point of view, we showed that by using the new formulation, a perfect detection of the speech components is possible if the noise originates from a point interference source, which can never be achieved with the single-channel case. In the case of incoherent noise, a coherent summation of the noise-free speech components is performed to allow for better speech detection, especially of low speech energy components as compared to the single-channel approaches. The proposed method applies for the general situation where the observed microphone signals are mixtures of a desired speech plus noise signals. The latter can be composed of interferences and other types of undesired signals (e.g., white noise).

REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[2] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1327–1339, May 2007.

[3] M. Souden, J. Benesty, and S. Affes, "New insights into non-causal multichannel linear filtering for noise reduction," in *Proc. IEEE ICASSP*, 2009, pp. 141–144.

[4] S. Gannot and I. Cohen, "Adaptive beamforming and postfitering," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. New York: Springer-Verlag, 2007, ch. 47, pp. 945–978.

[5] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1053–1065, Mar. 2007.

[6] P. C. Loizou, *Speech Enhancement: Theory and Practice*. New York: CRC Press, 2007.

[7] D. Middleton and R. Esposito, "Simultaneous optimum detection and estimation of signals in noise," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 3, pp. 434–444, May 1968.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[9] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[10] I. Potamitis, "Estimation of speech presence probability in the field of microphone array," *IEEE Signal Process. Lett.*, vol. 11, no. 12, pp. 956–959, Dec. 2004.

[11] G. Reuven, S. Gannot, and I. Cohen, "Dual source transfer-function generalized sidelobe canceller," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 711–726, May 2008.

[12] A. V. D. Bos, "The multivariate complex normal distribution—A generalization," *IEEE Trans. Inf. Theory*, vol. 41, no. 2, pp. 537–539, Mar. 1995.

[13] IEEE Subcommittee, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AE-17, no. 3, pp. 225–246, Sep. 1969.

[14] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, Apr. 1979.

[15] P. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, vol. 80, pp. 1527–1529, Nov. 1986.

[16] A. P. Varga, H. J. M. Steenekan, M. Tomlinson, and D. Jones, "The Noisex-92 study on the effect of additive noise on automatic speech recognition," *Tech. Rep., DRA Speech Research Unit*, 1992.

[17] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing.*. Berlin, Germany: Springer-Verlag, 2006.

Statistical Text-to-Speech Synthesis Based on Segment-Wise Representation With a Norm Constraint

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

***Abstract*—In statistical HMM-based text-to-speech systems (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 synthesized speech. In this correspondence, we propose a method to enhance a baseline STTS system by introducing a segment-wise model representation with a norm constraint. The segment-wise representation provides additional degrees of freedom in speech feature determination. We exploit these degrees of freedom for increasing the speech feature vector norm to match a norm constraint. As a result, statistically generated speech features are less over-smoothed, resulting in more natural sounding speech, as judged by listening tests.**

***Index Terms*—Segment-wise model representation, speech feature dynamics, statistical TTS, text-to-speech (TTS) synthesis.**

I. INTRODUCTION

Statistical TTS (STTS) systems employ statistical models for speech production, and speech is generated from previously learned statistical models. Contrary to concatenative TTS (CTTS), which may include discontinuities, particulary when small databases are used, STTS smoothly connects adjacent phonetic units.

However, STTS-generated speech is often over-smoothed, resulting in degraded speech quality in the form of muffled speech. A thorough review of STTS systems is provided in [1].

In this correspondence, we improve a baseline HMM-based STTS system by introducing 1) A robust model representation, based on a segment-wise representation, instead of the conventional frame-wise representation; and 2) A norm-regulated statistical speech feature vector that meets a norm constraint. These concepts are utilized in an iterative algorithm, proposed in this correspondence. 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.

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

Manuscript received August 25, 2009; revised December 14, 2009. First published January 19, 2010; current version published June 16, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gaël Richard.

S. Tiomkin and D. Malah are with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: stast@tx.technion.ac.il; malah@ee.technion.ac.il).

S. Shechtman is with the Speech Technologies Group, IBM Haifa Research Lab, Haifa University Campus, Mount Carmel, Haifa 31905, Israel (e-mail: slava@il.ibm.com).

Digital Object Identifier 10.1109/TASL.2010.2040795

II. HMM-Based Text-to-Speech Synthesis

A. Speech Feature Representation

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

$$\mathbf{c} = \left[\mathbf{c}_1^T, \mathbf{c}_2^T, \ldots, \mathbf{c}_N^T\right]^T \tag{1}$$

where $\mathbf{c}_i = (c_i(1), c_i(2), \ldots, c_i(M))^T$ are the expansion coefficients of the speech log-spectral amplitude, $log(\mathbf{A}(f'))$, by triangular basis functions, $\mathbf{B}_n(f')$.[1] This representation is successfully used in IBM's state-of-the-art CTTS system, detailed in [2]. A corresponding speech reconstruction unit is detailed in [3]. $\mathbf{c}_i$ denotes the static feature vector of dimension $M \times 1$ of the ith frame, where $M = 32$. In this research, we used frames of the length of 20 ms with a frame overlap of 10 ms. The prosody, (pitch, energy, and duration), is modeled by a context-dependent regression tree, detailed in [4], and [5].

The static speech features along with their first and second differences between frames, denoted dynamic features, constitute an augmented speech feature space, which is the conventional space for speech modeling. The static and dynamic features are combined into a vector $\mathbf{o}$

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

where $\mathbf{o}_i = (\mathbf{c}_i^T, \Delta^1\mathbf{c}_i^T, \Delta^2\mathbf{c}_i^T)^T$, with $\mathbf{\Delta}^1\mathbf{c}_i^T = 0.5(\mathbf{c}_{i+1} - \mathbf{c}_{i-1})$, and $\mathbf{\Delta}^2\mathbf{c}_i^T = 2\mathbf{c}_i - (\mathbf{c}_{i+1} + \mathbf{c}_{i-1})$, as detailed in [6], [7], and [8].

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

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

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

B. Statistical Model

Given a continuous mixture HMM, η, the optimal observation vector $\mathbf{o}$ over an entire utterance is derived by [9]

$$\mathbf{o}^{\text{opt}} = \underset{\mathbf{o}}{argmax}\, P(\mathbf{o}|\eta) \tag{4}$$

where $P(\mathbf{o}|\eta) = \sum_{\forall \mathbf{q}} P(\mathbf{o}, \mathbf{q}|\eta)$, and $\mathbf{q} = (q_1, q_2, \ldots, q_N)$ is the state sequence.

As mentioned in Section II-A, the prosody is modeled by a context-dependent regression tree, which provides the phonetic identities of states and their durations. Hence, we can reduce the general problem of solving (4) to the following problem, in which it is assumed that the state sequence $\mathbf{q}$ is given: $\mathbf{o}^{\text{opt}} = \underset{\mathbf{o}}{argmax}\ P(\mathbf{o}|\mathbf{q}, \eta)$.

In this correspondence, as in others described in a review on TTS systems [1], we use a single Gaussian model with a diagonal covariance matrix. Under such assumptions, the logarithm of $P(\mathbf{o}|\mathbf{q}, \eta)$ is

$$\begin{aligned} ln\left(P(\mathbf{o}|\mathbf{q}, \eta)\right) &= \frac{1}{2}(\mathbf{o} - \mathbf{m})^T\mathbf{U}^{-1}(\mathbf{o} - \mathbf{m}) \\ &= \frac{1}{2}\left\|\mathbf{U}^{-\frac{1}{2}}(\mathbf{o} - \mathbf{m})\right\|_2^2 \end{aligned} \tag{5}$$

with

$$\mathbf{m}_{3MN\times 1} = \left[\mathbf{m}_{q_1}^T, \mathbf{m}_{q_2}^T, \ldots, \mathbf{m}_{q_N}^T\right]^T \tag{6}$$

[1] $log(\mathbf{A}(f')) = \sum_{n=1}^{M} c_n \cdot \mathbf{B}_n(f')$, where f' denotes a mel-scale frequency.

and

$$\mathbf{U}^{-1}_{3MN\times 3MN} = \text{diag}\left[\mathbf{U}_{q_1}^{-1}, \mathbf{U}_{q_2}^{-1}, \ldots, \mathbf{U}_{q_N}^{-1}\right] \tag{7}$$

where the state q_t has duration d_t frames, its mean vector $\mathbf{m}_{\mathbf{q_t}}$ and its inverse covariance matrix $\mathbf{U}_{q_t}^{-1}$ are replicated d_t times within $\mathbf{m}_{3MN\times 1}$ and $\mathbf{U}^{-1}_{3MN\times 3MN}$, respectively. This aspect of the conventional representation will be considered in Section III.

Taking into consideration the relation between the static and dynamic features, defined by (3), the cost function over an entire utterance is

$$\begin{aligned} J(\mathbf{Wc}) &= -ln\left(P(\mathbf{Wc}|\mathbf{q}, \eta)\right) \\ &= \frac{1}{2}\left\|\mathbf{U}^{-\frac{1}{2}}(\mathbf{Wc} - \mathbf{m})\right\|_2^2. \end{aligned} \tag{8}$$

To find the optimal solution $\mathbf{c}^{\text{opt}}$ over an entire utterance, we set the first derivative of $J(\mathbf{Wc})$ with respect to $\mathbf{c}$ to 0. Consequently, we get $\mathbf{W}^T\mathbf{U}^{-1}\mathbf{Wc} = \mathbf{W}^T\mathbf{U}^{-1}\mathbf{m}$, and the optimal solution $\mathbf{c}^{\text{opt}}$ is given by

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

We can see in Fig. 1(a) that, typically, the optimal solution (9) is over-smoothed and has much less dynamics (inter-frame variations), as compared to the corresponding natural speech features. The natural eighth 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 sound, as indicated by listening, and as also reported in [8].

Fig. 1(b) provides zooming into the word "Many," partitioned into the marked HMM-states, $M_1, M_2, \ldots, IY_3$, having duration in frames of $d_{M_1} = 2$, $d_{M_2} = 3$, $d_{M_3} = 2$, $d_{EH_1} = 3$, $d_{EH_2} = 3$, $d_{EH_3} = 2$, $d_{N_1} = 1$, $d_{N_2} = 2$, $d_{N_3} = 1$, $d_{IY_1} = 1$, $d_{IY_2} = 3$, and $d_{IY_3} = 3$, respectively. The state means (solid gray line) are replicated according to the state durations; e.g., the state "M_3" lasts two frames. Thus, $\mathbf{\Delta}^{1,2}\mathbf{c}_i$ do not appear to fully capture the features dynamics, as also indicated by listening. We conclude from Fig. 1(a) and (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 the following section, we introduce a new concept of segment-wise model representation, which is found to improve the naturalness of generated speech.

III. Segment-Wise Model Representation

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}, respectively. In the conventional approach the respective augmented space speech feature frames $\mathbf{o}_{n_t}, \ldots, \mathbf{o}_{n_t+d_t-1}$ and $\mathbf{o}_{n_t+d_t}, \ldots, \mathbf{o}_{n_t+d_t+d_{t+1}-1}$ 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, $\mathbf{c}_{n_t}, \ldots, \mathbf{c}_{n_t+d_t-1}$, approximate the same static feature model mean, and at the same time, the corresponding dynamic features, $\mathbf{\Delta}^{1,2}_{n_t}\mathbf{c}_{n_t}, \ldots, \mathbf{\Delta}^{1,2}_{n_t+d_t-1}\mathbf{c}_{n_t+d_t-1}$, 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

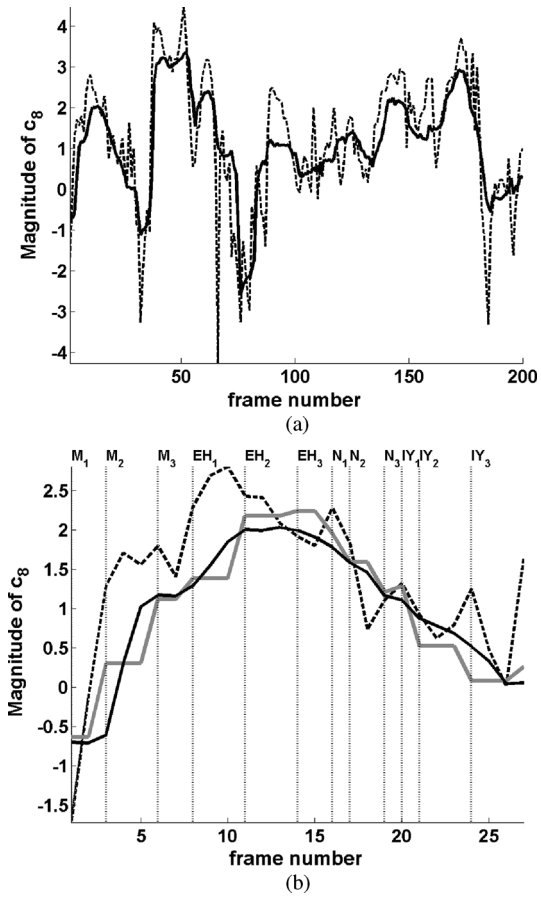

Fig. 1. Demonstrating conventional statistically generated speech feature over-smoothing in time, compared to a reference natural speech feature. (a) Variation in time of the eight expansion coefficient, c_8, in the utterance "**Many problems in reading and writing are due to old habits**": c_8^{opt} 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.

natural variances, represented by model covariance matrices, as seen in Fig. 1(b). The conventional model just connects smoothly adjacent models, involving a computationally complex matrix inversion, and redundant data storage required to store the statistics of $\Delta^{1,2}\mathbf{c}_{n_t}$, which do not have a sufficient effect, as depicted in this figure.

The above-mentioned conventional representation drawbacks often result in 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 $\widetilde{\mathbf{W}}$, defined in (12) and detailed in [10].

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, $\mathbf{o}_{n_t},\ldots,\mathbf{o}_{n_t+d_t-1}$, by the model mean of state q_t, as follows:

$$\bar{\mathbf{o}}_{n_t} = \frac{1}{d_t}\sum_{i=n_t-\left\lfloor\frac{d_t}{2}\right\rfloor}^{n_t+\left\lfloor\frac{d_t}{2}\right\rfloor}\mathbf{o}_i^T \tag{10}$$

and

$$J\left(\bar{\mathbf{o}}_{n_t}\right) = \frac{1}{2}\left\|\mathbf{U}_{q_t}^{-\frac{1}{2}}\left(\bar{\mathbf{o}}_{n_t}-\mathbf{m}_{q_t}\right)\right\|_2^2 \tag{11}$$

where $\bar{\mathbf{o}}_{n_t}$, $\mathbf{m}_{q_t}$, 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 $J(\bar{\mathbf{o}}_{n_t})$ is the corresponding cost function, constructed without replication of the model of state q_t.

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

$$\widetilde{\mathbf{W}}_{q_t} \triangleq \frac{1}{d_t}\begin{pmatrix} \mathbf{0} & \mathbf{1} & \cdots\underset{d_t-2}{\mathbf{1}}\cdots & \mathbf{1} & \mathbf{0} \\ -\frac{\mathbf{1}}{\mathbf{2}} & -\frac{\mathbf{1}}{\mathbf{2}} & \cdots\underset{d_t-2}{\mathbf{0}}\cdots & \frac{\mathbf{1}}{\mathbf{2}} & \frac{\mathbf{1}}{\mathbf{2}} \\ -\mathbf{1} & \mathbf{1} & \cdots\underset{d_t-2}{\mathbf{0}}\cdots & \mathbf{1} & -\mathbf{1} \end{pmatrix}_{3M\times M(d_t+2)} . \tag{12}$$

All the matrix elements in (12) are diagonal block matrices of dimension $M\times M$ each.

The segment-wise cost function $J^{sw}(\widetilde{\mathbf{W}}\mathbf{c})$, (where $\widetilde{(\cdot)}$ denotes non replication of state models, but rather approximation on average of state models), over an entire utterance is

$$J^{sw}(\widetilde{\mathbf{W}}\mathbf{c}) = \frac{1}{2}\left\|\widetilde{\mathbf{U}}^{-\frac{1}{2}}(\widetilde{\mathbf{W}}\mathbf{c}-\widetilde{\mathbf{m}})\right\|_2^2 \tag{13}$$

where $\widetilde{\mathbf{m}}_{3MK\times 1} = [\mathbf{m}_{q_1}^T,\mathbf{m}_{q_2}^T,\ldots,\mathbf{m}_{q_K}^T]^T$, $\widetilde{\mathbf{U}}^{-1}_{3MK\times 3MK} = \mathrm{diag}[\mathbf{U}_{q_1}^{-1},\mathbf{U}_{q_2}^{-1},\ldots,\mathbf{U}_{q_K}^{-1}]$, and $\widetilde{\mathbf{W}} = \mathrm{diag}[\widetilde{\mathbf{W}}_{q_1},\widetilde{\mathbf{W}}_{q_2},\ldots,\widetilde{\mathbf{W}}_{q_K}]$ are the non-replicated model mean vector, the non-replicated model covariance matrix, and the segment-wise transformation, respectively. Here, K is the total number of states in a synthesized utterance. $\widetilde{\mathbf{m}}_{3MK\times 1}$ and $\widetilde{\mathbf{U}}^{-1}_{3MK\times 3MK}$ 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 covariance matrix, $\mathbf{U}^{-1}_{3MN\times 3MN}$, defined in (6) and (7), respectively, which contain replicated terms (note the different dimensions). Using the proposed segment-wise representation, the model is less restricted and enables more dynamics in generated speech features, as compared to the conventional model.

Consequently, the conventional frame-wise cost function in (8) 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 (13) is derived as in (9)

$$\mathbf{c}^{opt,sw} = (\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{W}})^{-1}\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{m}}. \tag{14}$$

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 $\mathbf{c}^{\mathrm{opt},sw}$ for states having duration more than one frame. In such a case, the number of equations is smaller than the number variables, so, 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 the following section.

IV. Norm Constraint

We have observed[2] that the squared-norm of statistically generated speech feature vectors of entire utterances, $\|\mathbf{c}^{stt}\|_2^2$, is often quite lower than the squared-norm of natural speech feature vectors of entire utterances, $\|\mathbf{c}^{\mathrm{nat}}\|_2^2$. This is because, first, the conventional solution, shown in (9), is the minimal norm least squares solution, and, second, due to

[2]A set of 40 arbitrary sentences was generated, whose speech feature vector norms were examined, and compared to 1) corresponding speech feature model mean vector norms, and 2) corresponding natural speech feature vector norms.

insufficient speech feature dynamics, a statistically generated speech feature vector norm is quite close to the model means norm $\|\mathbf{c}^{mdl}\|_2^2$

$$\|\mathbf{c}^{stt}\|_2^2 \approx \|\mathbf{c}^{mdl}\|_2^2. \tag{15}$$

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 (8), we add a norm-dependent auxiliary term, constraining the speech feature vector norm, thus avoiding the norm reduction.

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 $\|\mathbf{c}^{\mathrm{nat}}\|_2^2$ by a factor $\gamma_0 > 1$

$$\gamma_0 = \frac{\|\widetilde{\mathbf{c}^{\mathrm{nat}}}\|_2^2}{\|\widetilde{\mathbf{c}^{stt}}\|_2^2} \tag{16}$$

denoted as the enhancement factor, where $\widetilde{\|\cdot\|}$ is an averaged norm over a set of utterances generated from a particular voice. Consequently, using (15), a constraint on the norm of speech features $\|\mathbf{c}^{stt}\|_2^2$ should be set to $\Gamma = \gamma_0 \cdot \|\mathbf{c}^{mdl}\|_2^2$, in order to compensate for the norm reduction, achieving in our case $\|\mathbf{c}^{stt,sw}\|_2^2 \approx \|\mathbf{c}^{\mathrm{nat}}\|_2^2$. In the following subsection, we provide a systematic approach for speech feature dynamics enhancement by applying this constraint.

A. Norm-Constrained Cost Function

Our goal is to find an optimal norm-constrained feature vector, $\mathbf{c}^{\mathrm{opt}}$, over an entire utterance, which minimizes the model error and possesses sufficient features dynamics.

For that end, we propose to regulate the solution by adding a squared-norm term of the feature vector to the model-error term of the cost function of (13), using a factor λ to balance the contribution of the two terms. Thus, the cost function of (13) is replaced by

$$J_c^{sw}(\widetilde{\mathbf{W}}\mathbf{c}) \triangleq \frac{1}{2}\left\|\mathbf{U}^{-\frac{1}{2}}(\widetilde{\mathbf{W}}\mathbf{c}-\mathbf{m})\right\|_2^2 + \frac{\lambda}{2}\|\mathbf{c}\|_2^2. \tag{17}$$

In the proposed method, the norm term provides a solution with enhanced dynamics, by using prior information on λ, as elaborated below. We propose an iterative algorithm that minimizes the model cost function value, while assuring sufficient dynamics in the resulting solution. The minimization is done by means of a gradient descent algorithm as follows:

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \alpha_n \nabla(\mathbf{c}_n) \tag{18}$$

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 $\alpha_n = (1/\|\nabla(\mathbf{c}_n)\|_2^2)$. From (17)

$$\nabla(\mathbf{c}_n) = \widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\widetilde{\mathbf{W}}\mathbf{c}_n - \widetilde{\mathbf{W}}^T\mathbf{U}^{-1}\mathbf{m} + \lambda\mathbf{c}_n. \tag{19}$$

A final feature vector should approximate well the models, and have a norm value that is compatible with the enhancement factor, defined in (16). 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. That is, we replace λ in (19) by λ_n, which is updated according to

$$\lambda_{n+1} = \theta\lambda_n, \quad 0 \le \theta < 1. \tag{20}$$

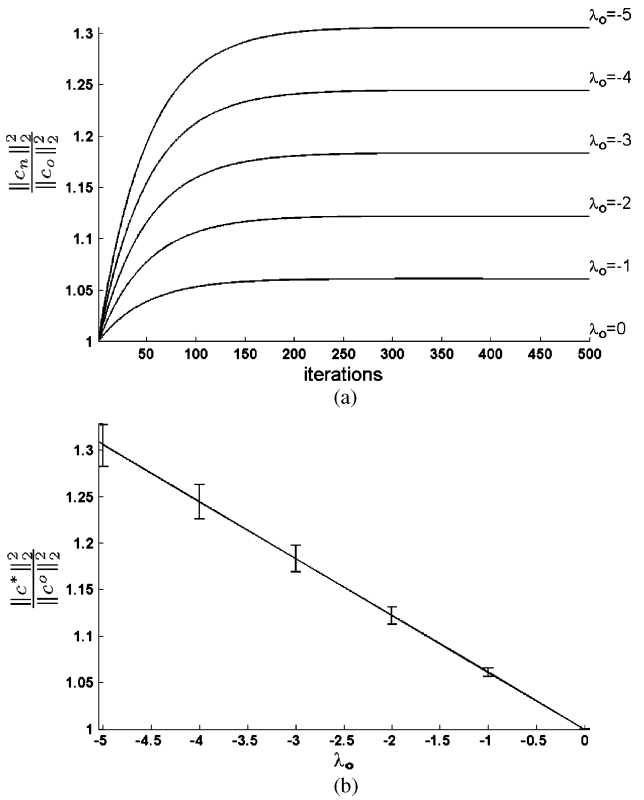

Fig. 2. Evolution of $\|c_n\|_2^2$ as function of λ_o. (a) An increase in a feature vector norm $\|c_n\|_2^2$ as a function of an initial value for λ_o, where $\|c_o\|_2^2$ is the norm of an initial vector. (b) Relation between λ_o and the final feature vector norm $\|c^*\|_2^2$. The error bars depict the standard deviations in $\|c^*\|_2^2$ for given values of λ_o.

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

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

An empirical relation between λ_0 and the ratio of the feature vector norm after n iterations, $\|c_n\|_2^2$, to the initial feature vector norm $\|c_0\|_2^2$ is represented in Fig. 2(a) and (b). This relation was obtained by averaging λ_0 over a large set of iteratively generated utterances. The standard deviations of the final speech feature vector norm, for given values of λ_0, are represented by the error bars in Fig. 2(b). For λ_0 equal to -5, which is consistent with the needed 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 the 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. Setting λ_0 according to the above mentioned empirical relation enables

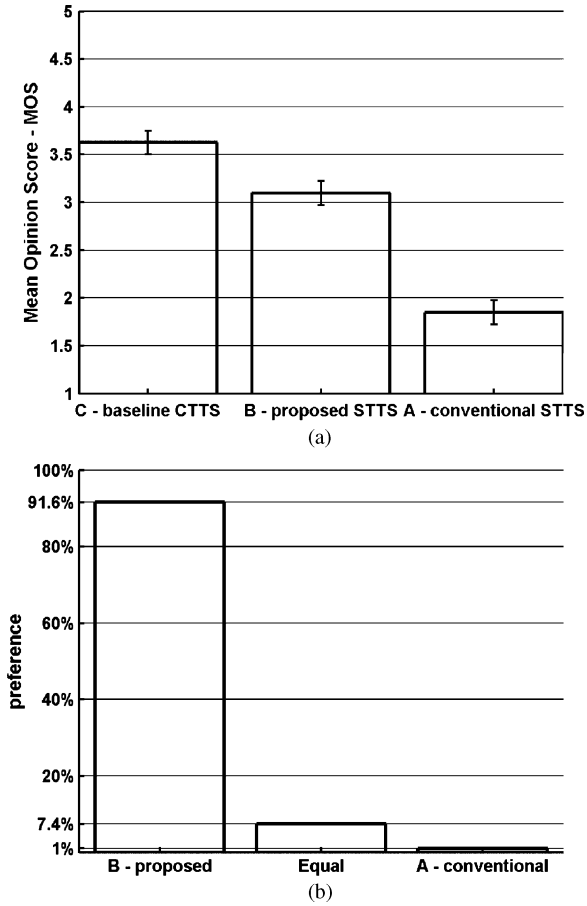

Fig. 3. Subjective evaluation. (a) Mean opinion score (MOS) test. (b) "A versus B comparison test."

an increase in the norm of $\mathbf{c}_n$ that is consistent with the norm enhancement factor introduced in (16), resulting in enhanced dynamics in generated speech, as confirmed by listening tests described in Section V.

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

V. Experimental Results

We have evaluated the proposed STTS algorithm by the two following subjective tests.

Test I: In this test, we have evaluated the mean opinion score (MOS) of a set of nine arbitrary sentences, where each sentence was generated in three versions: 1) conventional statistical speech generation algorithm, mentioned in Section II, (group A); 2) proposed speech generation scheme (group B), and 3) IBM's CTTS system, detailed in [2] and [5], (group C). Thus, 27 samples were included in the test, each of which was evaluated by 20 listeners. The same target prosody was used in the synthesis of all the tested versions of a particular sentence, so as to not be affected by different prosody targets in different systems.

Fig. 3(a) 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. The error bars indicate 95% confidence interval, computed using the "t-test"

Test II: In this test, a set of 11 arbitrary sentences was used, each of which was evaluated by 20 listeners. Each of the sentences was generated in two versions: 1) conventional statistical speech generation algorithm, mentioned in Section II, (group A); and 2) proposed speech generation scheme (group B). The two versions of each sentence were compared using an "A versus B" comparison 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. Again, same target prosody was used in the synthesis of all the tested versions of a particular sentence, so as to not be affected by different prosody targets in different systems.[3]

Comparing the speech quality generated by the proposed method to that of the global variance (GV) approach, detailed in [11] and [12], we got similar MOS-test results. The proposed method is advantageous in the following aspects: 1) It is more efficient in terms of memory storage since it avoids an increase in the memory footprint of about 30%, (each model requires $6M$ numbers for acoustic features, (M is speech feature vector dimension), and the GV requires to store additional $2M$ numbers for global variance). 2) The segment-wise linear transformation requires less real-time memory, than does the conventional frame-wise representation because the former and the later representations models dimensions are $\widetilde{\mathbf{W}}_{3MK\times MN}$, $\widetilde{\mathbf{m}}_{3MK\times 1}$, $\widetilde{\mathbf{U}}_{3MK\times 3MK}$, and $\mathbf{W}_{3MN\times MN}$, $\mathbf{m}_{3MN\times 1}$, $\mathbf{U}_{3MN\times 3MN}$, respectively, where N is the number of frames and K is the number of models (segments) in a synthesized utterance, and $N \geq K$. 3) The computational complexity of the proposed iterative method, as summarized in Section IV, is lower than GV [11].

VI. Summary

We proposed in this correspondence a method for the enhancement of statistically generated speech feature dynamics, which alleviates the over-smoothing of speech features, and as a result, improves the statistically generated speech quality. The proposed method is based on a segment-wise representation of the augmented space for speech features.

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

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

Acknowledgment

The authors would like to thank R. Hoory, the head of the Speech Technologies Group at IBM Haifa Research Lab (HRL), Z. Kons (HRL), A. Sagi (formerly with HRL), and A. Sorin for useful discussions in the course of the work. They would also like to thank the Signal and Image Processing Lab (SIPL) staff, N. Peleg, Z. Avni, A. Rosen, and Y. Moshe, for their technical support and the anonymous reviewers for their comments and suggestions that helped much to improve the correspondence. This research is part of a joint research project conducted at SIPL, Technion–Israel Institute of Technology, and HRL.

[3]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.

REFERENCES

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, pp. 1039–1064, 2009.

[2] D. Chazan, R. Hoory, Z. Kons, A. Sagi, S. Shechtman, and A. Sorin, "Small footprint concatenative text-to-speech synthesis using complex envelop modeling," in *Proc. Interspeech'05*, Lisbon, Portugal, Sep. 2005, pp. 2569–2572.

[3] D. Chazan, R. Hoory, A. Sagi, S. Shechtman, A. Sorin, Z. Shuang, and R. Bakis, "High quality sinusoidal modeling of wideband speech for the purpose of speech synthesis and modification," in *Proc. ICASSSP'06*, Toulouse, France, May 2006, pp. 1303–1306.

[4] R. E. Donovan, "Trainable speech synthesis," Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., Jun. 1996.

[5] R. E. Donovan and E. M. Eide, "The IBM trainable speech synthesis system," in *Proc. ICSLP'98*, Sydney, Australia, vol. 5, pp. 1703–1706.

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

[7] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," in *Proc. ICASSP'86*, Tokyo, Japan, Apr. 1986, pp. 877–880.

[8] K. Tokuda, T. Yoshimura, T. Masuko, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP'00*, Istanbul, Turkey, pp. 1315–1318.

[9] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[10] S. Tiomkin, "A segment-wise hybrid approach for improved quality text-to-speech synthesis" M.Sc. thesis, Technion-Israel Inst. of Technol., Haifa, May 2009 [Online]. Available: http://sipl.technion.ac.il/siglib/FP/Stas_Tiomkin.pdf

[11] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in *Proc. Interspeech'05*, Lisbon, Portugal, Sep. 2005, pp. 2801–2804.

[12] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *Trans. IEICE*, vol. E90-D, pp. 816–824, May 2007.

Channel Robust Feature Transformation Based on Filter-Bank Energy Filtering

Claudio Garreton, Nestor Becerra Yoma, and Matias Torres

***Abstract*—This correspondence proposes a novel feature transform for channel robustness with short utterances. In contrast to well-known techniques based on feature trajectory filtering, the presented procedure aims to reduce the time-varying component of channel distortion by applying a bandpass filter along the Mel frequency domain on a frame-by-frame basis. By doing so, the channel cancelling effect due to conventional feature trajectory filtering methods is enhanced. The filtering parameters are defined by employing a novel version of relative importance analysis based on a discriminant function. Experiments with telephone speech on a text-dependent speaker verification task show that the proposed scheme can lead to reductions of 8.6% in equal error rate when compared with the baseline system. Also, when applied in combination with cepstral mean normalization and RASTA, the presented technique leads to further reductions of 9.7% and 4.3% in equal error rate, respectively, when compared with those methods isolated.**

***Index Terms*—Channel robustness, robust features, speaker recognition, text-dependent speaker verification (TD-SV).**

I. INTRODUCTION

Robustness to channel mismatch between training and testing conditions is one of the most important problems faced by speaker verification (SV), speech recognition (ASR), language recognition, and phonetic quality assessment systems in real applications. Also, due to operating and usability restrictions in telephone services, the amount of adaptation data to remove or reduce convolutional noise is limited. For instance, enrolling and verification in text-dependent speaker verification (TD-SV) systems over the telephone network should be fast and efficient.

The motivation of channel canceling or compensation techniques is to reach the error rate observed in channel matched conditions by minimizing the requirements of extra data. The approaches to tackle the problem of channel mismatch can be clustered into two main areas [1]: feature compensation [2]–[4]; and model adaptation [5], [6]. The most widely accepted model for channel distortion corresponds to a cepstral or log-spectral bias that results from the following hypotheses: H1, the channel response is signal independent; and H2, the channel can be modeled as a linear filter. The aim of current feature compensation methods is to estimate the original undistorted signal by removing a bias constant or a low-frequency component in the cepstral or log-spectral domain [2]–[4]. Usually, these approaches can dramatically reduce the error rate in channel mismatch condition but also show a significant efficacy lost with limited data. For instance, cepstral mean normalization (CMN) attempts to remove the bias component in the cepstral domain, but its effectiveness is reduced with short utterances [7]. Moreover, bias removal methods based on the expectation–maximization (EM) algorithm can also provide significant reductions in error rate [3]. Nevertheless, the EM algorithm is also sensitive to utterance

Manuscript received April 09, 2009; revised December 09, 2009. First published May 03, 2010; current version published June 16, 2010. This work was supported by Conicyt-Chile under Grants Fondef D05I-10243 and Fondecyt 1070382/1100195. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Haizhou Li.

The authors are with the Speech Processing and Transmission Laboratory, Department of Electrical Engineering, Universidad de Chile, Santiago 837-0451, Chile (e-mail: nbecerra@ing.uchile.cl).

Digital Object Identifier 10.1109/TASL.2010.2049671