Uniform Speech Parameterization for Multi-form Segment Synthesis

Alexander Sorin[1], *Slava Shechtman*[1], *Vincent Pollet*[2]

[1] Speech Technologies, IBM Haifa Research Lab, Haifa, Israel
[2] Text-To-Speech Research, Nuance Communications, Merelbeke, Belgium
sorin@il.ibm.com, slava@il.ibm.com, vincent.pollet@nuance.com

Abstract

In multi-form segment synthesis speech is constructed by sequencing speech segments of different nature: model segments, i.e. mathematical abstractions of speech and template segments, i.e. speech waveform fragments. These multi-form segments can have shared, layered or alternate speech parameterization schemes. This paper introduces an advanced uniform speech parameterization scheme for statistical model segments and waveform segments employed in our multi-form segment synthesis system. Mel-Regularized Cepstrum derived from amplitude and phase spectra forms its basic framework. Furthermore, a new adaptive enhancement technique for model segments is presented that reduces the perceived gap in quality and similarity between model and template segments.

Index Terms: speech synthesis, multi-form segment, speech parameterization, statistical speech synthesis, text to speech, mel-regularized cepstrum.

1. Introduction

Model-based and concatenative Text-To-Speech (TTS) systems differ in many aspects. In terms of principle, concatenative [1] systems select basic units in the form of waveform segments from a large speech inventory and stitch them together with a minimum of signal manipulation. Model-based [2] systems rely heavily on signal manipulation; speech waveforms are decomposed into speech parameters, modeled and reconstructed. In terms of speech quality, concatenative systems produce variable quality speech, which at best is highly natural. This variation in quality is dependent on the length of continuous speech segments selected from the unit inventory and the concatenation quality. Limited domain concatenative systems, which tend to return long stretches of stored speech, produce highly natural synthesis. A speech model as such is a simplified representation of the structure of natural speech. Hence model-based systems have more consistent speech quality but with a synthetic "processed" character. The generated speech is smooth, stable and the systems show good predictable behavior with respect to unseen contexts.

Multi-form Segment (MFS) [3] TTS combines the benefits of the model-based and the concatenative synthesis approaches. A statistical framework forms the basis for selecting and generating multi-form segment sequences. As a consequence it superposes automatic processing, consistency, auto-tuning and generalization with high segmental quality. Both natural speech fragments and models are employed in the speech construction process resulting in high quality and naturalness. Using the statistical framework, *template segments* (i.e. natural speech fragments) are selected and used as observations to optimize the generation of *model segments* so they can be seamlessly combined with the selected template segments. *Template segments* are used as final segments where the optimized model segments fail to represent natural speech adequately. The output sequence of best multi-form segments is a mixed sequence of template and model segments which is motivated by perceptual and statistical foundations. The output speech waveform is obtained by concatenating the waveforms resulting from synthesizing the acoustic parameter trajectories of the model segments with the waveforms of template segments.

In the MFS system the Model-Template Ratio (MTR) expresses how much of the output speech is generated from model segments on average. The probability of a multi-form segment being a model segment depends on three categories of cues: phonologic cues, acoustic cues and channel cues. In [3] it was observed that a MTR of 40% can produce highly natural speech. For achieving higher ratios, the system is predominately constrained by the channel cues: the quality of the model segments and the speech parameterization and reconstruction quality. Improving these aspects implicates an increased flexibility and quality of MFS synthesis.

This paper introduces a *uniform speech parameterization* scheme for the multi-form segment TTS system. The main motivation for adopting this parameterization scheme is that template segments can be 'spliced in' with higher flexibility which results in increased uniform speech quality in case of a higher MTR. In section 2, a speech parameterization scheme, based on a Mel-Regularized Cepstrum representation [4] for both template and model segments is introduced. Template segments have an additional layer for storing phase information. Section 3 presents model analysis insights and introduces a new *adaptive model segment enhancement* technique. The focus is to improve the modeled speech quality and to reduce the gap in voice character between the modeled and the template-based speech. In Section 4, the uniform parameterization and the proposed model enhancement technique are evaluated. MOS results for the "all model-segments" operating point and model-template similarity results are presented to emphasise the impact of this work. The results are compared against an earlier experimental MFS US English baseline. Finally in section 5, conclusions are drawn from this work.

2. Uniform speech parameterization

Mel Regularized Cepstral Coefficients (MRCC) parameterization of speech applied to the constant frame sinusoidal model [4] produced encouraging results when adopted for a statistical TTS system [5]. The MRCC vector $\mathbf{c} = \{c_k\}$ parameterizes the amplitude spectrum $A(\tilde{f})$ of a speech frame:

$$\log A(\tilde{f}) \approx c_0 + 2\sum_{k=1}^{K} c_k \cdot \cos(2\pi k \cdot \tilde{f}) \qquad (1)$$

where $\tilde{f}$ is the normalized Mel-scaled frequency. The MRCC vector is obtained by solving following quadratic minimization problem:

$$\arg\min_{\mathbf{c}} \left\{ \sum_{i=0}^{N} \left| \log A(\tilde{f}_i) - \log A_i \right|^2 + \lambda \int_0^{0.5} \left[\frac{d \log A(\tilde{f})}{d\tilde{f}} \right]^2 d\tilde{f} \right\} \quad (2)$$

where $\{\tilde{f}_i\}$ and $\{A_i\}$ are respectively the Mel-frequencies and amplitudes of the line spectrum components given by the sinusoidal model. The minimization criterion (2) establishes a balance between the approximation accuracy represented by the first term and the degree of spectral smoothness represented by the second term. A detailed description of speech reconstruction from the MRCC vectors can be found in [5].

In our MFS prototype, the MRCC parameterization is used as a uniform representation for both the model and template segments. To further improve the synthesized speech, in particular for template segments, a novel phase spectrum parameterization approach is introduced which logically complements the MRCC parameterization of the amplitude spectrum. The continuous phase spectrum $\Phi(\tilde{f})$, given by the imaginary part of the logarithm of the complex spectrum, can be expressed in cepstral domain as a sine series combined with a linear in frequency term. Replacing the infinite sine series by a finite sum the following approximation is obtained:

$$\Phi(\tilde{f}) \approx \alpha + \beta \cdot \tilde{f} - 2\sum_{k=1}^{K} d_k \cdot \sin(2\pi k \cdot \tilde{f}) \quad (3)$$

Where α is a constant phase offset equal to either *0* or π depending on the polarity of the time-domain waveform, β is a time offset of the waveform and $\mathbf{d}=\{d_k\}$ is the vector of the phase cepstral coefficients. Adopting the idea of the MRCC regularization, the vector $\mathbf{d}$ of phase parameters is then estimated as follows:

$$\mathbf{d} = \arg\min_{\mathbf{d},\alpha,\beta} \left\{ \sum_{i=0}^{N} \left| \Phi(\tilde{f}_i) - \phi_i \right|^2 A_i^{\mu} + \nu \int_0^{0.5} \left[\frac{d\Phi(\tilde{f})}{d\tilde{f}} \right]^2 d\tilde{f} \right\} \quad (4)$$

where $\{\phi_i\}$ are the phases of the line spectrum components available from the sinusoidal model and further unwrapped.

The first term of the minimization criterion (4) represents the distance between the parameterized phase spectrum and the line spectrum phases. The distance calculation incorporates weighting factors dependant on the respective line spectrum amplitudes. The second regularizing term imposes smoothness on the phase spectrum. The strength of the amplitude dependent weighting and the smoothness degree are controlled by parameters μ and ν respectively.

The minimization of the quadratic form (4) is achieved by solving a set of linear equations in $\mathbf{d}$, α and β. Thus the linear phase term is optimized together with the phase cepstral parameters $\mathbf{d}$. Next, the linear term is abandoned and only the vector $\mathbf{d}$ is stored. The components of the vector $\mathbf{d}$ are further referred to as Weighted Mel-Regularized Cepstral Coefficients (WMRCC). Important to note is that no constraints are imposed on the phase offset α. Not fixing α yields an extra degree of freedom which may contribute to a better approximation accuracy and smoothness of the resulting phase spectrum. This was inspired by informal listening tests which indicated that adding an arbitrary offset to all harmonic phases of all frames is practically inaudible. To reconstruct the waveform from the parameters, the optimal linear phase parameters α and β are computed by maximizing the cross-correlation between the current and previous frame. Both the parameters can be easily calculated from the analytical signal corresponding to the cross-correlation function as shown in [6]. An example of the phase parameterization is depicted in Figure 1. It shows that the resulting phase spectrum is smooth and the approximation of the sinusoidal phases gets better at lower frequencies and at higher amplitudes. The phase spectrum is represented within the 0-4 kHz frequency band. At reconstruction time the phase spectrum is extrapolated to full band using the method described in [6].

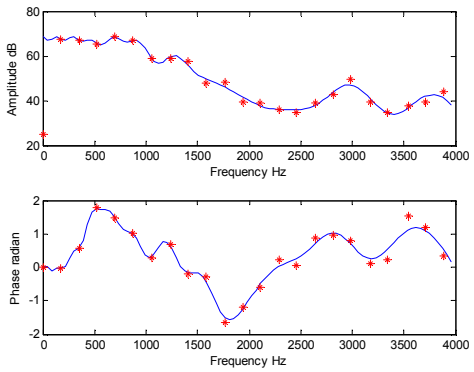

Figure 1: *WMRCC phase parameterization (bottom) and MRCC amplitude parameterization (top) based on the sinusoidal phases and amplitudes shown by asterisks.*

Table 1 summarizes the results of a subjective preference test performed with a goal to compare the quality of the WMRCC phase against a minimum-phase representation. 10 pairs of US English sentences (5 female and 5 male) recorded at 22 kHz were analyzed and reconstructed using the MRCC amplitude spectrum (of order 33) combined with either the minimum-phase or the WMRCC phase (of order 24, 30 for female, male respectively). Each pair was evaluated by 7 subjects. The results clearly indicate a significant preference ($p<0.001$) for the WMRCC phase representation.

Table 1. *A-B preference test: WMRCC phase vs. MRCC minimum-phase*

WMRCC	Equivalent	Min-phase
68.6%	25.7%	5.7%

Employing the WMRCC phase representation for the model segments did not indicate an advantageous effect. It was observed that the WMRCC phase representation was poorly modeled. Consequently in our MFS prototype, the WMRCC phase representation is only available to the template segments. In case of model segment reconstruction, minimum-phase derived from the amplitude MRCC is applied.

3. Analysis and improvement of model segments generation

The flexibility of the MFS system is constrained by the quality of the model segments. Further reducing the quality gap between model and template segments implicates the usage of more 'flexible' MFS schemes. Therefore quality enhancement of the MRCC-based model segments generation constitutes the primary focus of this work.

The over-smoothed model generated speech is accounted to spectral shape smearing as a result of Gaussian modeling of cepstral vectors at each HMM state. The smearing of the MRCC-parameterized spectra is depicted in Figure 2. The

MRCC vectors emitted from the Gaussian model (*model vectors*) show flatter spectra with lower peaks and higher valleys compared to the spectra of the original MRCC vectors (*raw vectors*) that were used for the model estimation. Figure 2 shows subsets of raw and model spectra for a certain HMM state of an US English female voice model. Graphically, the model spectrum lines (drawn in thin black) merge together to a single black bold line due to the fact that the variability in the model vectors is low compared to the one observed in the raw vectors.

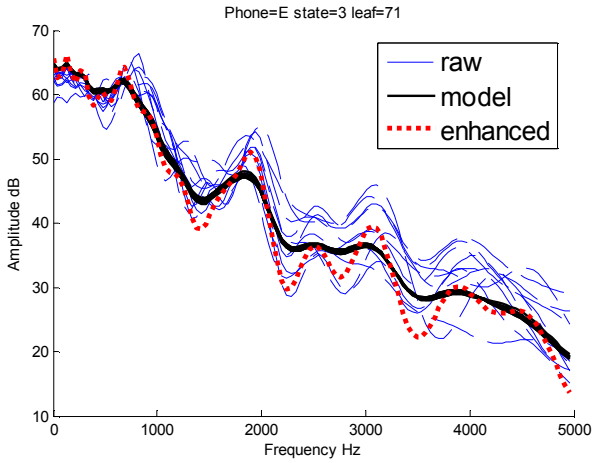

Figure 2*: Raw spectra, model spectra and enhanced model spectra associated with a selected HMM state*

The spectrum flattening is closely related to an increased attenuation of the cepstrum with *quefrency*. Insight of this relation can be gained using the rational representation of the vocal tract transfer function:

$$S(z) = \prod_{m=1}^{M}(1 - z^{-1}z_m) \Big/ \prod_{k}^{K}(1 - z^{-1}p_k),\ |p_k| < 1, |z_m| < 1 \qquad (5)$$

where $\{p_k\}$ and $\{z_m\}$ are respectively poles and zeros of $S(z)$. Taking the logarithm of the right-side of (5) and applying the Maclaurin series expansion to the additive logarithmic terms, the cepstrum of the vocal tract impulse response can be expressed as following:

$$\hat{s}(n) = \sum_{k=1}^{K}\frac{p_k^n}{n} - \sum_{m=1}^{M}\frac{z_m^n}{n}, n = 1, 2, \ldots \qquad (6)$$

From (6) follows that when the poles and zeros of the transfer function move away from the unit circle and towards the origin of Z-plane (flattening spectral peaks and valleys) the cepstrum attenuation increases. The MRCC parameterization is a finite length approximation of the cepstrum transform applied to the frequency-warped spectrum. Only the radial location of spectral peaks and valleys is indicative for the analysis insights. Hence the frequency warping can be ignored when applying the above consideration to the MRCC vectors. In other words, it's expected that statistically modeled MRCC vectors have higher attenuation in quefrency than raw MRCC vectors. This hypothesis is supported by the statistical observation which compares the L2-norm distribution over the MRCC components measured on raw and model vectors. For a given HMM state the *L2-norm ratio vector* **R** is defined as:

$$R(k) = \sqrt{M_{raw}^2(k) \big/ M_{\mathrm{mod}}^2(k)},\ k = 1, \ldots, K \qquad (7)$$

where $\mathbf{M}_{raw}^2$ and $\mathbf{M}_{\mathrm{mod}}^2$ are the empirical second moments of the raw and model MRCC vectors correspondingly. $\mathbf{M}_{raw}^2$ can easily be calculated from the Gaussian mixture parameters (i.e. means, variances and mixture weights). In order to calculate $\mathbf{M}_{\mathrm{mod}}^2$, a large number of sentences is synthesized and the MRCC vectors emitted from the selected HMM state are collected. The second moment vectors are smoothed along the quefrency axis with a 5-tap moving average operator before calculating the ratio vector (7). The **R** vector components calculated for the HMM state analyzed on Figure 2 are represented by the stemmed plot in Figure 3.

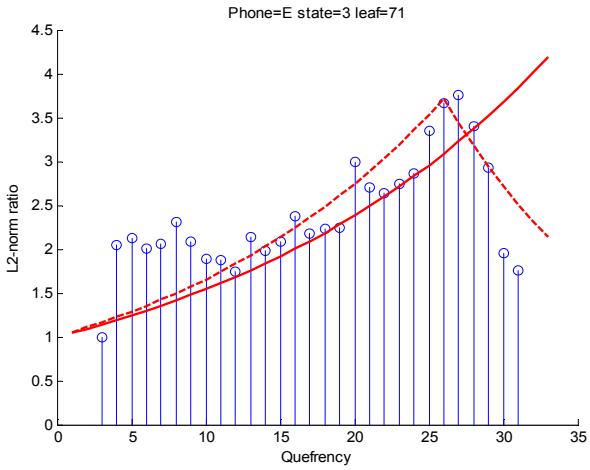

Figure 3: *L2-norm vector and optimal enhancement functions estimated for the analyzed HMM state*

The ratio vector components exhibit an increasing trend along the quefrency axis which means that the model vectors have a stronger attenuation than the raw vectors on average. This statistical observation was validated on all the HMM states of several male and female voice models in three languages summing up to about 7000 states. The observations and statistical evidence above hints to compensate for this stronger attenuation of model vectors by *liftering* the MRCC vectors with an enhancement function prior to synthesis. Relation (6) suggests a simple and tractable enhancement liftering by following exponential function:

$$\hat{s}(n) \cdot \rho^n = \sum_{k=1}^{K}\frac{(\rho p_k)^n}{n} - \sum_{m=1}^{M}\frac{(\rho z_m)^n}{n}, n = 1, 2, \ldots \qquad (8)$$

$$1 < \rho < \frac{1}{\max\{|p_k|, |z_m|\}} \qquad (9)$$

The exponential liftering results in the uniform radial migration of poles and zeros towards the unit circle of the complex plane that directly relates to spectrum sharpening without changing the location of the peaks and valleys on the frequency axis. The degree of the spectrum sharpening depends on the selected exponent base ρ. A too high ρ may overemphasize the spectral formants and even render the inverse cepstrum transform instable. On the other hand, a too low ρ may not yield the expected enhancement effect. It seems obvious that ρ may vary significantly across different states. Inequality (9) imposes a theoretical limit on the exponent base value within the rational representation of the vocal tract transfer function without giving a practical recipe for choosing an appropriate value of this parameter.

Our approach to this problem consists in the estimation of ρ for each state by utilizing the statistical properties of the state represented by the L2-norm ratio vector (7). ρ is calculated so that the enhanced model vectors exhibit the attenuation observed in the corresponding raw vectors on average. The enhancement parameter (i.e. the exponent base

ρ) for a given HMM state can be estimated by the linear regression of the log L2-norm ratio vector:

$$\log \rho = \sum_k k \cdot \log R(k) \Big/ \sum_k k^2 \qquad (10)$$

On Figure 3, the exponential liftering function estimated for the analyzed HMM state is shown by the solid red line. An observation of typical shapes of the L2-norm vectors motivated an alternative, less tractable mathematically, enhancement function in the form of two concatenated exponents. In this scenario, an exhaustive search is performed for finding optimal concatenation point in combination with the log-linear regression based estimation of the two exponent base values. The piece-wise exponential liftering function is shown by the dashed red line on Figure 3.

The result of the optimal exponential enhancement applied to the model vectors is illustrated by a representative example on Figure 2. It can be seen that the enhanced model spectrum exhibits emphasized peaks and valleys and resembles the raw spectra much better compared to the original model spectra.

The optimal enhancement parameters are estimated and stored for each HMM state. During synthesis time, each model MRCC vector is liftered with the enhancement parameters estimated for the HMM state emitting that vector. Hereafter term *enhanced MRCC* (*eMRCC*) is used for the MRCC parameterization of model segments combined with the adaptive enhancement.

4. Experimental setup and results

Subjective listening tests were performed to assess the proposed uniform parameterization on two important aspects for the MFS synthesis: the overall quality of the modeled speech and the voice character similarity between the modeled and template-based speech. In both evaluations the eMRCC parameterization (with order of 33) was compared to an earlier experimental *baseline* which employs an enhanced Mel-warped LPC-based cepstrum representation. Experimental US English male and female voice models were trained with the eMRCC and the baseline parameterizations using respective datasets, each containing approximately 1.5 hours of 22 kHz speech. To produce the test samples both the systems were ran at the MTR=100% operating point, i.e. in "all model segments" mode.

The modeled speech quality was assessed by a standard MOS test. 10 out-of-training domain samples were synthesized with each voice model. Over 20 participants judged on the quality of the samples. The MOS results, depicted in Figure 4, indicate a significant advantage ($p<0.001$) of the new eMRCC speech parameterization over the baseline in terms of the modeled speech quality.

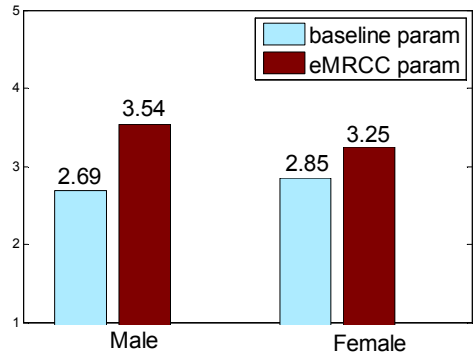

Figure 4: *MOS test results*

To assess the voice character similarity between the model and template segments, five in-training domain PCM signals from the male and female datasets were re-synthesized in "all model segments" mode. In this case the original phoneme durations were imposed. 15 participants judged on the similarity between the synthesized and corresponding PCM samples using following 5-point scale: 1 - "no similarity", 2 - "slight similarity", 3 – "fair similarity", 4 – "good similarity", 5 - "high similarity". The mean similarity scores, presented on Figure 5 show that the new eMRCC parameterization significantly outperforms ($p<0.001$) the experimental baseline. In absolute terms the similarity achieved by the new parameterization was qualified by the listeners between "fair" and "good" on average.

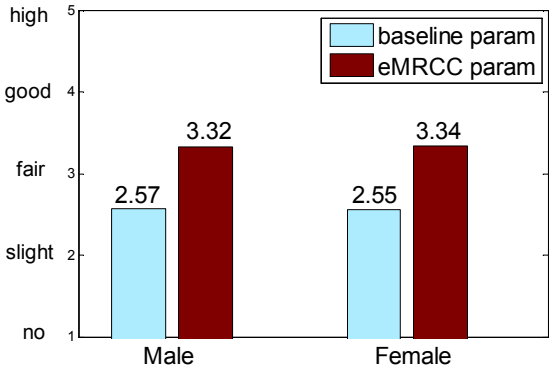

Figure 5: *Model-template similarity test results*

5. Conclusions

A uniform MRCC-based speech parameterization aiming at MFS TTS quality improvement especially for high Model-Template Ratio operating points was presented. The WMRCC phase parameterization contributes to the quality improvement of template segments. The adaptive model segment enhancement technique operating on the uniform speech parameterization significantly improves the quality of statistically modeled speech reconstruction and further reduces the perceived gap in voice character between model and template segments.

6. References

[1] Hunt, A. and Black, A., "Unit selection in a concatenative speech system using a large speech database", in Proc. ICASSP, 1996.

[2] Black, A., Zen H. and Tukoda K., "Statistical parametric speech synthesis", in ICASSP, 2007.

[3] Pollet, V. and Breen, A., "Synthesis by generation and concatenation of multiform segments", in Proc. *Interspeech* 2008.

[4] Chazan, D., Hoory, R., Sagi, A., Shechtman, S., Sorin, A., Shuang, Z. and Bakis, R., "High quality sinusoidal modeling of wideband speech for the purpose of speech synthesis and modification", In Proc. ICASSP 2006.

[5] Shechtman, S. and Sorin, A., "Sinusoidal model parameterization for HMM-based TTS system", in Proc. Interspeech 2010.

[6] Chazan, D., Hoory, R., Kons, Z, Silberstein, D, Sorin, A., "Reducing the footprint of the IBM trainable speech synthesis system", In Proc. ICSLP 2002.