gradient of cost function with regulatory constraint.
$J_c^{sw}(\widetilde{\mathbf{W}}\mathbf{c})$	Cost function with regulatory constraint.

Chapter 1

Introduction

Rapid development in computer systems and devices requires development of human-machines interfaces as well. There are many devices that would be inconvenient to use, if their input-output were restricted by the standard I/O devices, such as a keyboard, a mouse, or a monitor. Today's machines can interact with people in very different ways. Beyond the conventional I/O devices, additional interfaces like direct human-machine spoken dialog systems are being developed. The input stage of human-machine spoken dialog system should include speech and language recognition modules. While the output device should be able to translate machine commands and respond by speech, typically generated by a text-to-speech (TTS) system.

There is great interest in improving the quality of Text-To-Speech (TTS) systems as the number of applications using TTS increases. 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 [16], [24], [26], [30], [27]. Efforts invested in handling the over-smoothing problem are reported in [16], [24].

In this research 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)* An alternative 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 proposed segment-wise method is denoted as 'SW-STTS'.

While investigating the speech feature over-smoothing technique, we developed an additional method (to 'SW-STTS' method) for improving speech feature dynamics. By this technique speech feature dynamics are enhanced in the transform domain. Resulting generated speech quality is better than the quality of speech generated by the conventional model. This technique is inferior in its quality to the quality of SW-STTS, however, its less computational complex. In addition, examining speech feature dynamics provides additional insights to the speech feature over-smoothing. The method for improving speech feature dynamics in the transform domain was published and is detailed at [28].

Also, in this research 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 as HTTS. An HTTS system interweaves natural segments with statistical model-generated segments via a proposed hybrid dynamic path algorithm.

The proposed hybrid dynamic path algorithm aims to introduce statistical models to certain positions within an utterance at which CTTS suffers from discontinuities. As a result, natural segment sequences are bridged by boundary constrained statistically generated segments. This concatenation scheme is realized by the proposed hybrid dynamic path algorithm described in this work.

The main components constituting the proposed HTTS system are the mentioned hybrid dynamic path algorithm, enabling allocation of natural segments along with statistical model-generated segments within an entire utterance, and, a proposed hybrid speech features generating algorithm.

The proposed HTTS system inherits the naturalness of CTTS systems and the smooth transitions of STTS, while, on one hand, having a lower footprint than CTTS, and, on the other hand, requiring less computational resources than pure STTS systems. 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 controls the ratio of the numbers of natural segments to the numbers of statistically generated segments comprising a synthesized utterance. Speech generated in an intermediate (hybrid) mode consists of natural and statistically generated segments, interweaved within an utterance.

Obviously, the quality of a HTTS system depends on the qualities of the baseline CTTS and STTS systems. Particulary, different STTS systems result in HTTS systems having different qualities. Consequently, an enhanced STTS system will results in a better HTTS. To demonstrate this aspect we compared the quality of HTTS composed of a conventional STTS and a baseline CTTS to the quality of HTTS composed of a SW-STTS with improved dynamics and the baseline CTTS. Speech features, generated by SW-STTS, are less smoothed, compared to those generated by a conventional STTS, and as a result, the generated speech sounds less muffled and buzzy. As confirmed by listening tests, combining SW-STTS with CTTS results in a HTTS system with better naturalness, as compared to HTTS that is composed of conventional STTS and CTTS.

This thesis is organized as follows. In Chapter 2 we provide the essentials of the baseline CTTS, in Chapter 3 we provide the essentials of the baseline STTS used in this research. In the rest of the thesis we represent the methods which were developed in this research. In Chapter 4 we present a technique for improving speech feature dynamics in the transform domain. In Chapter 5 we present the improved quality statistical text-to-speech synthesis based on segment-wise representation with a norm constraint. In Chapter 6 we present the proposed hybrid text-to-speech system. Finally, in Chapter 7 we provide a summary and directions for continuation of this research.

Chapter 2

Concatenative Text-to-Speech Synthesis

In concatenative systems, speech is synthesized by concatenating natural speech segment features, denoted in literature as candidates, units, or segments. These segments are basic segments for speech synthesis. In different systems these segments represent phonemes or even sub-phonemes. The current research is based on the IBM concatenative text to speech system that uses a sub-phoneme as a basic segment/units, as described in Section 2.1, and detailed at [6], [4], [7], and [5]. The main functional blocks of CTTS are: 1) Acoustic context-dependent decision trees for each sub-phoneme, holding in their leaves parametric representation for natural speech segment features, represented by 'Unit Repository' block in Fig.2, 2) Target-prediction trees, holding context dependent target energy, pitch and duration, and, a phonetic text analyzer, providing graphemes to phonemes transcription and dictating target values for pitch, energy and duration, represented by

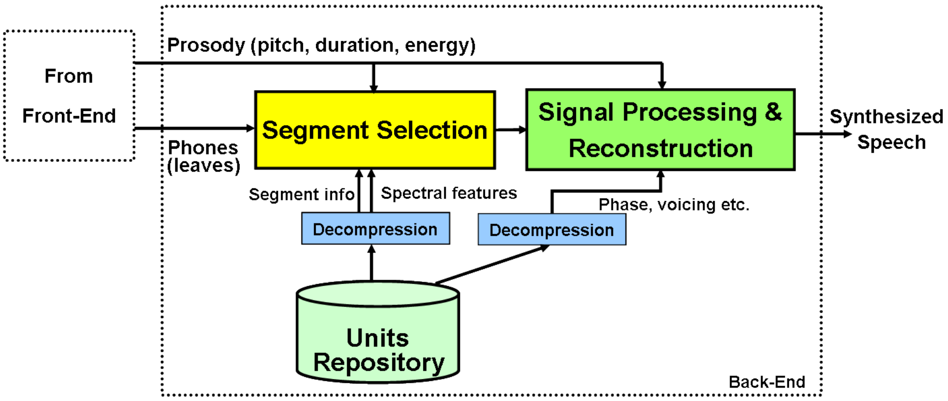


Figure 2.1: The main blocks of the reference CTTS system used in our research.

’Front-End’ block in Fig.2, 3)A dynamic search algorithm, finding an optimal sequence of natural speech segment features according to target values, and concatenative distance between segments, represented by ’Segment Selection’ block in Fig.2, 4) A speech generator that composes speech from the speech feature sequence found by the dynamic search, represented by the ’Signal Processing and Reconstruction’ block in Fig.2. Typically, natural speech segment features are stored in a compressed form, and are decompressed during synthesis. This is done in ’Decompression’ block in Fig.2.

2.1 Acoustic context decision tree

Each sub-phoneme context-dependent tree is built off-line in the following fashion: All occurrences of a particular sub-phoneme are assigned to a root node of a tree. Afterwards, this root node is split into two other nodes, which may be split further and so forth. There are many possible options for splitting a certain node. These options are dictated by a sub-phoneme context. A sub phoneme context is defined by a set of questions concerning neighbors of that sub-phoneme. Each question splits all occurrences of a node into two groups. A question, which divides a node into two most homogenous groups, defines a correct split and a node is divided according to that question. A question is asked concerning immediate neighbors of a phoneme. Asking questions about more distant contexts may give slightly more accurate acoustic models. However, it results in appearing leaves which don't have segments concatenating smoothly with neighboring segments. A node's split occurs when the total homogeneity of its sibling nodes is lower than its own homogeneity. This process stops when a tree achieves a predefined number of leaves or when there are not any nodes which may be split into two nodes with a lower total homogeneity than their parents. Where a metric for homogeneity is related to variance of a node.

.

2.2 Target prediction trees

2.2.1 Energy prediction trees

An energy prediction decision tree is built for each sub-phoneme in a fashion like acoustic trees are built, but the questions asked for splitting nodes concerns just two neighboring phonemes. A stored median value of each leaf is used for the prediction during synthesis.

2.2.2 Duration prediction trees

A duration prediction decision tree is built for each sub-phoneme asking questions concerning two nearest neighbors like an energy tree is built. A mean value for each leaf is used for duration prediction at the synthesis stage.

2.2.3 Pitch prediction trees

For each segment in the database a pitch delta is computed at the beginning and at the end of the segment divided by the pitch at the beginning of the segment. The geometric mean of these deltas is computed for each sub phonemes and recorded for later use in pitch contour generation.

During synthesis the words to be synthesized are converted to a phone sequence by dictionary lookup, with the selection between alternatives for words with multiple pronunciations being performed manually. The decision trees are used to convert a phone sequence into an acoustic, duration, and energy leaf for each sub phoneme in the sequence; and pitch contour is defined as well. The median training values in the duration and energy leaves are used as the predicted duration and energy values for each sub-phoneme.

2.3 Dynamic search

Each acoustic leave, which represents a particular phoneme in a phoneme sequence comprising a generated text, holds a number of candidates (segments), to be concatenated to an entire speech waveform. Clearly, the more candidates are in any given acoustic leave the higher quality of generated speech. However, there is an exponential number of possible combinations of candidates to combine an utterance. Dynamic programming with an appropriate cost function is applied to find an optimal combination of candidates.

The IBM cost function includes two basic components. The first one is the similarity of segments inherent prosody to the front-end dictated target prosody. The second is a concatenation cost of each segment with its adjacent segments in the acoustic leaves sequence.

The overall cost for a particular candidate is composed from the target cost plus the concatenation cost. A dynamic search, to obtain the optimal path, is performed using this overall cost:

$$\begin{aligned} \widehat{\Theta} &= \underset{\Theta}{argmin}\, d(\Theta, T) \\ d(\Theta, T) &= \sum_{j=1}^{N} d_u(\theta_j, t_j) + \sum_{j=1}^{N} d_c(\theta_j, \theta_{j+1}), \\ d_u(\theta_j, t_j) &= \sum_{i=1}^{I} \omega_i d_u^i(\theta_j, t_j), \end{aligned} \qquad (2.1)$$

where, $\widehat{\Theta}$ is the optimal sequence of segments, θ_j is a segment, being considered at the *j-th* stage, t_j is a required prosody for the *j-th* phoneme, $d_c(\theta_j, \theta_{j+1})$ is a spectral distance between θ_j and θ_{j+1}, $d_u^i(\theta_j, t_j)$ is a cost for the *i-th* prosody component of θ_i, and, ω_i is a weight of the *i-th* prosody component.

All possible segment concatenations are examined during a forward pass of the dynamic search. The best path is found on this full connected trellis by back-tracing. The best path segments are sent to the speech generator, using overlap and add synthesis. In Section 6.2 we propose a modified dynamic search, enabling an optimal interweaving of natural segments with statistically generated segments.

One of our research objectives is to find a way to integrate a statistical model into the dynamic search, and to define a hybrid synthesis system that uses data-base segments along with segments derived from statistical models. The next section describes the statistical model.

2.4 Speech parameters

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:

$$log(\mathbf{A}(f')) = \sum_{n=1}^{M} c_n \cdot \mathbf{B}_n(f'), \tag{2.2}$$

where f' denotes a mel-scale frequency[1].

This representation is successfully used in IBM's state-of-the-art CTTS system, detailed in [3]. However, it was not previously used in HMM-based speech synthesis systems. Examination of the suitability of this represen-

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

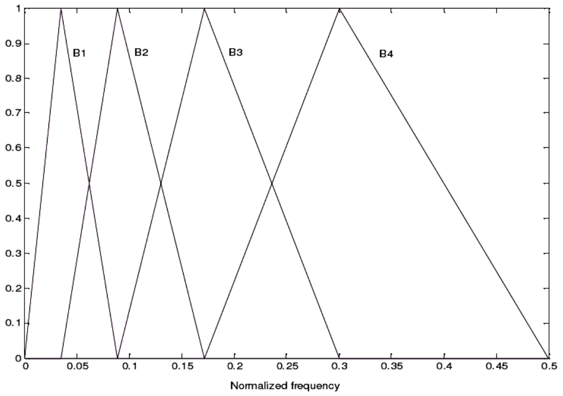


Figure 2.2: Triangular basis functions.

tation 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 [16] and [22], respectively.

2.5 Speech reconstruction

In this research speech waveform is represented by sinusoidal model, [2]:

$$s_{\omega}(n) \cong \tilde{s}_{\omega}(n) = \omega(n) \sum_{k=0}^{k=L} A_k sin(\Theta_k n + \varphi_k), \tag{2.3}$$

where A_k, Θ_k, φ_k is the k-th harmonic amplitude, frequency and phase, respectively. $\omega(n)$ is a window function. However, not all the frequencies are stored, but only the pitch frequency, Θ_0. At the reconstruction stage all the frequencies are generated as multiples of Θ_0, $\tilde{\Theta}_k = \Theta_0 k$.

The phase information is stored for the low band frequencies only, while the high bad phase information is generated from the low bands phase by a non-linear operation (such as wive rectification) in the time domain.

To improve the reconstructed speech naturalness a random frequency dither is applied to $\tilde{\Theta}_k$. The dither is applied above a predefined threshold frequency and gradually increases towards high frequencies.

A short time spectrum is reconstructed in the frequency domain according to (2.3). Speech waveform is generated by OLA (overlap-add) method from a short time wave forms.

Chapter 3

Statistical Text-to-Speech Synthesis

In this chapter we briefly describe the conventional approach for deriving the entire utterance speech feature vector in statistical HMM-based TTS.

3.1 Statistical speech features representation

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

$$\mathbf{c} \;=\; [\mathbf{c}_1^T, \mathbf{c}_2^T, \ldots, \mathbf{c}_N^T]^T, \tag{3.1}$$

where $\mathbf{c}_i = (c_i(1), c_i(2), \ldots, c_i(M))^T$ are the expansion coefficients, introduced in (2.2). $\mathbf{c}_i$ 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 trees, detailed in [6], [4], and [5].

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 [20]. 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 [9], [21], [16], [24]. 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:

$$\mathbf{o} = [\mathbf{o}_1^T, \mathbf{o}_2^T, \ldots, \mathbf{o}_N^T]^T, \tag{3.2}$$

where,

$$\mathbf{o_i} = (\mathbf{c}_i^T, \Delta^1\mathbf{c}_i^T, \Delta^2\mathbf{c}_i^T)^T. \tag{3.3}$$

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 [8]. When using only the two-sided first and second order differences, the dynamic features are computed as:

$$\Delta^{1,2}\mathbf{c}_i^T = \sum_{\tau=L_-^{(1,2)}}^{L_+^{(1,2)}} \omega^{1,2}(\tau)\mathbf{c}_{i+\tau}^T, \tag{3.4}$$

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:

$$\mathbf{o}_{3M\cdot N\times 1}{=}\mathbf{W}_{3M\cdot N\times M\cdot N}\mathbf{c}_{M\cdot N\times 1}, \tag{3.5}$$

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.

3.2 Statistical model

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

$$\mathbf{o}^{opt} = \underset{\mathbf{o}}{argmax}\;\; P(\mathbf{o} \mid \lambda) \tag{3.6}$$

and

$$P(\mathbf{o}|\lambda) = \sum_{\forall \mathbf{q}} P(\mathbf{o}, \mathbf{q}|\lambda), \tag{3.7}$$

where $\mathbf{q} = (q_1, q_2, \dots, q_N)$ 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 [20]. 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 [7].

As mentioned in Section 2.4, 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 equation (3.6) to the following problem, which assumes that the state sequence, $\mathbf{q}$, is given:

$$\mathbf{o}^{opt} = \underset{\mathbf{o}}{argmax} \;\; P(\mathbf{o} \mid \mathbf{q}, \lambda), \tag{3.8}$$

Methods for full HMM-based speech feature synthesis appear at [16], [24], [25].

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:

$$log(P(\mathbf{o} \mid \mathbf{q}, \lambda)) \;\; = \;\; \frac{1}{2}(\mathbf{o} - \mathbf{m})^T \mathbf{U}^{-1} (\mathbf{o} - \mathbf{m}), \tag{3.9}$$

with

$$\mathbf{m} \;\; = \;\; [\mathbf{m}_{q_1}^T, \mathbf{m}_{q_2}^T, \ldots, \mathbf{m}_{q_N}^T]^T \tag{3.10}$$

and

$$\mathbf{U}^{-1} \quad = \quad diag[\mathbf{U}_{q_1}^{-1}, \mathbf{U}_{q_2}^{-1}, \ldots, \mathbf{U}_{q_N}^{-1}], \tag{3.11}$$

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, $\mathbf{m_{q_t}}$ 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 5.1.

Clearly, given equation (3.9), expression (3.8) 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 (3.8), taking into consideration the relation between the static and dynamic features, defined by equation (3.5) (note that $\mathbf{W}$ is not invertable):

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

Consequently, the cost function over an entire utterance is:

$$\begin{aligned} J(\mathbf{Wc}) \quad &= \quad -lnP(\mathbf{Wc} \mid \mathbf{q}, \lambda) \\ &= \quad \frac{1}{2}(\mathbf{Wc} - \mathbf{m})^T \mathbf{U}^{-1} (\mathbf{Wc} - \mathbf{m}) \\ &= \quad \frac{1}{2}\|\mathbf{U}^{-\frac{1}{2}}(\mathbf{Wc} - \mathbf{m})\|_2^2. \end{aligned} \tag{3.12}$$

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:

$$\begin{aligned}\frac{\partial J(\mathbf{Wc})}{\partial \mathbf{c}} &= -\mathbf{W}^T\mathbf{U}^{-1}\mathbf{Wc} + \mathbf{W}^T\mathbf{U}^{-1}\mathbf{m} \\ &= \ 0 \end{aligned} \tag{3.13}$$

consequently,

$$\mathbf{W}^T\mathbf{U}^{-1}\mathbf{Wc} = \mathbf{W}^T\mathbf{U}^{-1}\mathbf{m}. \tag{3.14}$$

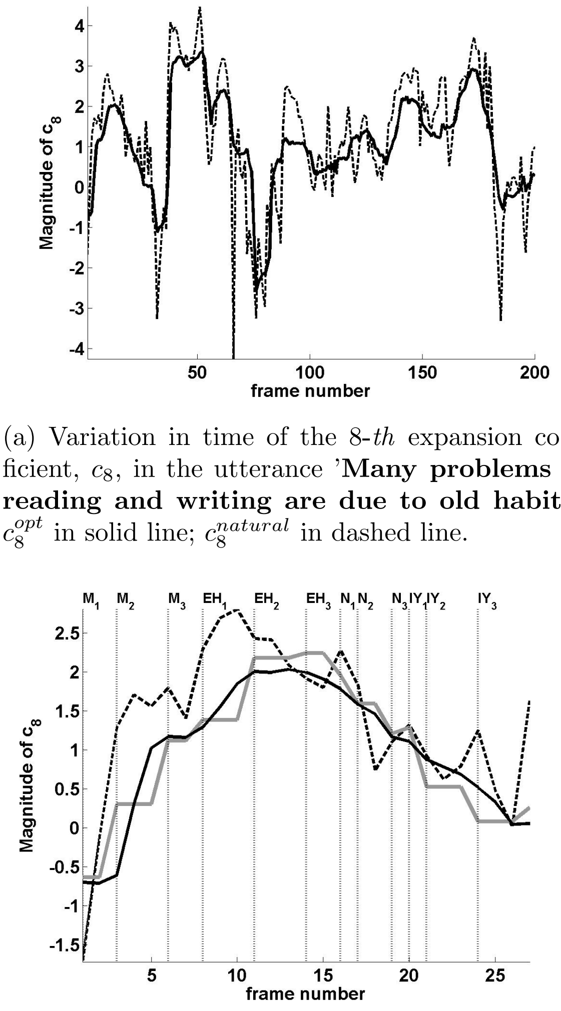
Assuming that the matrix $\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W}$ is invertible, the optimal solution $\mathbf{c}^{opt}$ is given by:

$$\mathbf{c}^{opt} = (\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{U}^{-1}\mathbf{m}. \tag{3.15}$$

To solve (3.15) directly requires $O(N^3M^3)$ computations. However, utilizing the special structure of $\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W}$, (3.15) can be solved by the Cholesky decomposition or the QR decomposition with $O(NM^3L^3)$, where L:

$$L = \max\{L^1_{+,-}, L^2_{+,-}\}.$$

We can see in Fig. 3.1(a) that, typically, the optimal solution (3.15) is over-smoothed and has much less dynamics (inter-frame variations), as compared to the corresponding natural speech features. The natural 8-*th*



(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.

Figure 3.1: Demonstrating conventional statistically generated speech feature over-smoothing in time, compared to a reference natural speech feature.

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 [16], [24], [26], [30], [27]. The buzziness is not caused by source/filter decomposition, because speech does not sound buzzy after just analyzing the speech and synthesizing it back from its features.

Fig. 3.1(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, $\boldsymbol{\Delta}^{1,2}\mathbf{c}_i$ do not appear to fully capture the features dynamics, as also indicated by listening. We conclude from Fig. 3.1(a) and Fig. 3.1(b) that generated speech features should approximate the model means but, at the same time, they should fluctuate about the model 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 addition, the speech feature dynamics may be enhanced by a different approach to dynamic speech feature modeling.

In the next chapter we propose to model the speech feature dynamics in the transform domain, rather than by the conventional approach based on the first and the second deltas, (3.4).

In Chapter 5 we introduce a new concept of segment-wise model representation, which is found to improve the naturalness of generated speech.

The second approach is superior to the first approach in terms of the generation speech quality, but it is more computational complex.

Chapter 4

Improving STTS Dynamics In The Transform Domain

We found that speech features in contiguous frames, as generated by a STTS system, do not vary much, while those in natural speech vary much more and thus are more dynamic, as shown in Fig. 4.4. We propose to represent speech features dynamics in the transform domain and not directly in terms of frame to frame variation. In the transform domain, the insufficient dynamics is characterized explicitly by a marked attenuation in inter-harmonic components. We found that the quality of speech generated by a STTS system is improved by enhancing these attenuated components, making the synthesized speech sound less buzzy and less muffled. We also propose to differently treat inter- and intra-phoneme (or sub-phoneme) frames, where the dynamics of intra-phoneme frames is improved by enhancing inter-harmonic amplitude components, while inter-phoneme transitions are smoothed by constraining phonemes boundary differences. This approach is published in the international speech conference INTERSPEECH 2008 [28].

This chapter is organized as follows: In Section 4.1 we demonstrate the proposed approach to modeling speech features dynamics in the transform domain. In Section 4.2 we show how to combine enhancement of intra-phoneme dynamics with inter-phoneme transition smoothing, deriving an optimal solution for the speech features of an utterance. In section 4.3 we provide experimental results, and, finally, we summarize the chapter in Section 4.4.

4.1 Modeling feature dynamics in the transform domain

Features representation in the transform domain

To analyze the inter-frame speech features dynamics we propose to consider a phoneme of T_i frames as a quasi-periodic sequence with a period of d samples, where a phoneme of T_i frames is represented as a one-dimensional coefficients sequence of length dT_i, as in Fig.4.1. In that figure we can see that the statistically generated one-dimensional sequence is almost periodic, with a repeating pattern every d samples, while the natural one-dimensional sequence varies much more from frame to frame. In Fig.4.1, the inter-frame dynamics of the statistically generated frames (middle plot) is compared to the inter-frame dynamics of the natural phoneme (top plot), and it is clearly seen that the statistical features have a much lower dynamics.

To investigate the inter-frame dynamics of each phoneme we apply a DFT of length dT_i to the whole set of T_i feature frames representing p_i (*i-th* phoneme), $i = 1, 2, \ldots, L$. Obviously, in the transform domain the inter-frame dynamics in p_i is expressed by the *inter-harmonic* frequencies: $k+1, k+2, \ldots, k+T_i-1$, where $k = 1, T_i, 2T_i, \ldots, (d-1)T_i$.

Comparing the variation from frame to frame of statistically generated and natural phonemes in the transform domain, one observes an essential difference between the two. The spectrum of the statistically generated phoneme features, represented as one-dimensional sequence, has spectral components that are mostly located at the harmonic frequencies $k = lT_i$, $l = 1, 2 \ldots, d$, while the transformed natural phoneme coefficients sequence occupies inter-harmonic frequencies as well, as seen in Fig.4.1. It is seen in this figure that the inter-harmonic content of the statistical phoneme (dot-dashed line) is much lower (by $\sim 20 - 30$dB) than in the natural phoneme (solid line). This inter-harmonic content describes the variation from frame to frame within a particular phoneme. This confirms our assumption that inter-frames dynamics of statistical phonemes is too low.

Consequently, we propose to improve the inter-frame dynamics by enhancing in each phoneme the transform components at the inter-harmonic frequencies. Thus, the inter-frame dynamics can be better modeled by the non-harmonic components instead of by $\underline{\triangle}^{1,2}$.

We propose to enhance the amplitude of the inter-harmonic frequencies in the transformed features sequence by learning the statistics of the inter-harmonic content in a training stage for every phoneme and, afterwards,

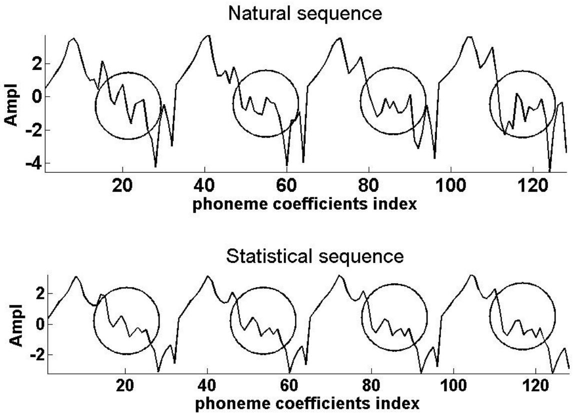


Figure 4.1: *Features frames of a natural sequence (4 frames on the same plot) at the top; statistical sequence at the bottom. The frame-to-frame variations between circled regions demonstrate the low dynamics in the statistical sequence as compared to the natural sequence.*

to match the inter-harmonic content at the synthesis stage to the acquired statistics, as described below. In Fig. 4.1 we see on the top plot that conventional generated speech features (solid line) include less dynamics, as compared to natural speech features (dashed line). In the bottom plot we see that speech feature enhanced in the transform domain (dashed line) have more dynamics compared to the natural speech feature.

Learning inter-harmonic content

For all natural segments pertaining to a particular phoneme p_i, where a segment consists of the features of contiguous natural frames from the database assigned to p_i, we learn inter-harmonic amplitude statistics as follows. We apply a DFT of length dT_i to the sequence of coefficients of a natural segment. The transformed sequence has d harmonic components and $d(T_i - 1)$

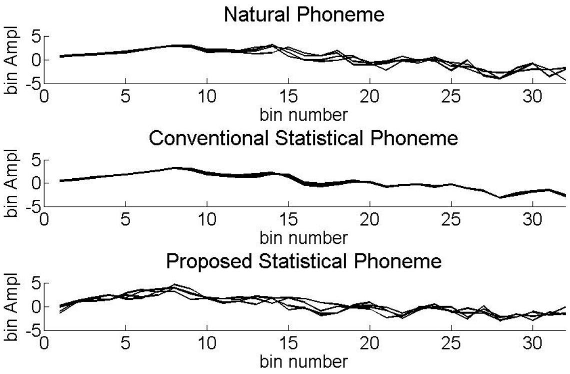


Figure 4.2: Features frames of natural phoneme (4 frames on the same plot) at the top; Conventional statistically generated phoneme in the middle; Proposed statistically generated phoneme at the bottom.

inter-harmonic components (that is, $T_i - 1$ components between every two harmonic components). The mean and variance of the inter-harmonic amplitudes located between every two harmonic component are computed. Thus each element $\widetilde{m}_k$, $k = 1, 2, \ldots, d$, of the inter-harmonic content mean vector $\underline{\widetilde{M}}_{d\times 1}$, is the mean value of the amplitudes of the inter-harmonic component located between *k-th* and *(k+1)-th* harmonic components. The static features statistics, namely, the first d component of $\underline{M}_{p_i}$ in Section 3, are computed, as in the conventional model described in Section 3.2.

Phoneme-level synthesis with inter-harmonic content

In the synthesis stage of a segment (representing p_i) of T_i frames, the mean of the static features of its model is repeated T_i times in order to get a one-dimensional sequence of length dT_i. This one dimensional sequence is transformed by a DFT. The phase of the transformed sequence is stored. Clearly,

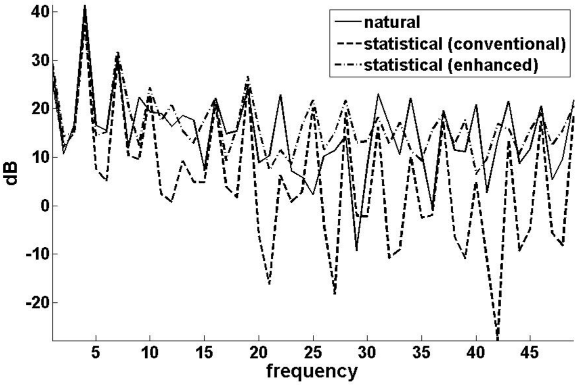


Figure 4.3: Magnitude of transformed natural sequence (3 frames) in thin solid line; Magnitude of transformed conventional statistical sequence in dashed line; Magnitude of transformed statistical sequence with enhanced inter-harmonic content in dot-dashed line. Because of symmetry of the magnitude sequence, only the 48 (=$32 \cdot 3/2$) positive frequencies are shown.

the inter-harmonic components of the transformed sequence are exactly zero because no dynamics is present in the one dimensional sequence due to its construction by replication. We propose to compute the components within the k-th inter-harmonic interval by a least squares approximation by a polynomial of order 2 of the points H_k, $\widetilde{m}_{\{k, replicated\ (T_i-1)\ times\}}$, H_{k+1}, where H_k is the k-th harmonic component and $\widetilde{m}_k$ is mean of the k-th interval inter-harmonic amplitudes obtained in the training stage. The dot-dashed line in Fig.4.1 depicts the enhanced amplitudes, which are very close to that of the natural amplitudes. A gain factor of of T_i is applied to inter-harmonic component amplitudes to match their level to the number of frames T_i. Finally, the inter-harmonic and harmonic components are combined appropriately and inverse-transformed by means of the IDFT, using the original phase stored earlier. As a result, we get a segment (representing phoneme) with

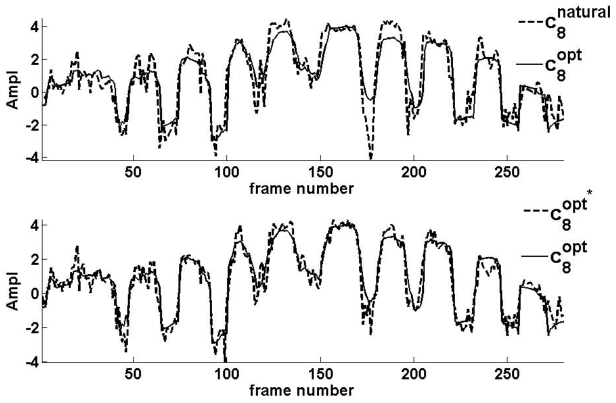


Figure 4.4: Demonstrating features over-smoothing. Top plot: variation in time of $c_8^{natural}$ (dashed line) and of c_8^{opt} (solid line). Bottom plot: variation in time. The optimal solution is over-smoothed and has much less dynamics (inter-frame variations) as compared to the natural segment.

the required static features and enhanced inter-frame dynamics, as seen in the bottom plot of Fig.4.1.

4.2 Utterance-level synthesis

Problem setting

In conventional statistically generated speech features, the inter-frame transitions are smoothed both within phonemes (intra-phoneme) and at the inter-phoneme boundaries. Obviously, intra-phoneme frames transitions should not be smoothed but rather be synthesized according to their dynamics, as modeled above by inter-harmonic components. On the other hand, inter-phoneme boundaries transitions should indeed be smoothed in order to avoid discontinuities. Consequently, these two types of frames should be subject to different treatment, which is not possible in the conventional statistical speech synthesis. In order to derive an optimal solution over an entire utterance with intra- and inter-phoneme frames being treated differently, we propose to modify the linear transformation W in (3.5).

Modified linear transformation

For a particular sequence of phonemes $(p_1, p_2, \ldots, p_L)$ of lengths $(T_1, T_2, \ldots, T_L)$, respectively,we propose to model the intra-phoneme frames in the transform domain, as proposed in Section 4.1, while modeling inter-phoneme transitions by the conventional differences, $\underline{\triangle}^{1,2}$, and to combine them by applying a modified linear transformation $\widehat{W}_{(4 \cdot d \cdot (L-1)) \times d \cdot N}$ instead of W in (3.5):

$$\widehat{W}{=}(\omega^1; \beta^1; \omega^2; \beta^2 \ldots; \beta^{i-1}; \omega^i;\ \beta^{i+1} \ldots; \beta^{L-1}; \omega^L), \tag{4.1}$$

$$(;\ \ \textit{denotes}\ \ \textit{vertical}\ \ \textit{concatenation})$$

where $\omega^i = [\mathbf{0}_{d \cdot T_i \times d \cdot \sum_{k=1}^{i-1} T_k}\ \mathbf{I}_{d \cdot T_i \times d \cdot T_i}\ \mathbf{0}_{d \cdot T_i \times d \cdot \sum_{k=i+1}^{L} T_k}]$ is constructed to preserve the dynamics of intra-phoneme frames modeled in the transform domain, and β^i, shown in (4.2), smoothes the transitions between p_{i-1} and p_i by applying $\underline{\triangle}^{1,2}$:

$$\beta^i = \begin{bmatrix} \mathbf{0}_\rho & -\frac{\mathbf{1}}{\mathbf{2}}_\xi & \mathbf{0}_\xi & +\frac{\mathbf{1}}{\mathbf{2}}_\xi & \mathbf{0}_\xi & \mathbf{0}_\eta \\ \mathbf{0}_\rho & -\mathbf{1}_\xi & \mathbf{2}_\xi & -\mathbf{1}_\xi & \mathbf{0}_\xi & \mathbf{0}_\eta \\ \mathbf{0}_\rho & \mathbf{0}_\xi & -\frac{\mathbf{1}}{\mathbf{2}}_\xi & \mathbf{0}_\xi & +\frac{\mathbf{1}}{\mathbf{2}}_\xi & \mathbf{0}_\eta \\ \mathbf{0}_\rho & \mathbf{0}_\xi & -\mathbf{1}_\xi & \mathbf{2}_\xi & -\mathbf{1}_\xi & \mathbf{0}_\eta \end{bmatrix},$$

where $\rho = d \times (d \cdot \sum_{k=1}^{i-1} T_k - 2 \cdot d),\ \xi = d \times d,\ \eta = d \times (d \cdot \sum_{k=i+1}^{L} T_k - 2 \cdot d)$ and $(\cdot)_y$ denotes a block of size y of stated dimensions.

Utterance-level optimal solution

In order to derive the optimal solution $\underline{C}^{opt^*}$ over an entire utterance, we rearrange the model mean and the covariance matrix to be compatible with $\widehat{W}$. The intra-phoneme frames are modeled in the transform domain, while, to satisfy smooth transitions at the phoneme boundaries, $\underline{\triangle}^{1,2}$ are constrained at boundary frames. Consequently, for a particular sequence of phonemes $(p_1, p_2, \ldots, p_L)$ of lengths $(T_1, T_2, \ldots, T_L)$, respectively, the utterance model mean vector and covariance matrix are:

$$\underline{\widehat{M}} = [\underline{\widehat{M}}^{p_1}_{l_1}, \underline{\triangle}^*{}_q, \underline{\widehat{M}}^{p_2}_{l_3}, \underline{\triangle}^*{}_q, \ldots, \underline{\triangle}^*{}_q, \underline{\widehat{M}}^{p_L}_{l_L}]^T, \tag{4.2}$$

$$\scriptstyle l_i = d \cdot T_i \times 1,\ q = 4 \cdot d \times 1;$$

$$\begin{aligned}\widehat{U} = diag[&{}^s\widehat{U}^{p_1}_{\tilde{l}_1}, {}^{\triangle^1}\widehat{U}^{p_1}_{\tilde{q}}, {}^{\triangle^2}\widehat{U}^{p_1}_{\tilde{q}}, {}^{\triangle^1}\widehat{U}^{p_2}_{\tilde{q}}, {}^{\triangle^2}\widehat{U}^{p_2}_{\tilde{q}}, {}^s\widehat{U}^{p_2}_{\tilde{l}_2},\\ &{}^{\triangle^1}\widehat{U}^{p_2}_{\tilde{q}}, {}^{\triangle^2}\widehat{U}^{p_2}_{\tilde{q}}, {}^{\triangle^1}\widehat{U}^{p_3}_{\tilde{q}}, {}^{\triangle^2}\widehat{U}^{p_3}_{\tilde{q}}, \ldots, {}^{\triangle^1}\widehat{U}^{p_{L-1}}_{\tilde{q}},\\ &{}^{\triangle^2}\widehat{U}^{p_{L-1}}_{\tilde{q}}, {}^s\widehat{U}^{p_L}_{\tilde{l}_L}],\ \scriptstyle \tilde{l}_i = d \cdot T_i \times d \cdot T_i,\ \tilde{q} = d \times d.\end{aligned} \tag{4.3}$$

where $\underline{\widehat{M}}^{p_i}_{1 \times d \cdot T_i}$ is the mean vector of phoneme p_i in the features domain, with the dynamics that was enhanced in the transform domain; $\underline{\triangle}^*{}_{1 \times 4 \cdot d}$ constrains the values of $\underline{\triangle}^{1,2}$ at phonemes boundaries; ${}^s\widehat{U}^{p_i}_{d \cdot T_i \times D \cdot T_i}$ is the covariance matrix of the static features for p_i; ${}^{\triangle^1}\widehat{U}^{p_i}_{d \times d}$ and ${}^{\triangle^2}\widehat{U}^{p_i}_{d \times d}$; are the covariance matrices of the differences $\underline{\triangle}^{1,2}$ at boundary frames, respectively. $\underline{\widehat{M}}$ is column vector, $\widehat{U}$ is a block diagonal square matrix.

Consequently, using (4.1), (4.2) and (4.3) in (3.12), the optimal solution is $\underline{C}^{opt^*}=(\widehat{W}^T\widehat{U}^{-1}\widehat{W})^{-1}\widehat{W}^T\widehat{U}^{-1}\widehat{\underline{M}}$, where the intra-phoneme frames with enhanced dynamics are optimally combined with smoothed inter-phoneme transitions.

4.3 Subjective evaluation

To evaluate the proposed approach we checked: *a)* Whether the inter-frame variations in C^{opt^*} are consistently higher, as compared to those of C^{nat}. *b)* Whether the naturalness of speech generated from C^{opt^*} is improved, in comparison to speech generated by the conventional approach from C^{opt}. This aspect was evaluated by a subjective listening test.

To obtain an objective evaluation for the inter-frame variations of speech features, we computed the measure $\Lambda = mean(\sum_{i=1}^{N-1} \|\underline{c}_{i+1} - \underline{c}_i\|$ for 30 sentences generated from C^{nat}, C^{opt^*} and C^{opt}. The averaged Λ value over these sentences was 4.81, 4.37, and 1.5 for C^{nat}, C^{opt^*}, and C^{opt}, respectively. In the bottom plot of Fig.4.4 we see that the C^{opt^*} has much more dynamics than C^{opt} does. This provides an objective support to the proposed dynamics enhancement method.

As stated above, we also performed an informal listening test to evaluate subjectively the improvement in the naturalness of the proposed approach in comparison to conventional statistically generated sentences. The test includes 20 entries, where each entry is a triplet with the same sentence appearing three times, in an order related to C^{nat}, C^{opt^*}, C^{opt}. The same sentence appears in another entry but in an different order related to C^{nat}, C^{opt}, C^{opt^*}. The listeners were asked to compare the naturalness of speech generated from C^{opt^*} and C^{opt} to the same sentence generated from C^{nat} in a CTTS system, and indicate which of the two sounds closer to the CTTS sentence. The total preference score given to C^{opt^*} was 81.7%, while for C^{opt} it was just 18.3%. This provides a subjective support to the proposed synthesis method. Notwithstanding the promising results, the naturalness of C^{opt^*} is still worse than that of C^{nat}, so more work is needed to improve the naturalness of STTS.

4.4 Conclusion

In this chapter we have presented a method for enhancing intra-phoneme speech features dynamics in the transform domain and for smoothly combining phonemes into an utterance while maintaining the enhanced dynamics. The improvement in comparison to conventional STTS is supported by performed subjective tests results, without increasing much the computational complexity.

Chapter 5

Segment-Wise Representation

In this chapter we propose an alternative technique, to the technique presented in Chapter 4, for improving the statistically generated speech quality. This technique is superior in terms of the generated speech quality to the technique, presented in Chapter 4. Here, we propose a method to enhance a baseline STTS system by introducing a segment-wise model representation with a norm constraint.

5.1 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, $\boldsymbol{\Delta}_t^{1,2}\mathbf{c}_t, ..., \boldsymbol{\Delta}_{t+d_t-1}^{1,2}\mathbf{c}_{t+d_t-1}$, 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 3.1(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:

$$\bar{\mathbf{o}}_t = \frac{1}{d_t}\sum_{k=1}^{d_t}\mathbf{o}_k, \tag{5.1}$$

and

$$J(\bar{\mathbf{o}}_t) = \frac{1}{2}\|\mathbf{U}_{q_t}^{-\frac{1}{2}}(\bar{\mathbf{o}}_t - \mathbf{m}_{q_t})\|_2^2, \tag{5.2}$$

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:

$$\widetilde{\mathbf{W}}_{q_t} \triangleq \frac{1}{d_t}\begin{pmatrix} \mathbf{0} & \mathbf{1} & \underset{d_t-2}{\cdots\mathbf{1}\cdots} & \mathbf{1} & \mathbf{0} \\ -\frac{\mathbf{1}}{\mathbf{2}} & -\frac{\mathbf{1}}{\mathbf{2}} & \underset{d_t-2}{\cdots\mathbf{0}\cdots} & \frac{\mathbf{1}}{\mathbf{2}} & \frac{\mathbf{1}}{\mathbf{2}} \\ -\mathbf{1} & \mathbf{1} & \underset{d_t-2}{\cdots\mathbf{0}\cdots} & \mathbf{1} & -\mathbf{1} \end{pmatrix}_{3M\times M(d_t+2)} . \tag{5.3}$$

All the matrix elements in (5.3) 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 (5.4):

$$\widetilde{\mathbf{W}}_{MK\times MN} \;=\; \left[\begin{array}{ccccccccc} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & \cdots \\ \cdots & -\frac{1}{6} & -\frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{6} & 0 & 0 & \cdots \\ \cdots & -\frac{1}{3} & -\frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & \cdots \\ \cdots & 0 & 0 & 0 & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \cdots \\ \cdots & 0 & 0 & 0 & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{array}\right]_{MK\times MN} \tag{5.4}$$

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:

$$\widetilde{\mathbf{W}}\mathbf{c} - \widetilde{\mathbf{m}} \;=\; [\; \mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_t}, \ldots, \mathbf{w_K} \;]^T, \tag{5.5}$$

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

$$\mathbf{w_t} \;\triangleq\; \tfrac{1}{d_t} \sum_{i=t-\lfloor \frac{d_t}{2} \rfloor}^{t+\lfloor \frac{d_t}{2} \rfloor} \mathbf{o}_i^T - \mathbf{m}_{qt}^T. \tag{5.6}$$

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

$$J^{sw}(\widetilde{\mathbf{W}\mathbf{c}}) = \frac{1}{2}\|\widetilde{\mathbf{U}}^{-0.5}(\widetilde{\mathbf{W}\mathbf{c}} - \widetilde{\mathbf{m}})\|_2^2, \tag{5.7}$$

where

$$\widetilde{\mathbf{m}}_{3MK\times 1} = [\mathbf{m}_{q_1}^T, \mathbf{m}_{q_2}^T, \ldots, \mathbf{m}_{q_K}^T]^T, \tag{5.8}$$

and

$$\widetilde{\mathbf{U}}_{3MK\times 3MK}^{-1} = diag[\mathbf{U}_{q_1}^{-1}, \mathbf{U}_{q_2}^{-1}, \ldots, \mathbf{U}_{q_K}^{-1}] \tag{5.9}$$

are the non-replicated model mean vector and the covariance matrix, respectively, where $\widetilde{\mathbf{m}}_{3MK\times 1}$ and $\widetilde{\mathbf{U}}_{3MK\times 3MK}^{-1}$ consist of K state means, $\mathbf{m}_{q_t}^T$, and state covariance matrices, $\mathbf{U}_{q_t}^{-1}$, respectively, This is in contrast to the frame-wise model mean vector, $\mathbf{m}_{3MN\times 1}$, and the frame-wise model covariance matrix, $\mathbf{U}_{3MN\times 3MN}^{-1}$, defined in (3.10) and (3.11), 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 (3.12) 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, (5.7), is derived by the same steps as in (3.13), (3.14) and (3.2):

$$\begin{aligned}\frac{\partial J^{sw}(\widetilde{\mathbf{W}}\mathbf{c})}{\partial \mathbf{c}} =& -\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{W}}\mathbf{c} + \widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{m}} \\ =& \, 0 \end{aligned} \tag{5.10}$$

consequently,

$$\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{W}}\mathbf{c} = \widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{m}}. \tag{5.11}$$

Assuming the matrix $\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{W}}$ in (5.11) is invertible, the optimal segment-wise solution $\mathbf{c}^{opt,sw}$ is derived by:

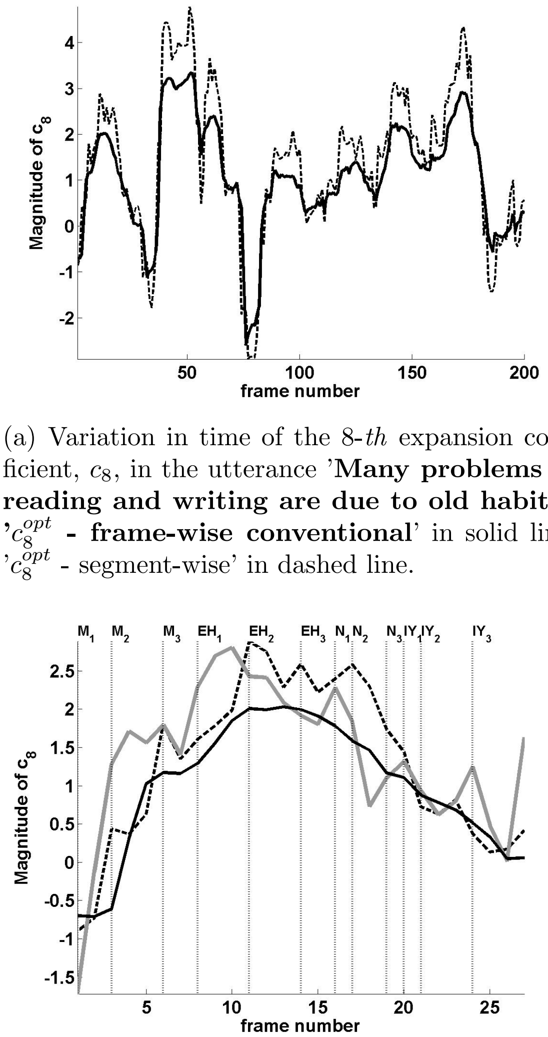
$$\mathbf{c}^{opt,sw} = (\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{W}})^{-1}\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{m}}. \tag{5.12}$$

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 5.2.

In Fig 5.1(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. 5.1(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. 3.1(b), is provided as a reference for expected speech feature dynamics. In the more detailed view of Fig. 5.1(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. 3.1(b). The last fact emphasizes our assumption regarding insufficient dynamics in the conventional frame-wise models.

The proposed segment-wise representation occupies less memory during synthesis than does the conventional frame-wise representation because the former and the later representations model dimensions are derived from: $\widetilde{\mathbf{W}}_{3MK\times MN}$, $\widetilde{\mathbf{m}}_{3MK\times 1}$, $\widetilde{\mathbf{U}}_{3MK\times 3MK}$, and $W_{3MN\times MN}$, $\mathbf{m}_{3MN\times 1}$, $\mathbf{U}_{3MN\times 3MN}$, respectively, where N is the number of frames and K is the number of models (segments) in a synthesized utterance.

In our experiments, we compared the empirical data fitting by the segment-wise cost function, $J^{sw}(\widetilde{\mathbf{W}}\mathbf{c})$, to the empirical data fitting by the conventional frame-wise cost function, $J^{fw}(\mathbf{Wc})$. We performed this comparison by computing the cost functions values on real speech examples, $\mathbf{c}^{natural}$. In Fig. 5.2 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



(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

Figure 5.1: Comparison of speech feature dynamics in a conventional frame-wise model to that of the proposed segment-wise representation.

states have lower value in the segment-wise model, due to its more flexible construction, providing more degrees of freedom. The longer the state du-

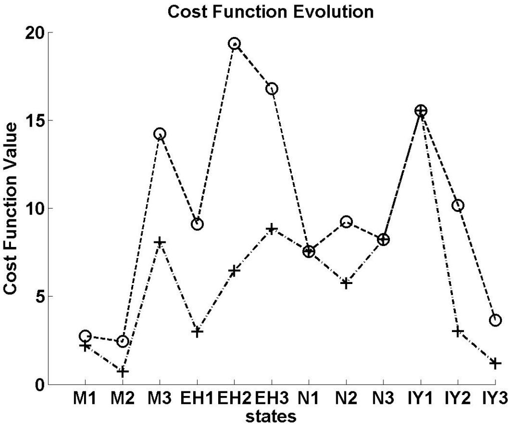


Figure 5.2: Evolution of cost function over states of the word '**Many**' in a real speech sample, $\mathbf{c}^{natural}$: the segment-wise cost function, $J^{sw}(\widetilde{\mathbf{W}}\mathbf{c}^{natural})$, in pluses; the frame-wise cost function, $J^{fw}(\mathbf{W}\mathbf{c}^{natural})$, in circles. x-axis - corresponding states, y-axis - cost function value.

ration, 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.

5.2 Norm constraint

We have observed 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 (3.2), 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 $\|\mathbf{c}^{mdl}\|_2^2$:

$$\|\mathbf{c}^{stt}\|_2^2 \approx \|\mathbf{c}^{mdl}\|_2^2 \tag{5.13}$$

In Fig. 3.1(a) and 3.1(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 (3.12), 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 [26] 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:

$$\gamma_0 = \frac{\widetilde{\|\mathbf{c}^{nat}\|_2^2}}{\widetilde{\|\mathbf{c}^{stt}\|_2^2}}, \tag{5.14}$$

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

Consequently, using (5.13), a constraint on the norm of speech features, $\|\mathbf{c}^{stt}\|_2^2$, should be equal to

$$\Gamma = \gamma_0 \cdot \|\mathbf{c}^{mdl}\|_2^2, \tag{5.15}$$

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

$$\|\mathbf{c}^{stt,sw}\|_2^2 \approx \|\mathbf{c}^{nat}\|_2^2 \tag{5.16}$$

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

5.3 Norm-constrained cost function

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.

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 (5.7), using a factor λ to balance the contribution of the two terms.

Thus, the cost function of (5.7) is replaced by:

$$\begin{aligned} J_c^{sw}(\widetilde{\mathbf{W}}\mathbf{c}) &\triangleq \frac{1}{2}\|\mathbf{U}^{-\frac{1}{2}}(\widetilde{\mathbf{W}}\mathbf{c} - \mathbf{m})\|_2^2 \\ &+\frac{\lambda}{2}\|\mathbf{c}\|_2^2, \end{aligned} \tag{5.17}$$

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

5.4 Iterative algorithm

We propose an iterative algorithm that minimizes the model cost function value, defined in 5.17, while assuring sufficient dynamics in the resulting solution. The minimization is done by means of a gradient descent algorithm as follows:

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \alpha_n \nabla(\mathbf{c}_n), \tag{5.18}$$

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:

$$\alpha_n = \tfrac{1}{\|\nabla(\mathbf{c}_n)\|_2^2}, \tag{5.19}$$

and from (5.17),

$$\begin{aligned} \nabla(\mathbf{c}_n) &= \widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\widetilde{\mathbf{W}}\mathbf{c}_n - \widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\mathbf{m} \\ &+\lambda\mathbf{c}_n. \end{aligned} \tag{5.20}$$

A final feature vector should approximate well the models, and have a norm value that is compatible with the enhancement factor, defined in (5.14). 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, (5.20) is replaced by :

$$\begin{aligned}\nabla(\mathbf{c}_n) &= \widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\widetilde{\mathbf{W}}\mathbf{c}_n - \widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\mathbf{m} \\ &\quad +\lambda_n\mathbf{c}_n,\end{aligned} \tag{5.21}$$

where λ_n is updated according to:

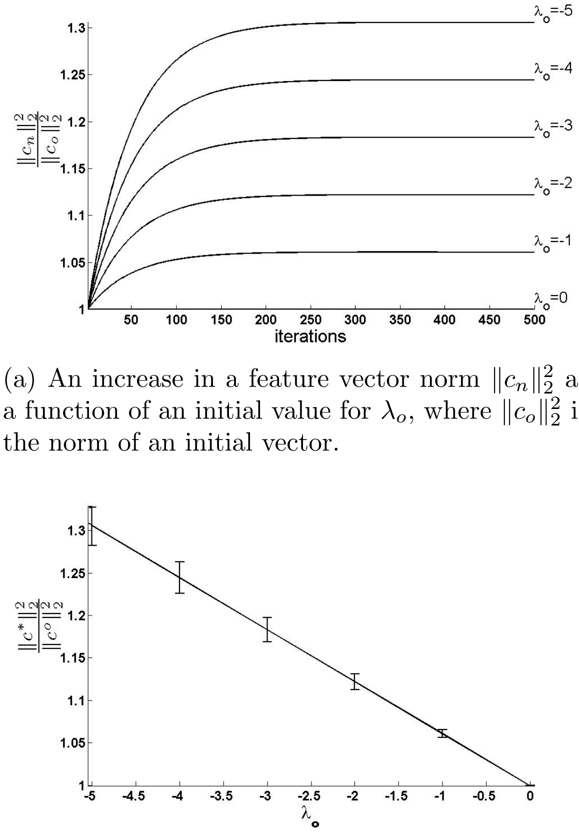
$$\lambda_{n+1} = \theta\lambda_n,\ 0 \le \theta \le 1, \tag{5.22}$$

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 (5.7), 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. 5.3(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



(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.

Figure 5.3: Evolution of $\|c_n\|_2^2$ as function of λ_o.

vectors of dimension $MN \times 1$.

The relation between λ_0 and, the attained maximal value of the feature vector norm $\|c^*\|_2^2$, is represented in Fig. 5.3(b) that is derived from Fig. 5.3(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. 5.3(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 (5.14), resulting in enhanced dynamics in generated speech, as confirmed by listening tests described in Section (5.5).

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

In Fig. 5.4, 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 conven-

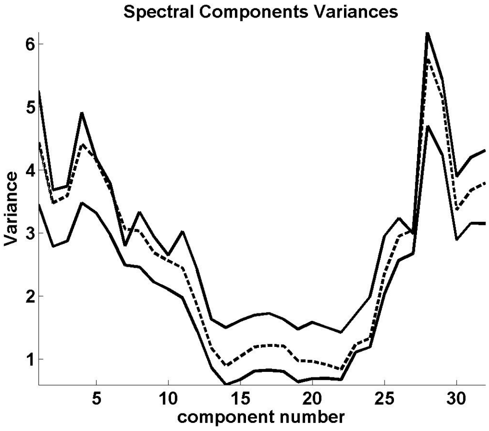


Figure 5.4: Frequency components variances of natural utterance (top solid line): conventional STTS generated (bottom solid line); and proposed STTS generated utterance (middle dashed line).

tional 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. 5.4 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 (3.2). Clearly, the iterative solution via (5.18) 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 (5.18) with the segment-wise model, converges to a vector that approximates the required norm, as shown in Fig. 5.3(a).

5.5 Subjective Evaluation

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

5.5.1 Mean opinion score (MOS) tests

5.5.1.1 Test I

In this test we have computed the Mean Opinion Score (MOS), according to [31], 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 3, (group A), *(ii)* by the proposed speech generation scheme (group B), and *(iii)* by IBM's CTTS system, detailed in [4], [3], (group C). Thus, 27 samples were included in the test, each of which was evaluated by 20 listeners. To eliminate prosody influence, the same target prosody was provided to all versions of a particular sentence.

Fig. 5.5 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.

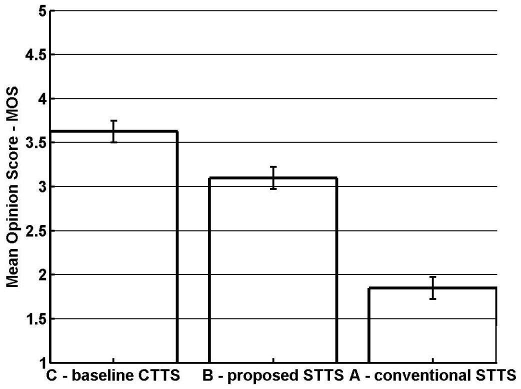


Figure 5.5: Mean Opinion Score (MOS) test, comparing the CTTS system, the proposed STTS method, and the base-line STTS. The error bars indicate 95% confidence interval, computed using the 't-test'.

5.5.1.2 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 [26]. 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 [32]:

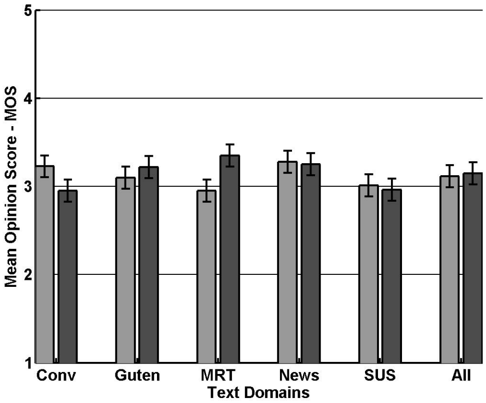


Figure 5.6: Mean Opinion Score - MOS test for different text domains: 'Conv', 'Guten', 'MRT', 'News', 'SUS'. 'All' - average score for all text domains. The samples by the proposed method and by the GV method are in light gray and dark gray, respectively'. The error bars indicate 95% confidence interval, computed using the 't-test'.

- Gutenberg novels - 'Guten'.
- Standard news text - 'News'.
- Conversational/dialog sentences - 'Conv'.
- Phonetically confusable words - 'MRT', detailed in [11].
- Semantically unpredictable sentences - 'SUS', detailed in [1].

Fig. 5.6 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 [2]. This high score for analysis-synthesized speech means that an 'encoder/decoder' introduces only a small speech quality degradation.

5.5.2 'A vs B' comparison test

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 3, (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. 5.7 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.

5.5.3 Subjective evaluation setup

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.

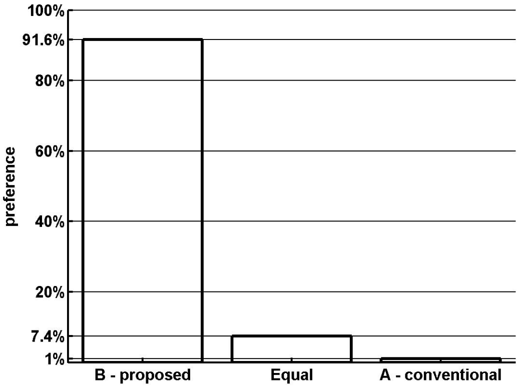


Figure 5.7: 'A vs B comparison test' **(20 listeners, 11 pairs of sentences)**: the proposed STTS (group B) was preferred in 91.6% of the cases, on average; 7.4% were judged as having the same quality; and the conventional STTS (group A) was preferred only in 1% of the cases

5.5.4 Conclusion

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.

However, statistically generated speech is inferior in terms of the generated speech naturalness, compared to the naturalness of speech generated by CTTS. In the next chapter we propose an hybrid text to speech synthesis method, which is comprised from STTS and CTTS. Speech generated by the proposed hybrid TTS synthesis method has less unpleasant audible discontinuities, than speech generated by the baseline CTTS system, described in Chapter 2. And, it is more natural than speech generated by the baseline STTS system.

Chapter 6

Hybrid Text-to-Speech Synthesis

One of the goals of the current research is to efficiently combine the advantages of CTTS and STTS into another class of TTS systems, named hybrid TTS system. The hybridism in the proposed system is in interweaving of natural segments with statistically generated segments, using a novel hybrid dynamic path algorithm. As a result, speech generated by the proposed HTTS includes less discontinuities than the baseline CTTS system does, and it sounds more natural than the baseline STTS.

The proposed system is based on *a)* Determination of a hybrid dynamic path that defines positions for statistical models within an utterance, aimed to include as many as possible long natural segments sequences, where natural sequences are smoothly connected in an optimal way by statistical generated segments, *b)* Representation of the hybrid speech feature vector over an entire utterance, *c)* A gradient descent algorithm with linear constraints, where statistical segments within a synthesized utterance are generated from constrained statistical model [29], while natural parts stay unchanged, yet they affect the statistically generated parts.

6.1 Different hybridism meanings

In this chapter we discuss the different meanings of the term 'hybridism' in the context of text to speech synthesis, as it appears in the literature.

The term hybrid TTS has different connotations when related to TTS synthesis. One class of TTS systems employs hybridism by combining different speech feature representations, as detailed in [14], [23], [10]. Another class of TTS systems employs hybridism by combining different modeling architectures, as detailed in [18], [13], [17], [12]. There are other systems which use statistically generated speech features as a target for segments selection in CTTS, as detailed in [15].

6.2 Hybrid dynamic path

6.2.1 Introduction

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 occur when initially contiguous natural segments can not be concatenated due to data pre-selection and/or data fragmentation, as detailed in [7], [5]. The concatenation of natural segments in CTTS systems is done by means of the Viterbi dynamic programming (DP) algorithm, as detailed in [6], [4].

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. The resulting perfect CTTS-generated speech is expected to have natural speech quality. Unfortunately, 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.

To justify this assumption we examined a concatenative TTS, having 1,300,000 recorded speech feature segments, which are clustered to 25,000 different acoustic clusters/acoustic tree leaves, as described in Section 2.1. Such a system is considered, as having a large footprint, since 25,000 acoustic leaves sufficiently cover phonetic contexts of English.

We synthesized a set of 40 arbitrary sentences six times. Each time the maximal number of allowable transitions between any two consecutive stages of the dynamic search, described in Section 2.3, was different. We tested 1, 10, 100, 1,000, 10,000, and 50,000 maximal allowable transitions between any two consecutive stages of the dynamic search. In any stage (in any acoustic leaf) of the dynamic search, more frequent segments appear before less frequent segments.

In Fig. 6.1 we see a relation between the discontinuities rate within synthesized utterances to the maximal number of allowable transitions in each stage of the dynamic search. A discontinuity is defined when spectral dis-

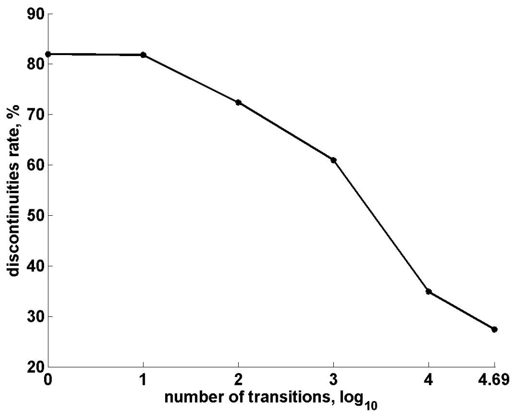


Figure 6.1: Relation between the discontinuities rate in percents, y-axis, to the logarithm of the maximal number of allowable transitions in each stage of the dynamic search, x-axis, where $log10(50,000) \approx 4.69$.

tance between two consequent segments is higher than a permitted spectral distance.

The discontinuities rate reflects the number of segments that are concatenated not to their original neighboring segments, but instead to other segments, resembling their original neighboring segments. We see that even for 50,000 transitions, there are still about 20% of segments which are connected to different segments than their original neighbors. The discontinuities rate was averaged over the set of 40 sentences.

Fig. 6.2 presents a relation between synthesis time to the maximal number of allowable transitions in each stage of the dynamic search. We see that increasing the the maximal number of allowable transitions in the dynamic search enlarges significantly the synthesis time[1]. The synthesis

[1] The represented times do not reflect the real synthesis times, but rather emphasize the big difference between the synthesis time, required for the dynamic search to find the optimal sequence of speech segments, using more than 100 allowable transitions to the

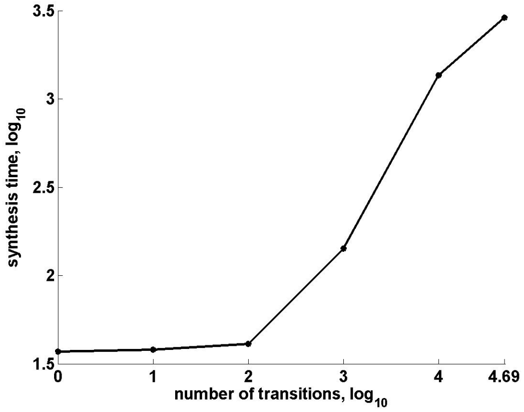


Figure 6.2: Relation between synthesis time, y-axis, to the logarithm of the maximal number of allowable transitions in each stage of the dynamic search, x-axis.

time was averaged over the set of 40 sentences.

Also, we examined how much the discontinuities rate is affected by the number of acoustic leaves/phonetic contexts in the system. We repeated the above experiment with different numbers of acoustic leaves: 5,000, 10,000, 15,000, 20,000, 25,000, 30,000, 35,000, 40,000, 45,000, 50,000. We found that for any of these numbers of acoustic leaves a low discontinuities rate is achieved only for a huge number of stage-to-stage transitions in the dynamic search, as shown in Fig. 6.3

Consequently, we conclude that the perfect CTTS system is not feasible, because in order to cover sufficiently phonetic contexts of a given language, the system needs a lot of recorded speech feature segments, that results in an exponential increase of the synthesis time.

time required to find the optimal sequence, using less than 100 transitions.

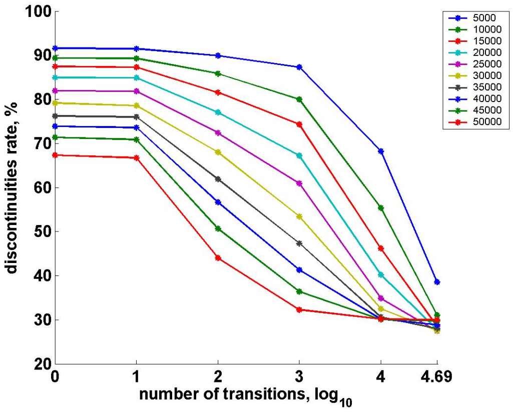


Figure 6.3: Relation between the discontinuities rate in percents, y-axis, to the logarithm of the maximal number of allowable transitions in each stage of the dynamic search, x-axis. Every line corresponds to a different number of acoustic leaves. The numbers of acoustic leaves is shown in the legend, in the order the lines appear on the plot, where the top line corresponds for 5000 acoustic leaves.

6.2.2 Hybrid dynamic path algorithm

We develop the idea of the perfect CTTS system, discussed in Section 6.2.1, in the proposed HTTS system. We propose to interweave natural segments with statistically generated segments, where positions of statistical segments encourage as long as possible natural segments sequences, as they appear at the training database. Consequently, we try to approximate the behavior of the abstract CTTS, by concatenating as many as possible contiguous natural segments.

We propose to determine the positions of statistical segments by means of an Hybrid Viterbi DP algorithm, denoted HDP, as described below.

Assume that we have a sequence of contexts $L_1, L_2, \ldots, L_K$, representing the stages of the HDP, where the context L_i holds the segments $n^i_1, n^i_2, \ldots, n^i_{N_i}$, representing the hybrid nodes of the HDP, as shown Figure.6.4, where C_k^{i-1} and $e_{k,1}^i - 1$ denotes the cumulative cost of a survivor path of node n_k^{i-1} and the transition cost between n_k^{i-1} and n_1^i, respectively.

Any hybrid 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, as described in Section 2.2, to ensure smooth connections to adjacent natural segments.

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 for each node in each stage in order to find the optimal path by back tracking through the best survivor path. When computing a survivor for node n_1^i, the first node at the stage L_i in the HDP shown in Fig.6.4, the existence of the following condition is examined:

$$\forall j \;\; e_{j,1}^{i-1} > \epsilon, \tag{6.1}$$

where ϵ is a permitted spectral distance (error).

If such a case exists, any path passing through n_1^i includes a discontinuity at the transition from L_{i-1} to L_i. Consequently, there is a possible degradation in generated speech quality due to spectral discontinuity.

We propose to replace n_1^i to a boundary constrained statistical model s_1^i. The statistical node s_1^i connects smoothly to its neighbors, consequently, a

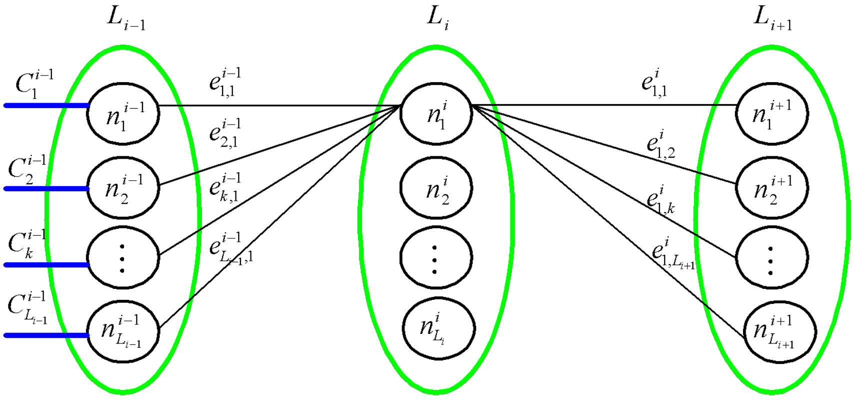


Figure 6.4: Hybrid dynamic path. $e_{j,k}^{i-1}$ is a spectral distance between the node n_j^{i-1} and n_k^i, C_j^i is the best partial path at the node n_j^i, L_i - i-th stage of the dynamic search.

survivor of s_1^i is defined as:

$$p_{s_1^i} = \underset{j}{argmin}\; C_j^{i-1}, \tag{6.2}$$

which means that s_1^i continues smoothly the best survivor path from stage L_{i-1} to stage L_i.

Although, the node n_1^i is replaced by the statistical node s_1^i, the spectral distance from the nodes at the stage L_{i+1} to the statistical node s_1^i is computed as:

$$e_{i,j}^1 \quad = \quad d(n_1^i, n_j^{i+1}), \quad j = 1, \ldots, L_{t+1},$$

instead of

$$e_{i,j}^1 \quad = \quad d(s_1^i, n_j^{i+1}), \quad j = 1, \ldots, L_{t+1},$$

in order to check whether n_1^i has its original right boundary neighbor, as they appeared in the training database, in L^{i+1}. If node n_1^i connects smoothly to some node n_j^{i+1} of L_{i+1}, 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}.

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

The number of statistical segments within a generated utterance can be controlled 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 every distance is positive or zero by definition, and every transition distance would be higher than any negative permitted distance, $\epsilon < 0$. While setting it to any positive number, results in HTTS, and, finally, setting it 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.

6.3 Hybridism ratio

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 controls the ratio of the numbers of natural segments to the numbers of statistically generated segments comprising a synthesized utterance. Speech generated in an in-

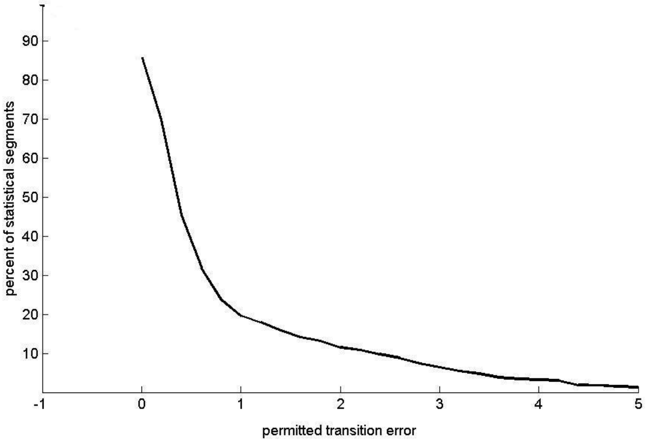


Figure 6.5: Relation between the *hybridism ratio* parameter, ξ, and the permitted spectral distance, ϵ. Any negative value for ϵ results in a pure statistical system.

termediate (hybrid) mode consists of natural and statistically generated segments, interweaved within an utterance.

We established an empirical relation between ξ and the permitted distance parameter, ϵ, mentioned in (6.2.2), for the used database, as shown in Fig.6.5. In Fig. 6.5 we see the ratio (in percent) of statistical segments to the overall number of segments in an utterance. Obviously, the statistical segments ratio is 100% for any negative value of the permitted distance. While requiring to concatenate only adjacent segments from the training database corresponds to a permitted spectral distance value, $\epsilon = 0$, which results in about 85% of segments to be statistical, as shown in Fig. 6.5. So, we can set a permitted distance parameter value according to this relation in order to get a hybrid utterance with a required number of statistical segments.

This relation was found by synthesizing an arbitrary set of 40 sentences in the proposed hybrid TTS, having footprint of 8.3MB. This set of sentences was generated using $\epsilon \in \{-1, 0 \text{ to } 5 \text{ in steps of } 0.2\}$. For each permitted spectral distance from the used range, a ratio of statistical segments within an utterance, ξ, was computed as an average ratio over this sentences.

6.4 Boundary constrained model

For the assumed model in (3.2), the optimal solution $\mathbf{c}^{opt}$ is the most probable statistically derived vector over the utterance of N frames. In a hybrid TTS 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 connections of natural segments to statistical generated segments within a whole speech feature vector $\mathbf{c}_{dN\times 1}$ are done by means of $\mathbf{\Delta}^{1,2}\mathbf{c}_i$ as follows. Assume that we have to connect the natural segment $\mathbf{c}^{nat} = [\mathbf{c}_1^{nat}, \mathbf{c}_2^{nat}, \ldots, \mathbf{c}_{T_i}^{nat}]$, having T_i frames, to the left boundary of a 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 the following constraints on the left boundary dynamic features $\mathbf{c}^{stt}$:

$$\widetilde{\mathbf{\Delta}}^1\mathbf{c}_1 = \frac{1}{2}(\mathbf{c}_2^{stt} - \mathbf{c}_{T_i}^{nat}), \quad (6.3)$$

$$\widetilde{\mathbf{\Delta}}_1^2\mathbf{c}_1 = (-\mathbf{c}_{T_i}^{nat} + 2\mathbf{c}_1^{stt} - \mathbf{c}_2^{stt}), \quad (6.4)$$

while in the unconstrained boundary synthesis they are:

$$\begin{aligned}\mathbf{\Delta}^1\mathbf{c}_1 &= \frac{1}{2}(\mathbf{c}_2^{stt} - \mathbf{c}_{T_i}^{stt}), & (6.5)\\ \mathbf{\Delta}_1^2\mathbf{c}_1 &= (-\mathbf{c}_{T_i}^{stt} + 2\mathbf{c}_1^{stt} - \mathbf{c}_2^{stt}). & (6.6)\end{aligned}$$

Here we present a general framework for generating hybrid speech feature vector over an entire utterance with arbitrary number and positions of natural frames, even if an entire utterance is composed either totally from natural segments or from statistical generated segments, as in CTTS and STTS systems, respectively.

Setting the optimization problem as a constrained optimization problem, a vector $\mathbf{c}^{opt}$ is derived not by the expression (3.2) but rather by:

$$\begin{aligned}\mathbf{c}^{opt} &= \underset{\mathbf{c}}{argmin}\ ln(P(\mathbf{Wc})). & (6.7)\\ \text{s.t.} \quad & \mathbf{Ac} = \mathbf{c}^*, &\end{aligned}$$

where $\mathbf{c}^*_{dk\times 1} = [\mathbf{c}^{*\,T}_{i_1}, \mathbf{c}^{*\,T}_{i_2}, \ldots, \mathbf{c}^{*\,T}_{i_k}]^T$ is a vector that includes the k constrained natural frames 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 $\mathbf{c}^*_{dk\times 1}$. For example, for $d = 1$, $N = 5$, $k = 3$, and $i_1 = 1$, $i_2 = 3$, $i_3 = 4$, the matrix $\mathbf{A}$ is:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

We derive the optimal solution for this constrained optimization problem (6.7) by means of the Lagrangian function with a vectorial Lagrange multiplier, γ:

$$L(\mathbf{c},\gamma) \quad = \quad ln(P(\mathbf{Wc})) + (\mathbf{Ac} - \mathbf{c}^*)^T\gamma, \tag{6.8}$$

where $\gamma_{dk\times 1}$ is the vectorial Lagrange multiplier. Consequently, assuming that $\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W}$ is invertible, the speech feature vector for the boundary constrained model is derived by:

$$\begin{aligned} \frac{\partial L(\mathbf{c},\gamma)}{\partial \mathbf{c}} &= \mathbf{0}, \quad \Rightarrow \\ \mathbf{c}^{opt} &= (\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{U}^{-1}\mathbf{M} \\ &+ (\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W})^{-1}\mathbf{A}^T\gamma, \end{aligned} \tag{6.9}$$

and using (6.9) in (6.8) γ is

$$\begin{aligned} \gamma &= (\mathbf{A}(\mathbf{W^TU^{-1}W})^{-1}\mathbf{A}^T)^{-1}\mathbf{c}^* \\ &- (\mathbf{A}(\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W})^{-1}\mathbf{A}^T)^{-1}\mathbf{A}(\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{U}^{-1}\mathbf{M}. \end{aligned} \tag{6.10}$$

Consequently, the optimal speech feature vector, $\mathbf{c}^{opt}$, is derived by substituting (6.10) into (6.9).

We can see in Figures 6.6 and 6.7 that the boundary constrained optimal solution has two obviously different types of frames. The hybrid utterance includes 42% of conventionally statistical generated frames. The segments, where the hybrid (dashed line) features do not coincide with the natural features, were not chosen from the natural segments inventory but rather were supplanted by the statistically generated segments according to the hybrid dynamic path, as described in Section 6.2. The natural segments within the hybrid utterance are set as constraints for the statistical model, gen-

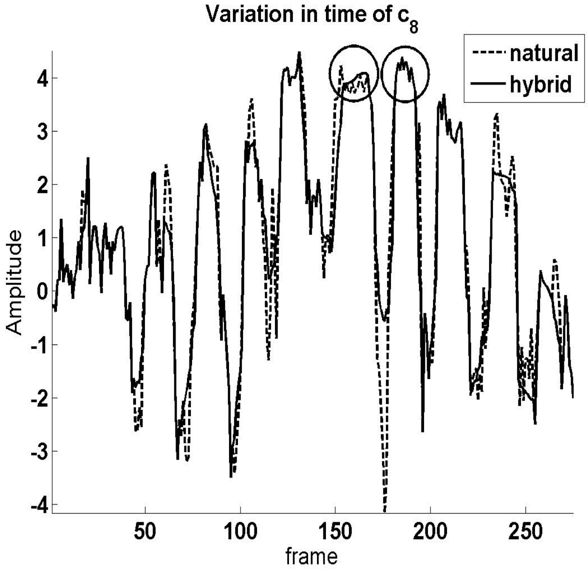


Figure 6.6: Comparison of a hybrid-generated utterance to a natural utterance.

erating the entire utterance. Frames pertaining to natural segments have much more variations than those pertaining to conventional statistically generated segments. This difference in the characteristics of natural segments and statistical segments generated by conventional STTS may result in unpleasant artifacts in the generated speech.

In the next section we demonstrate an approach which resolves the above mentioned quality discrepancy between natural and statistical segments by applying SW-STTS instead of the conventional STTS.

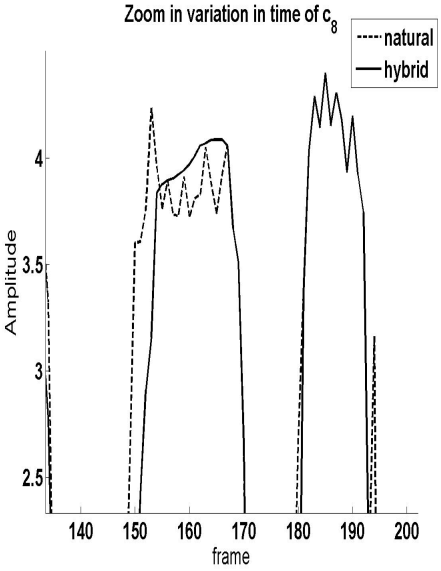


Figure 6.7: Zooming into the circled segments at the utterance at Fig.6.6.

6.5 Hybrid speech generation algorithm

In Section 6.4 we demonstrated the optimal solution (6.9) for a hybrid speech feature vector, where an arbitrary number of statistical segments may appear at arbitrary positions within an utterance, and natural segments are considered as boundary constraints for statistical models. The constraints are defined by the linear transformation A, defined in (6.8), and the optimization problem is set as the Lagrangian function (6.8) with the vectorial Lagrangian multiplier $\gamma_{dK\times 1}$, where the connections between statistical to natural sequences are optimal and smoothed by means of $\mathbf{\Delta}^{1,2}\mathbf{c}_i$.

The hybrid optimal speech feature vector over an entire utterance (6.9) includes over-smoothed conventional statistically generated speech features because no enhancement to speech features dynamics was applied. Although, the resulting speech naturalness is higher than the naturalness of speech generated by a pure statistical model due to the contribution of natural segments to the overall quality of the generated utterance, the low dynamics in statistically generated segment may introduce unpleasant artifacts.

On the other hand, the method introduced in Chapter 5, generates speech features having enhanced dynamics, resulting in the improved naturalness of the overall statistical generated speech. However, that algorithm introduces no natural segments to a generated utterance because no frames are constrained to the natural ones during the iterative generation.

We propose to combine the hybrid speech feature vector representation, shown in Section 6.4 with the algorithm shown in Section 5.4.

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 (5.17):

$$J_{c,c}(\widetilde{\mathbf{W}}\mathbf{c}) = \|\mathbf{U}^{-1}(\widetilde{\mathbf{W}}\mathbf{c} - \mathbf{M})\|_2^2 + (\mathbf{A}\mathbf{c} - \mathbf{c}^*)^T\gamma + \lambda\|\mathbf{c}\|_2^2, \qquad (6.11)$$

where the first term tends 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 features vector norm, as described in Section 5.4. Using the gradient descent algorithm, a hybrid speech feature vector after the n-*th* iteration is:

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \alpha_n \widetilde{\nabla}(\mathbf{c}_n), \tag{6.12}$$

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

$$\begin{aligned}\widetilde{\nabla}(\mathbf{c}_n) &= \widetilde{\mathbf{W}}^{\mathbf{T}}\mathbf{U}^{-\mathbf{1}}\widetilde{\mathbf{W}}\mathbf{c}_n - \widetilde{\mathbf{W}}^{\mathbf{T}}\mathbf{U}^{-\mathbf{1}}\mathbf{m} + \mathbf{A}^T\gamma + \lambda_n\mathbf{c}_n, \\ &= \mathbf{P}\mathbf{c}_n - \mathbf{Q} + \mathbf{A}^T\gamma + \lambda_n\mathbf{c}_n, \end{aligned} \tag{6.13}$$

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

Taking into consideration that $\mathbf{A}\widetilde{\nabla}(\mathbf{c}_n) = 0$[1], $\forall n$, we can compute the vectorial Lagrangian constraint γ from:

$$\mathbf{A}\mathbf{P}\mathbf{c}_n - \mathbf{A}Q + \mathbf{A}\mathbf{A}^T\gamma + \lambda_n\mathbf{A}\mathbf{c}_n = 0, \tag{6.14}$$

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

$$\begin{aligned}\gamma &= \mathbf{A}\mathbf{Q} - \mathbf{A}\mathbf{P}\mathbf{c}_n - \lambda_n\mathbf{A}\mathbf{c}_n \underset{\mathbf{A}\mathbf{c}_n=\mathbf{c}^*,\forall n}{=} \\ &= \mathbf{A}\mathbf{Q} - \mathbf{A}\mathbf{P}\mathbf{c}_n - \lambda_n\mathbf{c}^*, \end{aligned} \tag{6.15}$$

The gradient in the update step (6.12) is:

$$\widetilde{\nabla}(\mathbf{c}_n) = \mathbf{P}\mathbf{c}_n - \mathbf{Q} + \mathbf{A}^T\mathbf{A}\mathbf{Q} - \mathbf{A}^T\mathbf{A}\mathbf{P}\mathbf{c}_n - \lambda_n\mathbf{A}^T\mathbf{c}^* + \lambda_n\mathbf{c}_n, \tag{6.16}$$

where λ_n is updated by the rules described in Section 5.4.

Obviously, natural frames influence statistically generated segments,

[1] Multiplying the equation (6.12) by the constraints defining matrix $\mathbf{A}$, as defined in (6.8), we get that $\mathbf{A}\mathbf{c}_{n+1} = \mathbf{A}\mathbf{c}_n - \alpha_n\mathbf{A}\nabla(\mathbf{c}_n)$, where $\mathbf{A}\mathbf{c} = \mathbf{c}^*$, $\forall n$. Consequently, equation (6.12) becomes $\mathbf{c}^* = \mathbf{c}^* - \alpha_n\mathbf{A}\nabla(\mathbf{c}_n)$, and, $\mathbf{A}\nabla(\mathbf{c}_n) = 0$.

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 during the iterations of the algorithm. The proposed scheme combines the hybrid speech feature vector representation with the iterative speech feature vector algorithm. As a result, the overall quality of the proposed hybrid generated speech, using a statistical model with improved dynamics is better than the quality of hybrid generated speech that uses the conventional STTS, as confirmed by listening tests demonstrated in Section 5.5.

In Fig.6.8 we see speech features generated by the hybrid speech generation algorithm, presented in this section, as compared to the natural speech features, and to the hybrid speech features generated using the conventional statistical model, presented in Section 6.4, in solid grey line, in dashed line, and, in solid black line, respectively. The constrained frames are natural speech frames, where all three lines coincide, such as in frames: 7-12, 28-32, 39-44, 94-100, 102-108, and 113-118. Examining the frames 61-71 we see that, in general, statistical speech features generated using the conventional statistical model are less dynamic than statistical speech features generated by applying the segment-wise statistical model, SW-TTS.

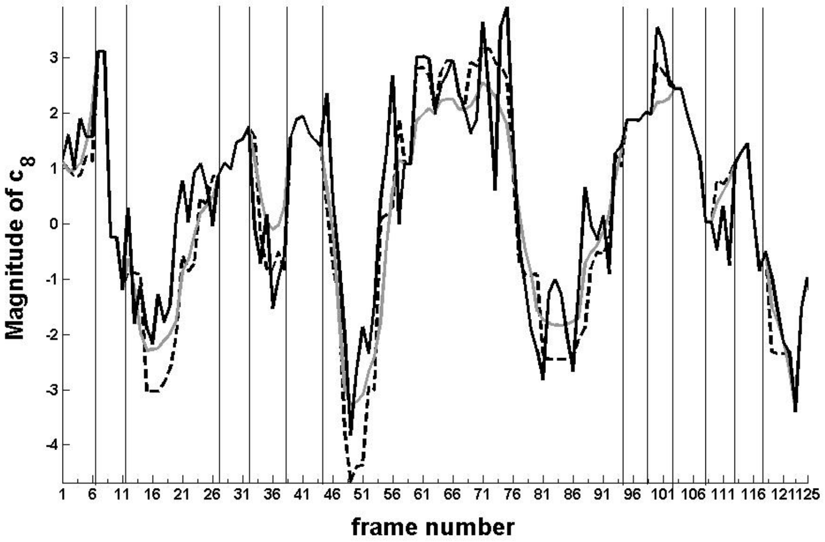


Figure 6.8: Comparison of a hybrid-generated utterance using the segment-wise representation statistical model, to a hybrid-generated utterance using the conventional statistical model, and to a natural utterance, in dashed line, in solid grey line, and, in solid black line, respectively. The vertical lines mark the constrained natural segments, where all the lines coincide.

6.6 HTTS Experimental Setup

We examined different compositions of the baseline CTTS and the baseline STTS in the proposed HTTS system. We simulated a CTTS system with different footprints (by using different numbers of speech feature segments), having memory size of 5MB, 7MB, 8.3MB, 12MB, and 22MB, respectively. The simulated system footprints were controlled by a number of stage-to-stage candidate transitions in the dynamic search.

All the systems had the same number of acoustic leaves/phonetic contexts, which was set to 25,000.

The size of the STTS models for 25,000 acoustic leaves was 1.3MB, approximately. The sizes of the corresponding HTTS systems were examined 6.3MB, 8.3MB, 12.3MB, and 23.3MB.

The HTTS system was simulated in a double pass fashion. In the fist pass a particular sentence was generated by the baseline CTTS, which provided spectral phases for the second pass, In the second pass, the hybrid dynamic path algorithm, described in Section 6.2.2, was applied. The hybrid dynamic-path algorithm determines a hybrid segment sequence, where some t of the segments are natural speech feature segments, while others are corresponding statistical models. These models generate statistical speech feature segments, applying either the conventional statistical speech feature generation algorithm of Section 3.2, with the boundary constrains, Section 6.4, or, the segment-wise model speech feature generation algorithm, described in Section 5.4 with the boundary constrains, demonstrated in Section 6.4. Simulating a HTTS system in a double pass fashion enables us to compare utterances generated by the HTTS system to those generated by the CTTS, where the only difference in generated utterances is in the quality of spectral speech features. The proposed HTTS can work in a single pass fashion as well, (independent of the spectral phases of the underlined CTTS system), in which case a linear phase is used for statistical generated utterances with a dithering in high frequencies, as described in [2].

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 appear in generated utterances.

Examining different compositions of hybrid utterances, we found that almost all natural segments are replaced by statistical models in a HTTS based on the 5MB baseline CTTS. On the other hand, almost all natural segments remain in a hybrid utterances for a HTTS based on the 22MB CTTS. We conclude that the HTTS system is less useful when it is based on CTTS system that have small sizes, e.g. 5MB, or, on a big size CTTS, such as 22MB.

The HTTS system, having an intermediate size (among the examined HTTS systems) of around 7MB-8MB, interweaves a marked amount of both segments types (natural and statistical), where the ratio between natural segments to statistical models varies from 30% to 70%. Consequently, we chose this working point for the HTTS in our experiments.

6.7 Subjective evaluation

We have performed listening tests to evaluate the quality of speech generated by the proposed HTTS method.

6.7.1 Mean opinion score (MOS) test

In this test we have computed the Mean Opinion Score (MOS), according to [31], for a set of 10 sentences, where each sentence was generated in four versions: *(i)* CTTS system, using up to 15 possible candidate segments in each step of the dynamic programming, (a memory size of $22MB$), described in equation (2.3), Section 2, (group A). *(ii)* Proposed hybrid text-to-speech synthesis method, applying the segment-wise model representation and the iterative algorithm, demonstrated in Section 5.4, with a 7MB baseline CTTS, (group B). *(iii)* CTTS system, using up to 6 possible candidate segments in each step of the dynamic programming, (a memory size of $8.3MB$), described in equation (2.3), Section 2, (group C). *(iv)* Proposed hybrid text-to-speech synthesis method, applying the conventional statistical speech generation method, discussed in Section 3, with a 7MB baseline CTTS, (group D).

Fig. 6.9 shows the results of the MOS test for the four groups. We see that the proposed method outperforms the concatenative system that is using up to 6 candidate segments in each step of the dynamic programming step. However, it is inferior to the concatenative system with 15 candidates segments in each step of the dynamic programming search. The hybrid

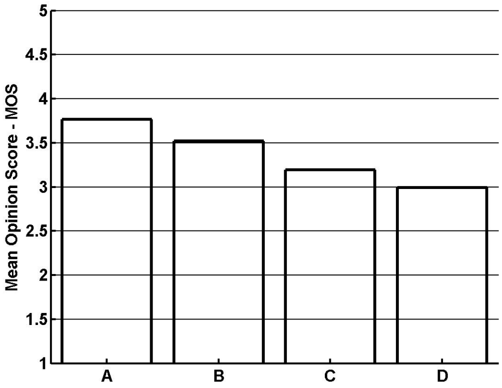


Figure 6.9: Mean Opinion Score (MOS) test, comparing CTTS system, using up to 15 possible candidate segments in each step of the dynamic programming search, (a memory size of $22MB$), (group A), *(ii)* by the proposed hybrid text-to-speech synthesis method, applying the segment-wise model representation and the iterative algorithm, with a 7MB baseline CTTS, (group B), *(iii)* CTTS system, using up to possible 6 candidate segments in each step of the dynamic programming search, described in equation (2.3), (a memory size of $8.3MB$), (group C), *(iv)* by the proposed hybrid text-to-speech synthesis method, applying the conventional statistical speech generation method, with a 7MB baseline CTTS, (group D).

system that uses the conventional statistical speech generation algorithm, with 7MB baseline CTTS, is worse than all the other examined methods.

6.7.2 Subjective evaluation summary

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.

Chapter 7

Summary and Future Work

In this chapter we discuss advantages of the proposed HTTS for both the baseline STTS and the baseline CTTS. In Section 7.1 we discuss the improvement of STTS by the proposed HTTS. In Section 7.2 we discuss the improvement of CTTS by the proposed HTTS. Also, in that Section we consider the effect of footprint size of the baseline CTTS system on the performance of the proposed HTTS. The advantages of the proposed HTTS are demonstrated by the results of the performed subjective evaluations. In Section 7.3 we summarize the entire work. And, finally, in Section 7.4 we provide possible continuations of this research.

7.1 Improvement of STTS by HTTS

As described previously, the main disadvantage of STTS synthesis is its non natural 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, provided in Section 6.4, enables to connect smoothly statistical generated speech- feature vectors to natural speech-feature vectors, where positions of natural speech feature-vectors are determined by the hybrid dynamic path algorithm, provided in Section 6.2.

Consequently, the overall naturalness of statistically generated speech can only be improved by the introduction of natural speech segments.

In Fig. 7.1 we see that introducing natural speech feature-segments to conventional statistical generated speech feature-vector results in a MOS increase of about one MOS unit.

The quality of segment-wise statistically generated speech increases by the hybrid scheme as well. However, the increase is lower than with conventional statistical generated speech. The reason is that the initial segment-wise quality is higher than the quality of conventional statistical generated speech.

In Fig. 7.2 we see that introducing natural speech feature-segments to segment-wise statistical generated speech feature-vector results in a MOS increase of about half a MOS unit.

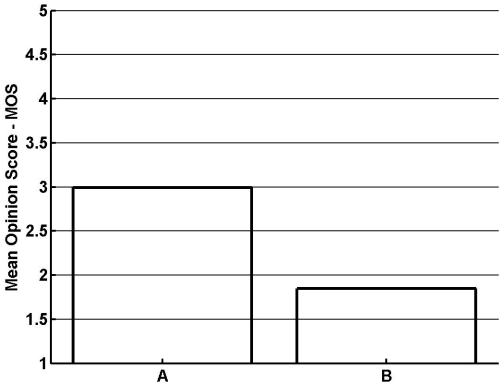


Figure 7.1: Contribution of CTTS to the quality of the conventional STTS. Mean opinion score (MOS) of the proposed HTTS, based on the conventional STTS, with 7MB baseline CTTS, column 'A'; MOS of the base line pure STTS, column 'B'.

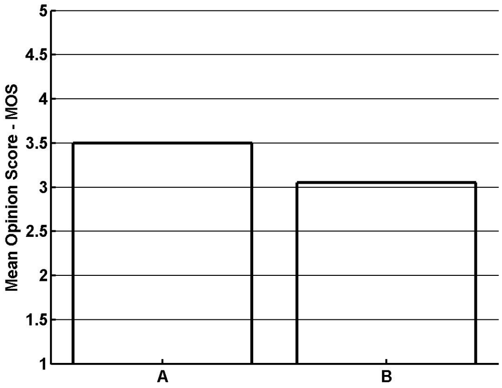


Figure 7.2: Contribution of CTTS to the quality of the SW-STTS. Mean opinion score (MOS) of the proposed HTTS, based on the SW-STTS, with 7MB baseline CTTS, column 'A'; MOS of the base line pure SW-STTS, column 'B'.

We conclude that a STTS system is improved by augmenting it by a CTTS system. The footprint of the CTTS system can be low, including only most frequent natural segments.

7.2 Improvement of CTTS by HTTS

The main drawback of CTTS synthesis is the possibility of obtaining abrupt transitions between adjacent speech feature segments. These abrupt transitions often cause unpleasant audible artifacts in the generated speech. The lower a CTTS system footprint, the more discontinuities appear in the generated speech. Yet, in spite of these artifacts it possesses natural features and is not muffled and/or buzzy, as compared to statistically generated speech.

In this research we found that the proposed HTTS method can reduce discontinuities in speech features by applying boundary constraints in the hybrid dynamic path, thus improving the overall generated speech quality. In Fig. 7.3 we see that the proposed HTTS system, using the segment-wise statistical speech model, column 'A', outperforms the CTTS system, column 'B'. The footprint of each of these two systems is approximately 8.3MB.

A HTTS system is preferable over a CTTS system when discontinuities in speech generated by a CTTS system cause an essential degradation in its quality. We found that a HTTS system, composed of a baseline CTTS system and of a STTS system, having footprints of 7MB, and 1.3MB, respectively, outperforms a CTTS system alone, having a footprint of 8.3MB. The smaller the footprint of the baseline CTTS system, the larger is the improvement in the quality of speech generated by the HTTS system. The bigger the footprint of the baseline CTTS is, the rarer are discontinuities in

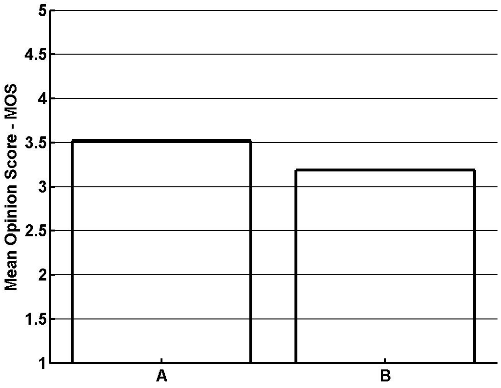


Figure 7.3: Mean opinion score (MOS) of the proposed HTTS, based on the segment-wise STTS, column 'A'; MOS of the CTTS system, column 'B'. Both systems have the same size of 8.3MB

generated speech, and, as a result the improvement obtained by the HTTS system decreases.

7.3 Summary

In this research we examined the characteristics of STTS systems. The main disadvantage of STTS systems is the over-smoothing of statistically generated speech features, which causes the generated speech to sound muffled and buzzy. We found that the over-smoothing problem can be alleviated by improving statistical speech features dynamics. We proposed two approaches for the enhancement of speech feature dynamics. The first one improves speech features dynamics in the transform domain, as described in Section 4. In the second approach we improved speech features dynamics by introducing a segment-wise representation, described in Section 5.2.

The segment-wise representation enables more speech feature dynamics by providing additional degrees of freedom for speech features evolution. This is realized by requiring the approximation of the model mean by all frames of a corresponding segment on average, rather than by its approximation by each frame of the corresponding segment, as it is performed in the conventional frame-wise representation. The additional degrees of freedom of the segment-wise representation are also employed for increasing the norm of generated speech feature-vector. The norm increase is regulated by introducing a norm-constraint, as described in Section 5.3. To generate a speech feature-vector over an entire utterance, we use an iterative algorithm, described in Section 5.4, where the speech feature-models and the speech feature-norm are balanced by a balancing factor, whose effect is decreased from iteration to iteration in the iterative algorithm. As a result, according to the performed MOS tests, described in Section 5.5, the generated speech sounds more natural and less buzzy. However, its naturalness is still worse than the naturalness of CTTS, having a big footprint size. The first approach (frequency domain) is less computational complex, however, it is inferior in term of the quality, in comparison to the second (segment-wise) approach.

In addition, we designed a hybrid TTS system by combining STTS with CTTS. The designed HTTS combines the advantageous characteristics of STTS, (optimal smoothed transitions between adjacent segments) with those of CTTS (the 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 6.2. In order to optimally connect natural segments to statistical models, boundary constrained statistical models are applied, 6.4. A hybrid speech feature-vector over an entire utterance is generated iteratively, where natural segments are fixed in the iterations, and constrain statistical segments, while statistical segments are updated according to the hybrid cost function gradient, as described in Section 6.5. As a result, according to the performed MOS tests (Section 6.7) hybrid generated speech sounds more natural than a corresponding pure statistical 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 a big footprint (20MB and more) are not improved by combining them with STTS, because in such a system, there is only a very small number of audible discontinuities in generated speech. On the other hand, CTTS systems having a small footprint (5MB and less), generate speech for which almost all segments are do not connect smoothly, resulting in a noticeable degradation in the generated speech quality, due to many audible discontinuities. In this research we show that the proposed HTTS is advantageous at a working point of 7MB (a baseline CTTS size). However, determination of the working point is 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 provided in the next section.

7.4 Future Work

The segment-wise representation provides additional degrees of freedom in the determination of the statistically generated 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, e.g., speech feature-vector entropy, by properly choosing an additional term in the cost function, like we did for regulating the spectral speech features norm.

In this research we employed the gradient descent algorithm to derive iteratively the speech feature-vector over an entire utterance. Other approaches for generation of the speech feature vector can be examined both in terms of complexity and in terms of the final solution characteristics, e.g., applying different iterative algorithms, or analytic determination of the norm-regulating balancing factor.

In this research spectral speech features are explored, while phase speech features are taken either from the original CTTS phases, in the case of HTTS, or a linear phase is used in the case of STTS. Statistical modeling of phase speech features could be an important continuation of this research.

The proposed dynamic path algorithm is based on a cost function derived from the spectral distance between consequent 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, will 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 research 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.

The performance of TTS systems is mostly evaluated by subjective listening tests. An objective metric for TTS system evaluation will improve an evaluation process of TTS systems. Such a metric can facilitate finding a working point for HTTS system as well.

Appendix A

Analysis of Cost Function Components

In this appendix we elaborate on the behavior of the iterative algorithm, detailed in Section 5.4, 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 $J_3(\Delta^2\mathbf{c})$. 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

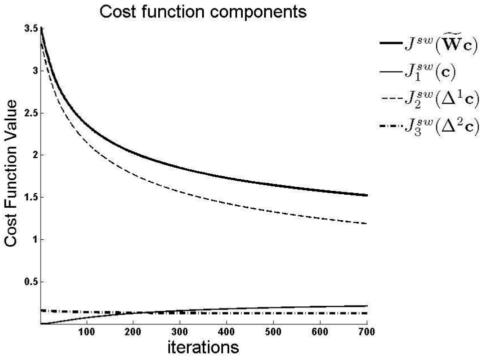


Figure A-1: The components of the unconstrained segment-wise cost function, $J^{sw}(\widetilde{\mathbf{W}\mathbf{c}})$, where $J^{sw}(\widetilde{\mathbf{W}\mathbf{c}}) = J_1^{sw}(\mathbf{c}) + J_2^{sw}(\boldsymbol{\Delta}^{\mathbf{1}}\mathbf{c}) + J_3^{sw}(\boldsymbol{\Delta}^{\mathbf{2}}\mathbf{c})$.

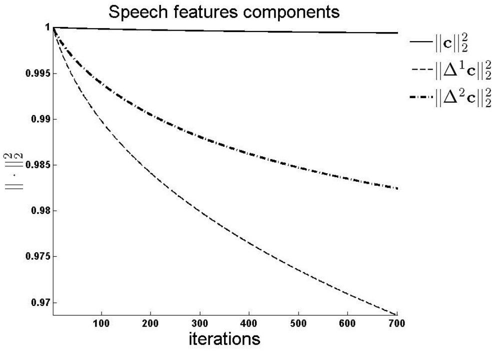


Figure A-2: The normalized static feature norm and the normalized dynamic feature norm, obtained by the minimization of $J^{sw}(\widetilde{\mathbf{W}\mathbf{c}})$.

(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 (3.4), 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 5.4.

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 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, $\frac{\lambda}{2}\|\mathbf{c}\|_2^2$,

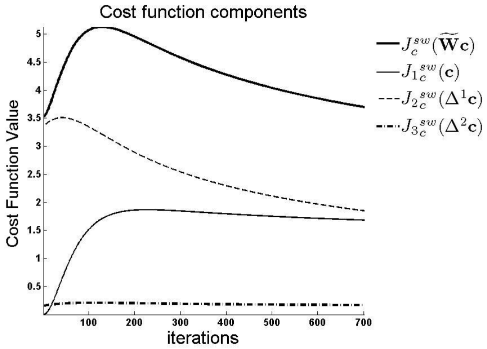


Figure A-3: The components of the constrained segment-wise cost function, $J_c^{sw}(\widetilde{\mathbf{Wc}})$, where $J_c^{sw}(\widetilde{\mathbf{Wc}}) = {J_1}_c^{sw}(\mathbf{c}) + {J_2}_c^{sw}(\boldsymbol{\Delta}^{\mathbf{1}}\mathbf{c}) + {J_3}_c^{sw}(\boldsymbol{\Delta}^{\mathbf{2}}\mathbf{c})$.

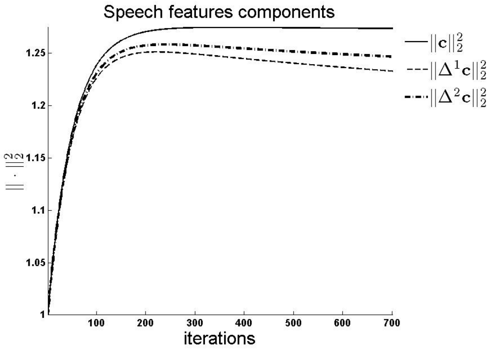


Figure A-4: The normalized static feature norm and the normalized dynamic feature norm, obtained by the minimization of $J_c^{sw}(\widetilde{\mathbf{Wc}})$.

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] 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. 59

[2] 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", ICASSP-2006, Toulouse, France. 19, 59, 81

[3] 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, Lisbon, Portugal, pp. 2569-2572. 18, 57

[4] R.E. Donovan, and E.M. Eide, "The IBM Trainable Speech Synthesis System", Proc. ICSLP-1998, Sydney, Australia, vol.5, pp. 1703-1706. 13, 22, 57, 64

[5] R. E. Donovan, "Text-to-speech using clustered context-dependent

phoneme-based units ", US Patent 6163769, issued on Dec. 19, 2000. 13, 22, 63

[6] R.E. Donovan, "Trainable speech synthesis", PhD thesis, Cambrige, June 1996. 13, 22, 64

[7] R. E. Donovan, "Topics in decision tree based speech synthesis", Computer Speech & Language Vol 17, no. 1, January 2003, pp. 43-67. 13, 24, 63

[8] D. Eberly, "Derivative Approximation by Finite Differences", 2001. 22

[9] S. Furui, "Speaker independent isolated word recognition based on dynamics emphesized cepstrum," Trans. IECE of Japan, vol. 69, Dec. 1986, pp. 1310-1317. 22

[10] M. Gary, C. Nishant, "Hybrid Speech Synthesizer, Method and use", United States Patent 20080195391, Oct. 2006. 63

[11] 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. 59

[12] L. Hongmao, "Hybrid-parameter mode speech synthesis system and method", United States Patent 20060161438, Jule 2006. 63

[13] S. Hei, S. Hawkins, "Hybrid approach to high-quality formant synthesis using HLsyn", 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, 1998. 63

[14] I. Iriondo, F. Alias, J. Sanchis, J. Melenchon, "A hybrid method oriented to concatenative text-to-speech synthesis", Eurospeech-2003, Geneva, Switzerland, pp, 2953-2956. 63

[15] J. Kominek, A. W Black, "The Blizzard Challenge 2006 CMU Entry introducing hybrid trajectory-selection synthesis". 63

[16] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis from HMMs using dynamic features," Proc. ICASSP-1996, Atlanta, Geargia, pp. 389-392. 9, 19, 22, 24, 28

[17] J. Matouek, Z. Hanzle, D. Tihelka, "Hybrid syllable-triphone speech synthesis“, Eurospeech-2005, Bonn,Germany, pp. 2529-2532. 63

[18] F. Malfrure, O. Deroo and T. Dutoit, "Phonetic Alignment: Speech Synthesis Based vs. Hybrid HMM/ANN". 63

[19] T. Okubo, R. Mochizuki, T. Kobayashi, "Hybrid Voice Conversion of Unit Selection and Generation Using Prosody Dependent HMM", IEICE Trans. on Information and Systems 2006 pp. 2775-2782.

[20] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, Feb. 1989, pp. 257-286. 22, 23

[21] F.K. Soong, A.E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," Proc. ICASSP-1986, Tokyo, Japan, pp. 877-880. 22

[22] K. Schnel and A. Lacroix, "Combination of LSF and Pole Based Parameter Interpolation for Model Based Diphone Concatenation", Interspeech-2007, Antwerp, Belgium, pp. 2897-2900. 19

[23] A. Spanias, "A hybrid model for speech synthesis", Circuits and Systems, May 1990, vol. 2, pp. 1521 - 1524. 63

[24] 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, Madrid, Spain, pp. 757-760. 9, 22, 24, 28

[25] K. Tokuda, T. Yoshimura, T. Masuko and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. ICASSP-2000, Istanbul, Turkey, pp.1315-1318. 24

[26] T. Toda, K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis", Proc. Interspeech-2005, Lisbon, Portugal, pp. 2801-2804. 9, 28, 50, 58

[27] K. Tokuda, T. Kobayashi and S. Imai, "Speech parameter generation from HMM using dynamic features," Proc. IEEE ICASSP-1995, Detroit, Michigan, pp. 660-663. 9, 28

[28] S. Tiomkin, D. Malah, ”Statistical Text-to-Speech Synthesis with Improved Dynamics”, Interspeech-2008, Brisbane, Australia, pp. 1841-1844. 10, 31

[29] S. Tiomkin, D. Malah, and S. Shechtman, ”Statistical Text-To-Speech Synthesis based on Segment-wise Representation with a Norm Constraint”, submitted to IEEE Trans, Audio, Speech and Language Processing, 2008. 63

[30] 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, Jan. 2007, pp. 153-173. 9, 28

[31] ”Mean Opinion Score (MOS)”, Recommendation P.800, Telecommunication Standardization Sector, International Telecommunication Union (ITU-T), freely available at: ”$http : //www.itu.int/rec/T - REC - P.800 - 199608 - I/en$”. 57, 83

[32] ”$http : //www.synsig.org/index.php/Main_Page$”, SynSIG, ISCA. 58

גישה משולבת לסינתזה משופרת של אותות דיבור מטקסט

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים

סטאס טיומקין

הוגש לסנט הטכניון - מכון טכנולוגי לישראל
אייר תשס״ט חיפה מאי 2009

תודות

המחקר נעשה בהנחיית פרופסור דוד מלאך בפקולטה להנדסת חשמל.

אני רוצה להביע את תודתי העמוקה לפרופ' דוד מלאך על הנחייתו המסורה, הדרכתו ותמיכתו בכל מהלך המחקר. למדתי הרבה מהידע והניסיון המחקרי שלו. בהנחייתו, למדתי בין היתר כיצד לתכנן ולנהל מחקר אקדמי, התנסיתי בכתיבה מדעית, וכן כיצד להגדיר את מטרות המחקר ולהשיגן.

המחקר היה בשיתוף פעולה בין המעבדה לעיבוד אותות בטכניון לבין מעבדת המחקר של IBM בחיפה (HRL). אני רוצה להודות אישית לרון חורי, ראש קבוצת עיבוד דיבור ב-HRL, סלבה שכטמן, צבי קונס, אריאל שגיא ואלכס סורין מHRL-IBM על הדיונים המועילים במהלך המחקר.

אני רוצה גם להודות לצוות של מעבדה לעיבוד אותות בטכניון (SIPL) על העזרה. בפרט, אני רוצה להודות לנמרוד פלג, זיוה אבני, אבי רוזן ויאיר משה. זאת הייתה חוויה נפלאה לעבוד עם הצוות של SIPL.

אני מודה מאוד לאימי סופיה ולאחי אנטון על העידוד והתמיכה להם זכיתי מהם.

אני מקדיש את התיזה באהבה לאוולינה א.

אני מודה ל טכניון ו- HRL-IBM על התמיכה הכספית הנדיבה בהשתלמותי.

תקציר

ההתפתחות המהירה במערכות מחשב ויישומיהם דורשת שכלול של הממשק בין אדם למכונה. הרבה מערכות תהיינה פחות נוחות לשימוש אם התקני הפלט-קלט שלהם יהיו מוגבלים להתקני פלט-קלט סטנדרטיים כגון: עכבר, מקלדת ומסך. למערכות חדישות יש צורות שונות של ממשקים. בנוסף להתקני פלט-קלט סטנדרטיים, מתפתחים ממשקים ישירים בין אדם למכונה, כמו מערכות שמפיקות פלט בצורה של אות הדיבור. התקני קלט במערכות כאלה מתבססים על זיהוי אותות הדיבור והשפה, והתקני הפלט צריכים לדעת לתרגם תגובות מכונה לשפה אנושית ולסנתז את אות הדיבור בהתאם.

יש חשיבות רבה לשיפור של איכות מערכות לסינתזת אותות דיבור, מפני שקיימות מערכות אשר משתמשות בפלט קולי. כל פעילות אשר דורשת שתי ידיים, (נהיגה, תפעול מכונות מורכבות, מנתחים בחדרי ניתוח וכו'), ובו זמנית דורשת לתקשר עם מכונה, תהיה יותר יעילה אם תהיה אפשרות לתת למכונה פקודות קוליות ולקבל פלט מהמכונה בצורה של אותות דיבור. כמן כן, מכונות עם פלט-קלט קולי תהיינה נגישות לאנשים עם בעיות ראייה, כמו כן, ישנם הרבה יישומי מחשב אשר יכולים להוציא פלט קולי. בנוסף לנוחות, פלט-קלט קולי יכול להיות יותר טבעי ונעים למשתמש כי תגובות קוליות של מכונה יכולות להינתן בקול בעל תכונות רצויות ומוכרות למשתמש, כמו קול של אדם מסוים. ישנם כיום יותר ויותר מערכות עם ממשק קולי הן בתעשייה והן ביישומים אישיים, כמו טלפונים ניידים ומשחקים. אלה הסיבות אשר מנחות קווי מחקר ופיתוח של מערכות עם פלט-קלט קולי.

קיימות שתי גישות עיקריות לפתרון פרדיגמת הסינתזה של אותות דיבור מטקסט. בגישה הראשונה, משתמשים במקטעים של אותות דיבור מוקלטים להרכבת אות דיבור על פי טקסט מסוים. במערכות כאלה שומרים מקטעים של אותות דיבור מוקלטים או ייצוג פרמטרי שלהם. המקטעים המוקלטים האלה יכולים להיות מילים שלמות, פונמות, (יחידות פונטיות בסיסיות של שפה), או ,אפילו חלקי פונמות. במחקר זה השתמשנו במערכת של סינתזת אות דיבור מטקסט עם יחידות בסיסיות של שליש פונמה. בגישה זאת, משרשרים מקטעים בהתאם לכללים מסוים אשר מתבססים על תכונות של מקטעים סמוכים במשפט. בהמשך, נתייחס למערכות כאלה בתור מערכות משרשרות (concatenative).

האיכות של מערכות משרשרות היא דומה לאיכות הטבעית של מקטעים הנשמרים במאגר של המערכת. האיכות של אות דיבור מסונתז במערכות משרשרות תלויה במספר של המקטעים במאגר. ככל שגודל המאגר גדול יותר, כך גם האיכות של אות הדיבור המסונתז גבוהה יותר. ככל שהאורך של היחידות הבסיסיות קצר יותר, כך גם האיכות של אות הדיבור המסונתז יותר גבוהה, כי יש יותר אפשרויות לחיבור של מקטעים סמוכים. כך למשל, האיכות של אות הדיבור המסונתז במערכות משרשרות, אשר משתמשות בשלישי פונמות, היא יותר גבוהה מאשר האיכות של אות הדיבור המסונתז במערכות משרשרות אשר משתמשות בפונמות שלמות.

אם המספר של המקטעים אינו מספיק גדול, אזי האיכות של אותות הדיבור המסונתז נפגעת. לכן, כדי לסנתז אות דיבור מטקסט נתון באיכות גבוהה, נדרש מספר גדול מאוד של מקטעים במערכת. הדבר גורם לגודל הכולל של מערכות כאלה להיות גדול מאוד אם נדרשת איכות גבוהה של אותות דיבור מסונתז.

החיסרון העיקרי של מערכות משרשרות, הינו קפיצות בין מקטעים סמוכים, אשר מופיעות בין כל שני מקטעים שלא מתחברים בצורה רציפה. הקפיצות (אי-רציפויות) מופיעות כאשר מחברים שני מקטעים אשר אמנם מתאימים ביותר בין כל המקטעים במאגר, אולם בכל זאת, יש ביניהם מרחק ניכר. הקפיצות האלה מורגשות באות הדיבור המסונתז ומפריעות למאזין. לכן, מערכות משרשרות עם מאגר מוגבל של מקטעים, מסנתזות אותות דיבור באיכות נמוכה יותר מאשר מערכות משרשרות עם מאגר יותר גדול של מקטעים.

בגישה אחרת, לא שומרים מקטעים מוקלטים, אלא לומדים תכונות של מקטעים השייכים לפונמות שונות וממדלים אותם בעזרת מודלים סטטיסטיים. גישה זאת נקראת גישה סטטיסטית לסינתזת אותות דיבור מטקסט. מערכות כאלה צריכות הרבה פחות זכרון ואין באותות הדיבור המסונתזים על ידיהן את הקפיצות הקיימות במערכות משרשרות. פרמטרי אותות דיבור במערכות סטטיסטיות עוברים באופן חלק ממודל למודל ואין קפיצות באות הדיבור המסונתז. אולם, לעתים קרובות, פרמטרי אותות סטטיסטיים הינם מוחלקים יתר על המידה. החלקת היתר במערכות כאלה גורמת לאותות דיבור להישמע עמומים ולא טבעיים. החלקת היתר הינה החיסרון העיקרי של מערכות סטטיסטיות לסינתזת אותות דיבור.

במחקר זה אנו מציעים מספר דרכים לשיפור מערכות סטטיסטיות לסינתזת אותות דיבור מטקסט. הדרכים המוצעות הן: א) ייצוג אלטרנטיבי למודלים סטטיסטיים; אנו מגדירים ייצוג סטטיסטי לשלישי פונמות שלמות, להבדיל מייצוג הסטטיסטי למסגרות בודדות בייצוג סטנדרטי, ב) אילוץ הנורמה של ווקטור פרמטרי אות הדיבור המסונתז כך שתתאים לנורמה של ווקטור פרמטרי אות דיבור טבעי. אנו מציעים אלגוריתם למציאת פרמטרי אותות הדיבור הסינטטי אשר מתבסס על עקרונות אלה. האלגוריתם המוצע יוצר פרמטרי אותות דיבור עם דינמיקה מוגברת. השיפור בדינמיקה מביא לשיפור כללי באיכות של אותות הדיבור המסונתזים בהשוואה לאיכות של אותות הדיבור המסונתזים בשיטה הסטטיסטית הקונבנציונלית. השיפור אומת בבדיקות סובייקטיביות, (MOS-Mean Opinion Score). אנו קוראים לייצוג האלטרנטיבי SW-STTS (Segment-wise STTS).

בנוסף לשיטת הייצוג האלטרנטיבי אנו מציעים שיטה נוספת לשיפור האיכות של אותות הדיבור בגישה הסטטיסטית. בשיטה המוצעת מתבצע שיפור הדינמיקה של פרמטרי אותות הדיבור במישור התדר. שיטה זאת מביאה לשיפור של אותות הדיבור המסונתזים בהשוואה לסינתזה בשיטה הסטטיסטית הקונבנציונלית . שיטה זאת היא פחות טובה משיטת SW-STTS, מבחינת איכות אותות הדיבור המסונתזים, אך היא בעלת סיבוכיות נמוכה יותר.

מרכיב חשוב בעבודה היא ההצעה לשלב את היתרונות של מערכות משרשרות עם היתרונות של מערכות סטטיסטיות. המערכת המשולבת, המכונה כאן מערכת היברידית, שוזרת מקטעים טבעיים עם מקטעים סטטיסטיים בעזרת אלגוריתם דינמי שפותח בעבודה ומוסבר בגוף החיבור.

האלגוריתם הדינמי לשילוב מקטעים טבעיים עם מקטעים סטטיסטיים משלב מקטעים סטטיסטיים במקומות שבהם מקטעים טבעיים לא מתחברים בצורה רציפה. המקטעים הסטטיסטיים מסונתזים עם אילוצים בקצוות, כאשר האילוצים הינם המקטעים הטבעיים הסמוכים.

המרכיבים העיקריים של המערכת ההיברידית לסינתזת אותות דיבור הינם א) אלגוריתם היברידי דינמי, שמאפשר שילוב מקטעים טבעיים עם מקטעים סטטיסטיים, ב) אלגוריתם היוצר את ווקטור פרמטרי אותות הדיבור המסונתזים במערכת ההיברידית.

המערכת ההיברידית המוצעת שומרת על הטבעיות של מקטעים טבעיים המתקבלים מהמערכת המשרשרת ועל מעברים חלקים בין המקטעים. המערכת ההיברידית גם דורשת

גודל זיכרון נמוך יותר מהמערכת המשרשרת. כמו כן, היא בעלת סיבוכיות נמוכה יותר ממערכות סטטיסטיות מקובלות. בנוסף, המערכת ההיברידית המוצעת הינה הכללה של שני סוגי המערכות, סטטיסטית ומשרשרת, כי היא מסוגלת לפעול הן במצב של מערכת משרשרת טהורה והן במצב שלמערכת סטטיסטית טהורה, כתלות בפרמטר היברידיזציה, שקובע את הכמות של מקטעים סטטיסטיים מתוך סך המקטעים במשפט.

אותות הדיבור המסונתזים במערכת היברידית מורכבים ממקטעים טבעיים וממקטעים המיוצרים סטטיסטית. האיכות של משפט היברידי תלויה באיכות של המערכת הסטטיסטית. מצאנו, שהמערכת ההיברידית, שמשתמשת במערכת סטטיסטית SW-STTS, מסנתזת אותות דיבור באיכות גבוהה יותר מאשר המערכת ההיברידית אשר משתמשת במערכת סטטיסטית קונבנציונלית. הדבר מוכח במבחנים סובייקטיביים (MOS), שנעשו בעבודה.

מצד אחד, המערכת ההיברידית מסנתזת אותות דיבור עם פחות אי-רציפויות מאשר במערכת עם מאגר סגמנטים טבעיים. מצד שני, אותות דיבור היברידיים הם יותר טבעיים בהשוואה לאותות דיבור סטטיסטיים, כפי שהוכח גם בבדיקות סוביקטיביות.

בעבודה אנו מגיעים למסקנה, שהמערכת ההיברידית המוצעת עדיפה על פני מערכות סטטיסטיות מקובלות והינה עדיפה על פני מערכות משרשרות בעלות גדלים קטנים ובינוניים, מכיון שמערכות משרשרות בעלות גודל גדול מסנתזות אותות דיבור בלי אי-רציפויות המפריעות למאזין.