# Statistical Text-to-Speech Synthesis with Improved Dynamics

*Stas Tiomkin*[1,2]*, David Malah*[1]

[1]Department of Electrical Engineering, Technion-I.I.T,
Israel Institute of Technology, Haifa 32000, Israel.
[2]Speech Technologies Group, Haifa Research Lab, IBM.
stast@tx.technion.ac.il, malah@ee.technion.ac.il

## Abstract

In statistical TTS systems (STTS), speech features 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 synthesized speech. To improve feature dynamics a Global Variance approach has been suggested. However, it is computationally complex. We propose a different approach for modeling feature dynamics based on applying the DFT to the whole set of feature frames representing a phoneme. In the transform domain the inter-frame feature dynamics is then expressed in terms of inter-harmonic content, which can be modified to statistically match the dynamics of natural speech. To synthesize a whole utterance we propose a method for smoothly combining the enhanced-dynamics phonemes, which improves synthesized speech quality of STTS with similar complexity to conventional STTS.

**Index Terms**: text-to-speech synthesis (TTS), statistical speech modeling, speech features dynamics, global variance.

## 1. Introduction

Concatenative (sample-based) synthesis and statistical synthesis are the two main approaches to text-to-speech synthesis. A concatenative TTS system [1], [2]. directly uses natural segments, selected from a recorded speech database. Consequently, concatenative TTS (CTTS) systems enable speech synthesis with natural quality. However, when trying to reduce the footprint of the system, segments that match the required characteristics are not always available, so that other segments having closer characteristics are used instead, resulting sometimes in audible discontinuities. Consequently, the smaller the footprint size of the CTTS system is, the lower is the quality of generated speech that is achieved.

On the other hand, a statistical TTS (STTS) system, although having a smaller footprint, generates speech that is free of such discontinuities, but in general, is of lower quality than CTTS, in terms of naturalness [4], [5], and often sounds muffled and buzzy. To improve speech features dynamics, the Global Variance (GV) approach was proposed in [3], by which a penalty is introduced for decreased variance of speech features. The Global Variance approach improves naturalness but involves a computationally complex iterative procedure.

In this paper we propose an alternative technique to GV for alleviating the over-smoothing effect in speech generated by a STTS system, but with a much lower computational complexity. 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. 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.

The paper is organized as follows: In Section 2 we briefly describe the speech representation scheme used in our research, according to [6]. In Section 3 we provide a short description of the conventional statistical speech generation algorithm, detailed in [5]. In Section 4 we demonstrate the proposed approach to modeling speech features dynamics in the transform domain. In Section 5 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 6 we provide preliminary experimental results, and, finally, we summarize the paper in Section 7.

## 2. Speech Representation

In this research, as in [6], each speech frame is represented by a complex spectral envelope, expressed in polar form as: $S(f) = A(f)\exp^{j\varphi(f)}$, where $f$ denotes frequency. The spectrum amplitude $A(f)$ is analyzed on a logarithmic scale and modeled by a linear combination of basis function: $log(A(f)) = \sum_{n=1}^{L} c_n B_n(f)$, where $B_n$ are functions of $f$ in a mel-frequency scale. Statistical models are based on the coefficients (*"bins"*) $c_n$, $n = 1, \ldots, 32$. Examining the appropriateness of these coefficients to statistical text-to speech modeling was one of the first aims of our research on STTS. This representation is successfully used in IBM's state-of-the-art CTTS system having a reduced footprint [6]. In this work the prosody and context analysis of a synthesized utterance is done by means of the front end of IBM's TTS system [1], [2].

## 3. Conventional Statistical Speech Generation Algorithm

We briefly describe here the conventional approach for deriving the entire utterance feature vector in statistical TTS as detailed in [6]. An entire utterance over *N* frames is represented

by:$\underline{c}=[\underline{c}_1^T,\underline{c}_2^T,\ldots,\underline{c}_N^T]^T$, where $\underline{c}_i=(c_i(1),c_i(2),\ldots,c_i(d))^T$ is the static feature vector of dimension $d$ of the *i-th* frame. The static and the dynamic features are combined into a vector $\underline{o}=[\underline{o}_1^T,\underline{o}_2^T,\ldots,\underline{o}_N^T]^T$, where $\underline{o}_i=(\underline{c}_i,\underline{\triangle}_i,\underline{\triangle}_i^2)$, $\underline{\triangle}_i=\frac{1}{2}\left(\underline{c}_{i+1}-\underline{c}_{i-1}\right)$, and $\underline{\triangle}_i^2=\left(-\underline{c}_{i-1}+2\underline{c}_i-\underline{c}_{i+1}\right)$. Consequently, the vector $\underline{o}$, over an entire utterance, can be obtained from $\underline{c}$ by a linear transformation:

$$\underline{o}_{3d\cdot N\times 1}=W_{3d\cdot N\times d\cdot N}\underline{c}_{d\cdot N\times 1}, \quad (1)$$

where the matrix $W$ is constructed according to the first and $2^{nd}$ difference vectors $\underline{\triangle}$ and $\underline{\triangle}^2$. A detailed description of the incorporation of dynamic features into statistical speech synthesis is described in [3], [4], [5]. Every phoneme $p_i$, $i=1,2,\ldots,L$, included in a given utterance is modeled by a Gaussian mixture $GM_{p_i}(\underline{o})\sim\sum_i\omega_i\mathcal{N}(\underline{o};\underline{m}_{p_i},U_{p_i})$, where $\underline{m}_{p_i}$ and $U_{p_i}$ are the $p_i$-*th* model mean and covariance matrices of dimension $3d\times 1$ and $3d\times 3d$ , respectively. To find the optimal vector $\underline{c}$, over an entire utterance, the following cost function (into which (1) is substituted for $\underline{o}$) is maximized with respect to $\underline{c}$ (in our work, as in [3], we used just a single Gaussian with a diagonal covariance matrix, instead of a Gaussian mixture with full covariance matrix):

$$ln(P(\underline{o}))=ln(P(W\underline{c}))=\tfrac{1}{2}(W\underline{c}-\underline{m})^TU^{-1}(W\underline{c}-\underline{m}), \quad (2)$$

where $\underline{m}=[\underline{m}_{p_1,T_1}^T,\underline{m}_{p_2,T_2}^T,\ldots,\underline{m}_{p_L,T_L}^T]_{d\cdot N\times 1}^T$, $U=diag[U_{p_1,T_1},U_{p_2,T_2},\ldots,U_{p_L,T_L}]_{d\cdot N\times d\cdot N}$ are the utterance model mean vector and covariance matrix, respectively. The sub-indexes $p_i,T_i$ denote here that the model for $p_i$ is replicated $T_i$ times, which is the length of $p_i$, dictated by the phonetic-analyzer at the front end. Consequently, the optimal vector is given by $\underline{c}^{opt}=argmin_{\underline{c}}ln(P(W\underline{c}))=(W^TU^{-1}W)^{-1}W^TU^{-1}\underline{m}$, as detailed in [3]. We can see in the top plot in Fig.1 that the optimal solution is over-smoothed and has much less dynamics (inter-frame variations) as compared to the natural segment. Thus $\underline{\triangle}^{1,2}$ do not appear to sufficiently model the features dynamics, as also indicated by listening.

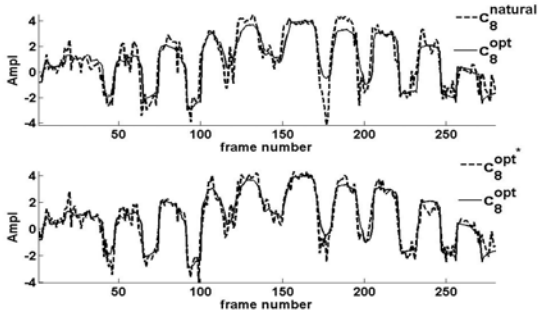


Figure 1: *Demonstrating features over-smoothing. Top plot: variation in time of $c_8^{natural}$ (dashed line) and of $c_8^{opt}$ (solid line). Bottom plot: variation in time of $c_8^{opt^*}$, as defined in Section 5.4, (dashed line) and of $c_8^{opt}$ (solid line).*

# 4. Modeling Feature Dynamics in the Transform Domain

## 4.1. 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.2. 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.3, 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. 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.

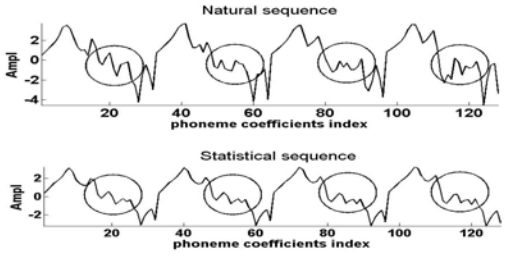


Figure 2: *Features frames of a natural sequence, $T_i=4$, $d=32$, (4 frames on the same plot) at the top; statistical sequence at the bottom. The frame-to-frame variations between circled regions demonstrate the low dynamics in the statistical sequence as compared to the natural sequence.*

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 $\underline{\triangle}^{1,2}$.

## 4.2. Improving speech features dynamics

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, to match the inter-harmonic content at the synthesis stage to the acquired statistics, as described below.

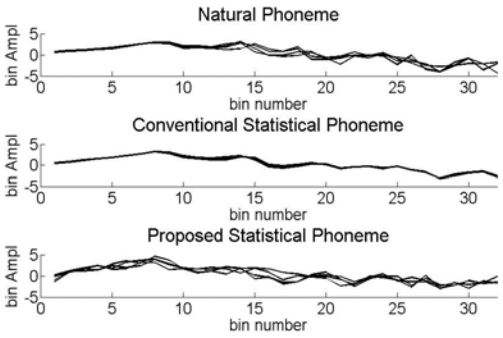


Figure 3: *Features frames of natural phoneme (4 frames on the same plot) at the top; Conventional statistically generated phoneme in the middle; Proposed statistically generated phoneme at the bottom.*

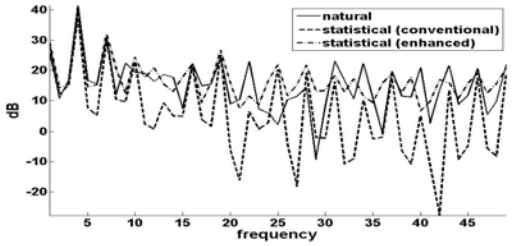


Figure 4: *Magnitude of transformed natural sequence (3 frames) in thin solid line; Magnitude of transformed conventional statistical sequence in dashed line; Magnitude of transformed statistical sequence with enhanced inter-harmonic content in dot-dashed line. Because of symmetry of the magnitude sequence, only the 48 (=$32 \cdot 3/2$) positive frequencies are shown.*

#### *4.2.1. 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 a corresponding length, being equal to a particular segment length $dT_i$, to the sequence of coefficients of every one of these natural segments. The transformed sequences have $d$ harmonic components and $d(T_i - 1)$ 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 [4].

#### *4.2.2. 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, 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 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 the required static features and enhanced inter-frame dynamics, as seen in the bottom plot of Fig.3.

## 5. Utterance-Level Synthesis

### 5.1. 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 (Section 2). 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 $W$ of (1).

### 5.2. 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, 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 (1):

$$\widehat{W}=(\omega^1;\beta^1;\omega^2;\beta^2\ldots;\beta^{i-1};\omega^i;\ \beta^{i+1}\ldots;\beta^{L-1};\omega^L), \quad (3)$$
$$(;\ denotes\ vertical\ concatenation)$$

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 $\beta^i$, shown in (4), smoothes the transitions between $p_{i-1}$ and $p_i$ by applying $\underline{\triangle}^{1,2}$:

$$\beta^i = \begin{bmatrix} \mathbf{0}_\rho & -\frac{\mathbf{1}}{\mathbf{2}}\xi & \mathbf{0}_\xi & +\frac{\mathbf{1}}{\mathbf{2}}\xi & \mathbf{0}_\xi & \mathbf{0}_\eta \\ \mathbf{0}_\rho & -\mathbf{1}_\xi & \mathbf{2}_\xi & -\mathbf{1}_\xi & \mathbf{0}_\xi & \mathbf{0}_\eta \\ \mathbf{0}_\rho & \mathbf{0}_\xi & -\frac{\mathbf{1}}{\mathbf{2}}\xi & \mathbf{0}_\xi & +\frac{\mathbf{1}}{\mathbf{2}}\xi & \mathbf{0}_\eta \\ \mathbf{0}_\rho & \mathbf{0}_\xi & -\mathbf{1}_\xi & \mathbf{2}_\xi & -\mathbf{1}_\xi & \mathbf{0}_\eta \end{bmatrix}, \quad (4)$$

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.

### 5.3. Utterance-level optimal solution

In Sections 5.1 and 5.2 $\widehat{W}$ is derived to enable different treatments of intra-phoneme frames and inter-phoneme transitions. 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:

$$\underline{\widehat{m}} = [\underline{\widehat{m}}^{p_1}_{l_1}, \underline{\triangle^*}_q, \underline{\widehat{m}}^{p_2}_{l_3}, \underline{\triangle^*}_q, \ldots, \underline{\triangle^*}_q, \underline{\widehat{m}}^{p_L}_{l_L}]^T, \quad (5)$$
$$l_i = d \cdot T_i \times 1, \;\; q = 4 \cdot d \times 1;$$

$$\widehat{U} = diag[{}^s\widehat{U}^{p_1}_{\tilde{l}_1}, {}^{\triangle^1}\widehat{U}^{p_1}_{\tilde{q}}, {}^{\triangle^2}\widehat{U}^{p_1}_{\tilde{q}}, {}^{\triangle^1}\widehat{U}^{p_2}_{\tilde{q}}, {}^{\triangle^2}\widehat{U}^{p_2}_{\tilde{q}}, {}^s\widehat{U}^{p_2}_{\tilde{l}_2},$$
$${}^{\triangle^1}\widehat{U}^{p_2}_{\tilde{q}}, {}^{\triangle^2}\widehat{U}^{p_2}_{\tilde{q}}, {}^{\triangle^1}\widehat{U}^{p_3}_{\tilde{q}}, {}^{\triangle^2}\widehat{U}^{p_3}_{\tilde{q}}, \ldots, {}^{\triangle^1}\widehat{U}^{p_{L-1}}_{\tilde{q}},$$
$${}^{\triangle^2}\widehat{U}^{p_{L-1}}_{\tilde{q}}, {}^s\widehat{U}^{p_L}_{\tilde{l}_L}], \;\; \tilde{l}_i = d \cdot T_i \times d \cdot T_i, \;\; \tilde{q} = d \times d. \quad (6)$$

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 (3), (5) and (6) in (2), the optimal solution is $\underline{c}^{opt^*} = (\widehat{W}^T \widehat{U}^{-1} \widehat{W})^{-1} \widehat{W}^T \widehat{U}^{-1} \underline{\widehat{m}}$, where the intra-phoneme frames with enhanced dynamics are optimally combined with smoothed inter-phoneme transitions.

## 6. Experimental Results

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 $\Lambda$ 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.1 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 with a small footprint to the naturalness of CTTS having a bigger footprint.

Notwithstanding the improvement in the overall naturalness of generated speech, the proposed statistical enhancement of speech dynamics may seldom cause increased variations in the low-band components as well. This issue should be investigated further to improve more the generated speech quality.

## 7. Conclusions

In this paper 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 preliminary subjective tests results, without increasing much the computational complexity. The spectral representation coefficients $c_n$ used in this study have been found appropriate for STTS modeling, as in the case in IBM's CTTS [6].

## 8. Acknowledgements

This research is a part of a joint research project by the Signal and Image Processing Lab (SIPL), Technion–I.I.T, and IBM's Haifa Research Lab (HRL). The authors are grateful to HRL-IBM for permission to use their speech databases and TTS software. The authors would like to thank in particular Zvi Kons, Slava Shechtman, Ron Hoory, Ariel Sagi, and Alex Sorin for useful discussions and valuable comments.

## 9. References

[1] R.E. Donovan, " Trainable speech synthesis", Phd thesis, Cambrige.

[2] R.E. Donovan, and E.m.Eide, "The IBM Trainable Speech Synthesis System", Proc. ICSLP98, Sydney.

[3] Tomoki Toda, Keiichi Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis", In INTERSPEECH-2005, 2801-2804

[4] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, Tadashi Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, vol.3, pp.1315-1318, June 2000.

[5] T.Masuko, K.Tokuda, T.Kobayashi, S.Imai, "Speech synthesis using HMMs with dynamic features ", Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference

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