Refined Inter-segment Joining in Multi-Form Speech Synthesis

Alexander Sorin\textsuperscript{1}, Slava Shechtman\textsuperscript{1}, Vincent Pollet\textsuperscript{2}

\textsuperscript{1} Speech Technologies, IBM Haifa Research Lab, Haifa, Israel
\textsuperscript{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 speech synthesis, speech output is constructed by splicing waveform segments and parametric speech segments which are generated from statistical models. The decision whether to use the waveform or the statistical parametric form is made per segment. This approach faces certain challenges in the context of inter-segment joining. In this work, we present a novel method whereby all non-contiguous joints are represented by statistically generated speech frames without compromising on naturalness. Speech frames surrounding non-contiguous joints between the waveform segments are re-generated from the models and optimized for concatenation. In addition, a novel pitch smoothing algorithm that preserves the original intonation trajectory while maintaining smoothness is applied. We implemented the spectrum and the pitch smoothing algorithms within a multi-form speech synthesis framework that employs a uniform parametric representation for the natural and statistically modeled speech segments. This framework facilitates pitch modification in natural segments. Subjective evaluation results reveal that the proposed smoothing methods significantly improve the perceived speech quality.

Index Terms: speech synthesis, text-to-speech, multi-form synthesis, hybrid speech synthesis, pitch smoothing, spectral smoothing

1. Introduction

Reducing perceptual discontinuities at non-contiguous inter-segment joints is a widely acknowledged challenge in the unit selection-based Text-To-Speech synthesis (TTS). In particular, discontinuities often occur due to the unit repository sparseness in the case of limited size voices or when attempting to synthesize speech that exhibits a highly expressive character. In such cases, the unit repository is lacking the segments to be able to generate a smooth and natural trajectory. Statistical-parametric TTS typically produces free of discontinuities speech but at the cost of muffled sound and loss of the character of the speaker. Multi-Form Segment (MFS) synthesis \cite{1, 2} interleaves segments derived from natural speech (template segments) with artificial segments generated from statistical parametric models (model segments). Our previous publications \cite{3, 4} present methods to increase the flexibility and improve the homogeneity of synthetic speech illustrating that MFS synthesis is a viable framework enabling highly natural speech that is free of discontinuities and preserves the character of the speaker. However when the system is put under pressure by the use of very small data sets, highly expressive and uncontrolled speech or in case of out-of-domain expressive prosody targets, discontinuities may still occur.

In general, the audible artifacts are attributed to discontinuities in temporal evolution of one or more of the following components around a non-contiguous joint: instantaneous spectral envelope, phase information and pitch. Following questions need to be addressed within the MFS synthesis framework to deal with such events: 1) How to determine which spectra has to be generated from the statistical models? 2) How to assure that the model generated spectra smoothly connect with the surrounding natural spectra? 3) How to augment the model spectra with phase information to guarantee smooth connection of the time-domain waveforms? 4) How to smooth the pitch contour?

An attempt has been made in \cite{2} to address the first two questions above. Within the approach developed in this work, a whole template segment is replaced by a model segment, once a spectral distance between this template segment and its neighboring template segments is above a pre-defined threshold. This approach has two obvious drawbacks. First, it does not guarantee that all the audible discontinuities will be detected and fixed. Secondly, the replacement of whole segments, especially exhibiting non-stationary and loud characteristics, generally results in muffled quality pertinent to the statistically generated speech \cite{4}. If the spectral distance threshold is too high then almost nothing is fixed. If the threshold is too low then almost all the segments are generated statistically and the output speech sounds muffled.

In response to the second problem, a boundary constrained generation of model spectra is proposed in \cite{2}. The boundary conditions are given by the adjacent natural frames and the delta and delta-delta relations at the edges of the model segment. However the mathematical formulation of this idea proposed in \cite{2} does not take into consideration the specific form of the boundary constraints and leads to an overcomplicated and inefficient solution that was improved in our previous work \cite{4}.

The questions related to handling the phase information and pitch smoothing are not addressed in \cite{2}.

In this work we address all the above aspects and describe an approach and algorithms that enable effective removal of the audible discontinuities in synthesized speech within MFS-framework minimizing the mulling effect to a negligible level.

The rest of the paper is structured as follows. In Section 2 we present an overview of the experimental MFS framework and our general approach to the segment joints refinement. In Sections 3 we present the spectrum and phase smoothing algorithms. In Section 4 we present the pitch smoothing method. In Section 5 we report the results of a subjective evaluation.

2. Reduction of discontinuities: general approach

The approach and algorithms described below are applicable to MFS TTS systems that employ a uniform parameterization for template and model segments and a harmonic-plus-noise type of reconstruction.

Without loss of generality we present our method in application to the experimental MFS system reported in our
previous work [3]. As a parameterization scheme, we adopt the multi-layered uniform parameterization based on a sinusoidal representation and Mel-Regularized Cepstral Coefficients (MRCC) parameterization of the spectral envelope and phases [3, 5]. An additional layer of residual noise components (represented by random codebook entries and gains) is optionally added for the template segments to further increase the waveform reconstruction quality. During the synthesis, the spectral and phase parameters (MRCC) are converted to a sinusoidal model representation (a.k.a. line spectrum). This framework allows a convenient frequency domain pitch modification [6] within template and model segments.

The system can be operated in a purely concatenative mode, i.e. at zero model-template ratio (MTR) working point. Alternatively, some segments may be generated statistically based on the criterion proposed in our work [4] or using another selection mechanism. If prosody modification is not required, the baseline MFS system can be operated without the parameterization of template segments which are represented by PCM waveforms.

By analyzing synthesized speech samples in detail, we observed that vast majority of audible discontinuities occur at voiced joints, i.e. the joints surrounded by voiced speech frames. This observation holds for both PCM-based and parameterized template segments. For technical reasons explained below we refine the notion of voiced joint requiring existence of at least two voiced frames at each side of the joint.

Hence, when dealing with spectral discontinuities we process all non-contiguous voiced joints between template segments (referred hereafter as target joints) and only these joints. At each target joint we select a small number of frames preceding and following the joint and regenerate them from corresponding statistical models. Hence we apply the model based synthesis at the sub-segment level. Thus all potential spectral discontinuities are handled while the MTR is kept reasonably low to minimize the muffling effect and maintain the character of the speaker.

We employ a boundary constrained generation of the MRCC frames using the efficient approach proposed in our previous work [4] and applying it to the sub-segment level. The details of the spectral discontinuity removal algorithm including the phases generation are given in the next section.

A similar approach could be applied to the pitch smoothing. However, our experiments reveal that a local pitch contour smoothing is not effective enough. This is why we developed a special pitch smoothing algorithm that considers a wider temporal scale and guarantees the desired level of the pitch contour smoothness. Finally not only model but also template frames are reconstructed using the smoothed pitch. This is underpinned by the parametric representation that enables pitch modification at the template frames.

3. Sub-segmental statistical spectral smoothing and phases generation

The process of model frames selection is illustrated by the schematic diagram depicted on Figure 1 where a non-contiguous voiced joint between two (initially template) segments is shown. Speech frames represented by the rectangular boxes are numbered and labeled as voiced (V) and unvoiced (U).

We consider the leading frames of Segment 2 within the vicinity of the joint specified by a half-window control parameter (equals to 5 frames in the example of Figure 1). On the figure these frames are indexed by the numbers 16 through 20. Within this frame set we identify the subset of consecutive voiced frames adjacent to the joint. It is comprised of the frames 16 through 18. All the frames excluding the last one (frame 18) are selected for the model-based regeneration. The mirrored version of the same procedure is applied to the trailing frames of Segment 1. Finally the frames 12 through 17 (shown by blue colored boxes) are selected for the model representation. Frames 10, 11, 18 and 20 establish boundary conditions for the model representation. As explained below, a set of consecutive model frames is fully determined by the boundary conditions established by the two pairs of surrounding template frames. Voiced frames 11 and 18 will be used also for harmonic phase generation at the model frames as explained below.

If two or more joint-wise sets of selected frames intersect they are merged into a single larger set. However this happens rarely; the unit selection mechanism is often set up to favor the longer contiguous sequences of segments while the half-window length is short (for 5ms frame size we typically use 5 frames as half-window size).

Once the model-template dichotomy is established at the frame level we generate model cepstra (MRCC) vectors from the respective statistical models so that the utterance level trajectory passes through the (known) template cepstra vectors. In other words, we compute the maximum likelihood (ML) trajectory that interpolates the fixed template cepstra vectors.

The classical (unconstrained) utterance level ML trajectory is obtained by solving quadratic optimization problem [7]:

$$c^* = \arg \min \{ c^T W^T \Sigma^{-1} W c - 2 e^T W^T \Sigma^{-1} \mu \}$$

(1)

where $$c = [c_1, ..., c_N]$$ is a concatenation of the cepstra vectors, $$N$$ is the number of frames in the utterance, $$\mu = [\mu_{11}, ..., \mu_{1N}]$$ is a concatenation of the Gaussian mean vectors corresponding to the HMM states (or acoustic leaves) $$[L1, ..., LN]$$ associated with the frames, the block-diagonal matrix $$\Sigma = \text{diag}([\Sigma_{11}, ..., \Sigma_{NN}])$$ is composed from the covariance matrices corresponding to the leaf sequence and the matrix $$W$$ implements the operator augmenting static features by dynamic components. Usually the 1st and 2nd order dynamic components (delta and delta-delta) are used. The leaf mean vectors and covariance matrices include these dynamic components.
Let \( \mathbf{m} = [\mathbf{c}_m^T, \ldots, \mathbf{c}_{m+n}^T]^T \) and \( \mathbf{t} = [\mathbf{c}_t^T, \ldots, \mathbf{c}_{t+n}^T]^T \) be concatenations of the model and template cepstra vectors respectively. Knowing the respective locations of the model and template vectors within the whole trajectory vector \( \mathbf{c} \), one can easily compose two sparse matrices \( \mathbf{T} \) and \( \mathbf{M} \) so that:

\[
\mathbf{T} \cdot \mathbf{t} + \mathbf{M} \cdot \mathbf{m} = \mathbf{c}
\]  

Combining (2) with (1) and treating the vector \( \mathbf{t} \) as known leads us to the following quadratic optimization problem:

\[
\mathbf{m}^* = \arg \min_{\mathbf{m}} \left[ \mathbf{m}^T \mathbf{M}^T \mathbf{W}^{-1} \mathbf{W} \mathbf{M} \mathbf{m} - 2 \mathbf{m}^T \mathbf{M}^T \mathbf{W}^{-1} (\mathbf{\mu} - \mathbf{W} \mathbf{T} \mathbf{t}) \right]
\]  

Finally, the sequence of the model vectors that lie on the constrained ML trajectory is obtained by solving the set of linear equations:

\[
\mathbf{M}^T \mathbf{W}^{-1} \mathbf{W} \cdot \mathbf{m}^* = \mathbf{M}^T \mathbf{W}^{-1} (\mathbf{\mu} - \mathbf{W} \mathbf{T} \mathbf{t})
\]  

Notice that the left-side matrix in (4) has a sparse diagonal band structure because only delta relations tie frames to each other. If the delta features calculation for a frame at time \( t \) relies only on its immediate neighbors (e.g. \( \Delta_i^t = 0.5(c_{i+1} - c_{i-1}) \), \( \Delta_i^t = c_{i+1} + c_{i-1} - 2c_i \) then any single equation cannot tie more than 3 consecutive frames. The model frames selection algorithm described above guarantees that each set of consecutive model frames (11 through 17 in our example) is surrounded by at least two template frames. It means that the equations system (4) can be decomposed to mutually independent smaller systems, each determining one consecutive set of model frames (typically surrounding a single joint). Similarly to the conventional ML trajectory case, the system (4) and its separate sub-systems can be decomposed and solved for each feature vector component independently (1D trajectories) in the case of diagonal covariance matrices.

A real example of the boundary constrained trajectory obtained for the 2nd MRCC component with our algorithm is shown on Figure 2. We used an experimental MFS voice with 5 ms frames and half-window = 25 ms. The coefficient values corresponding to template frames are shown as red stems. The original template trajectory (red curve) around the joint point features a discontinuity that was smoothed by the constrained ML trajectory (red curve).

Once a model MRCC vector is generated using the above algorithm, harmonic magnitudes are calculated by sampling the spectral envelope derived from this vector at the harmonic frequencies corresponding to the frame [5]. Harmonic phases are calculated using bilinear frequency-time interpolation of the complex harmonic amplitudes corresponding to the surrounding template frames (e.g. frames indexed by the numbers 11 and 18 in the example shown on Figure 1). Let \( t_i \), \( t_2 \) and \( t_3 \) be respective indices of a model frame, the last preceding template frame and the first following template frame, \( t_1 < t_2 < t_3 \). Then the phase value associated with the harmonic frequency \( f \) of the model frame is calculated as:

\[
\varphi_i(f) = \arg \left\{ \beta \left[ \alpha_i S_i(f) + (1 - \alpha_i) S_i(f) \right] + (1 - \beta) \left[ \alpha_i S_i(f) + (1 - \alpha_i) S_i(f) \right] \right\}
\]  

\[
\beta = (t_2 - t)/(t_3 - t_1)
\]  

\[
\alpha_i = (f - f)/\left(f_i - f\right).
\]

Harmonic frequencies for all the frames are derived from the smoothed pitch contour calculated as explained in the next section.

### 4. Smoothed pitch contour generation

A fragment of \( F_0 \) contour corresponding to the segments selection made by the baseline MFS system working in an expressive prosody target mode is presented on Figure 3 in blue. The fragment is composed of template segments only. The segments boundaries are depicted by black vertical stems. Contour discontinuities (jumps) are observed at some of non-contiguous joints. Our goal is to eliminate these discontinuities, while preserving the intonation contour. To this end, we generate a new smooth \( F_0 \) contour that alleviates the jumps and follows in general the shape of the original contour.

---

1 In general, 2m boundary template frames are required when the dynamic features calculation is based on 2m+1 consecutive frames.
following quadratic optimization problem with linear constraints:

\[ P^*(n) = \arg \min_{P(n)} \left[ \sum_{s,t} \frac{1-\lambda(s)}{\lambda(s)} \left( \vec{P}(s,t) - \vec{T}_s \right)^2 + \| \Delta P \| \right] \]

\[ \vec{P}_s = \sum_{s=1}^{N_s} P(n) \left( N_{s,n} - N_{s,1} \right) \]

\[ \Delta P(n) = 0.5 \left( P(n+1) - P(n-1) \right) \]

s.t. \[ \Delta P(n) < \delta, \ P_{\text{min}} \leq P(n) \leq P_{\text{max}} \]

where \( N_{s,1} \) and \( N_{s,n} \) are respectively the first and the last frame index within the segment \( s \) and \( V \) is the voiced region (i.e. a sequence of fully voiced segments) containing \( s \). Thus the smooth pitch contour is calculated for each voiced region separately. The minimization criterion establishes a tradeoff between the fit to the original intonation and smoothness controlled by the segment dependant smoothing parameter \( \lambda(s) \) concluded between 0 and 1. The two inequality constraints assure respectively a certain level of the pitch contour smoothness and a certain pitch range. We set \( \delta = fs \times \log 1.05 \), where \( fs \) is a frame size in ms. \( P_{\text{min}} \) and \( P_{\text{max}} \) equal to the highest and the lowest \( F_0 \) values observed in the voice dataset, correspondingly. The values of the segment dependent smoothing parameter \( \lambda(s) \) are determined automatically in such a way that the smoothing is stronger in a vicinity of the joints featuring large jumps. This idea is implemented by the following procedure. For each non-contiguous joint we calculate an \( F_0 \) jump measure \( d_j \) as the absolute log-ratio between the \( F_0 \) values at the adjacent frames and a joint weight \( w_j \) value as:

\[ w_j = \begin{cases} w_{\text{max}} & \text{if } d_j > G \cdot d_{\text{out}} \\ w_{\text{min}} & \text{if } d_j < d_{\text{out}} \text{ else} \\ w_{\text{min}} + (d_j - d_{\text{out}})(w_{\text{max}} - w_{\text{min}})/(d_{\text{out}}(G-1)) \end{cases} \]

where: \( w_{\text{max}}, w_{\text{min}} \) and \( G \) are configuration parameters (e.g. set to 0.15, 0.85 and 5 respectively); \( d_{\text{out}} \) is a measure of natural frame-to-frame \( F_0 \) jump calculated as 0.8-quantile of intra-segment jumps distribution either at a sentence level or a voice dataset level. The weights of contiguous joints are set to \( w_{\text{min}} \). All the weights are associated with the time moments of the respective joints. Then the weights of those contiguous joints that are located within certain vicinity (e.g. 50ms) of non-contiguous joints are recalculated using linear interpolation of the between other weights. Finally the \( \lambda \) value for a segment is set to the maximum of the weights associated with the two joints at the segment edges. The \( \lambda(s) \) values multiplied by 100 are depicted by the red stems on the Figure 3.

The algorithm described above produces inherently smooth pitch contours. However it smoothes also abrupt transitions related to natural vocal fry, resulting in quality deterioration. To resolve this, we identify the pitch transition regions related to vocal fry occurrences as intra-segment half-pitch jump locations where the pitch drops below a certain threshold (e.g. 95 Hz for a female voice). For these regions, the half-pitch values are doubled. Then the smooth pitch contour generation algorithm described above is applied. Finally the vocal fry region is restored by halving the smoothed values.

5. Subjective listening evaluation

We used an experimental MFS TTS system as reported in our previous work [3] with mid-sized US English female voice conveying a fast paced conversational style. The system was operated in the fully concatenative mode with out-of-domain expressive prosody targets (hence putting the synthesis system under pressure). Four versions of this MFS system were evaluated on a set of 32 stimuli. Identical segment selection was used for all the four versions. The reference version, denoted “MFS-PCM”, applies a non-parametric time-domain waveform concatenation. The second version, denoted “MFS-PRM”, applies the harmonic plus noise parameterization, as described in Section 2 with no smoothing operation. “MFS-PRM-F0” version employs fully parametric synthesis with the smoothed pitch contour. “MFS-PRM-F0-SPC” version implements both the pitch and the spectral smoothing.

The evaluation was performed using a MOS test publishing tool based on crowd-sourcing technique. A mix of speech experts and anonymous subjects provided by Amazon Mechanical Turk (AMT) platform participated in this evaluation. 71 subjects participated in the test. MOS values were estimated after the outlier-subjects rejection, as in [8], resulting in about 25% of the subjects removed. The results are presented graphically on Figure 4. It can be observed that the pitch smoothing increased the MOS by 0.09 and the spectral smoothing yielded 0.19 extra gain in MOS so that the cascade of the two techniques resulted in a 0.28 increase of MOS relatively to the MFS-PRM version without smoothing.

Switching from PCM to the parametric representation somewhat decreases the MOS. However the parameterization enables the pitch and spectra discontinuity removal and therefore eventually pays out. The final version MFS-PRM-F0-SPC outperforms the baseline PCM-based version by 0.2 MOS units. The differences between the final version and all the others are statistically significant with p-values below 0.001.

6. Conclusions

We developed an effective method for removal of discontinuities in Multi-Form Segment TTS operated in challenging settings such as sparse voice datasets, out-of-domain input and expressive prosody. The method utilizes the uniform speech parameterization and includes the statistical spectral smoothing at a sub-segmental level, time-frequency harmonic phase interpolation and the global pitch contour smoothing.
7. References


