Psychoacoustic Segment Scoring for 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 Technologies, IBM Haifa Research Lab, Haifa, Israel

sorin@il.ibm.com, slava@il.ibm.com, vincent.pollet@nuance.com

Abstract

In multi-form segment synthesis, output speech is constructed by splicing waveform segments with statistically modeled and regenerated parametric speech segments. The fraction of model-derived segments is called model-template ratio. The motivation of this work is to further increase flexibility of multi-form synthesis maintaining high speech quality for high model-template ratios. An approach is presented where the representation type of a segment is selected per acoustic leaf. We introduce a novel method for leaf representation selection based on a psychoacoustic segment stationarity score. Additionally, refinements in multi-form segment concatenation including boundary constrained statistical parametric synthesis and time-domain alignment based on multi-peak analysis of cross-correlation for high model-template ratio multi-form synthesis are presented.

Index Terms: speech synthesis, multi-form segments, speech stationarity, psychoacoustic segment scoring, statistical parametric synthesis, segment concatenation

1. Introduction

Multi-Form Segment (MFS) \cite{1} TTS is a speech synthesis technique that employs a statistical framework with segments of different representational structures. Both natural speech fragments and models are used in the speech construction process resulting in higher quality and naturalness compared to statistical parametric TTS and higher flexibility, generalization and smoothness compared to concatenative/unit-selection TTS.

In the MFS system, the Model-Template Ratio (MTR) expresses how much of the output speech is generated from model segments on average. The choice of representation (model or template) for a segment depends on three categories of cues: phonologic cues, acoustic cues and channel cues. In the initial work \cite{1} it was observed that highly natural speech can be produced at medium MTR levels. For achieving higher ratios, the system is predominately constrained by the psychoacoustic segment stationarity score and speech parameterization technique, based on sinusoidal representation and Mel-Regularized Cepstral Coefficients (MRCC) was introduced for MFS which allowed to further reduce the gap in perception between model-generated and template-based speech.

This paper introduces a psychoacoustic segment scoring technique for the multi-form segment TTS system. The novel scoring technique is assessed as a channel cue in isolation that determines the multi-form segment representation type choice solely, being not augmented by phonetic and acoustic cues as in \cite{1}. Next, a series of refinements are introduced for synthesizing speech at high MTR levels including boundary constrained generation of model segment acoustic parameters and waveform discontinuity reduction. Finally, conclusions are drawn from this work based on large scale subjective evaluations.

2. Psychoacoustic metrics and algorithm for choice of speech representation type

Speech segments generated from a statistical model, for example an HMM state, are highly stationary by definition. Model segments have a slowly in-time evolving spectral envelope and an excitation which is a mix of quasi periodic and random stationary components. In particular, statistical TTS systems fail to reproduce transient sounds adequately. This observation raises the idea of considering a stationarity statistics measured within a cluster of template segments as a suitability indicator for statistical parametric modeling.

We start with developing a stationarity score for a template segment. The segment associated with an HMM state representing an acoustic leaf of a statistical TTS system is divided into overlapping frames at a high frame rate, for example 1 kHz. The frame length is chosen wide enough to cover at least one pitch cycle of voiced speech. For example, for a 3-state phone-level HMM, the segment is typically longer than 25 ms and contains tens of frames.

Each frame is transformed to a perceptual loudness spectrum (PLS). The similar transformation is utilized in the Perceptual Linear Predictive ASR front-end \cite{3}. As a result, each frame is represented by a PLS vector whose components are proportional to the perceptual loudness levels associated with respective critical frequency bands. Let \( V(t) = [v_1(t), \ldots, v_g(t)] \) be the PLS vector derived from the \( t \)-th frame, \( V^2(t) \) be the squared PLS vector and \( T \) is the number of frames in the segment. Let \( M_1 \) and \( M_2 \) be respectively the 1\textsuperscript{st} and 2\textsuperscript{nd} empirical moments of the PLS vector distribution within the segment:

\[
M_1 = \frac{1}{T} \sum_{t=1}^{T} V(t) \quad M_2 = \frac{1}{T} \sum_{t=1}^{T} V^2(t)
\]

The segment non-stationarity measure \( R \) can be defined as integral relative variability of the PLS vector components:

\[
R = \frac{\sum_{t=1}^{T} (M_1^t - M_1)^2}{\sum_{t=1}^{T} M_2^t}
\]

where \( M_1^t \) and \( M_2^t \) are components of the vectors \( M_1 \) and \( M_2 \) respectively. We define the segment stationarity score \( S \) as

\[
S = \frac{1 - R - 1/T}{1 - 1/T} = \frac{1}{1 - 1/T} \frac{\sum_{t=1}^{T} M_1^t}{\sum_{t=1}^{T} M_2^t - 1/T}
\]
It can be proved that the stationarity score $S$ is defined within the range $[0,1]$ receiving the maximum value for a perfectly stationary segment with the identical PLS vectors and the minimum value for a singularly non-stationary segment with only one non-zero PLS vector.

The stationarity score histogram is depicted by the bar diagram and the LSM position is marked by the dashed stem.

Figure 1: Segmental stationarity scores.

In the example shown on Figure 1, the speech signal is segmented according to the HMM state-level alignment. The segment boundaries are depicted by vertical lines. The stationarity scores aligned with the respective segments are represented by the piece-wise constant line above the signal waveform.

Further we derive a stationarity measure of an acoustic leaf from the stationarity scores of all the template segments associated with this leaf. The Leaf Stationarity Measure (LSM) is set close to the lower bound of segmental stationarity scores within the leaf, for example 0.1-quantile of their empirical distribution. In the examples shown on Figure 2 the stationarity score histogram is depicted by the bar diagram and the LSM position is marked by the dashed stem.

Figure 2: Leaf Stationarity Measure calculation. Top - beginning of phone “p”, LSM=0.56. Bottom - middle part of phone “E”, LSM=0.96.

An attempt to choose the leaf representations based just on their LSM values results in noticeable quality degradation even at low MTR levels. This can be explained by the fact that the most stationary leaves typically represent the louder parts of vowels. Hence the template-model joints and the modeled character of voice can become audible. To include this sensitivity into the psychoacoustic score, we augment the LSM with a Leaf Loudness Measure (LLM). The LLM is derived similarly to the LSM. First we define a loudness score $L$ of a template segment as:

$$L = \sum_{i=1}^{N} M_i$$

The LLM is set close to the upper bound of the segmental loudness scores associated with the leaf, for example 0.9-quantile of their empirical distribution.

Once the stationarity LSM, and loudness LLM, measures are calculated for all the acoustic leaves $i=[1,...,I]$ we normalize them at the voice dataset level:

$$NM_i = \frac{M_i - \min(M_i)}{\max(M_i) - \min(M_i)}$$

where $M_i$ is either LSM or LLM, and $NM_i$ designates the normalized versions $NLSM_i$ or $NLLM_i$.

Finally the Leaf Modeling suitability Factor $LMF$ is defined by mixing the respective stationarity and loudness measures:

$$LMF = 0.5 \cdot \left[ NLSM_i + (1 - NLLM_i) \right]$$

According to this definition, the more stationary and the more silent the leaf is, the more favorable it is for the model-based representation.

Finally all the leaves are sorted in the descending order of their LMF values. Given an arbitrary target MTR level $P\%$, we mark the required number of the top leaves as being a “model”. The number of the model leaves is calculated as such that the durations of the segments comprising these leaves are summed up to $P\%$ of the total duration of all the segments in the voice dataset. At synthesis time, segments associated with model leaves are generated from the respective statistical parametric models while other segments are considered as templates.

Figure 3: Leaf modeling suitability scoring for a female US English voice.

The LMF mapping and model-template leaf dichotomy is illustrated by the example shown on Figure 3. All the acoustic leaves of a female US English voice are depicted by points (the circle centers) at the NLSM-NLLM plane. The LMF value of a leaf can be obtained by the projection of the corresponding point to the LMF axis. For MTR=30% chosen as an example, the line $LMF=0.72$ separates between the modeled leaves (located below the line) and the template leaves (located above the line).
3. Boundary-constrained generation of model segment acoustic parameters

In statistical parametric TTS, once the state sequence is determined, the maximal likelihood trajectory in the space of spectral parameter vectors is obtained by solving the well known quadratic minimization problem

$$\hat{x} = \min \|Ax - b\|^2, \quad A = \Sigma^{-0.5}W, \quad b = \Sigma^{-0.5}\mu$$ (7)

where $x = [x_1, x_2, \ldots, x_J]^T$ is a concatenation of the state output static feature vectors $x_j$, $\Sigma$ is a block matrix built up from the state covariance matrices, $W$ is a matrix augmenting a static feature vector $x$ by dynamic components. In MFS synthesis those state output vectors corresponding to the template segments are known, i.e. can be obtained by parameterization of the template segments. Then the minimization criterion (7) can consequently be rewritten as:

$$\min_m \|Am - (b - At)\|^2$$ (8)

where vector $m$ contains only the model sub-vectors while the template sub-vectors are replaced by zero-vectors, and vector $t$ contains only the template sub-vectors while the model sub-vectors are replaced by zero-vectors. The minimization is carried out with respect to the model feature vectors only. The solution of the constrained minimization problem (8) can be considered as the maximal likelihood trajectory over the model segments that interpolates the selected template segments. This approach reduces discontinuities at the model-template joints and makes speech quality more homogeneous. The optimization problem (8) can be solved very efficiently. A similar technique can be applied for the boundary-constrained pitch contour generation at model segments.

4. Waveform discontinuity reduction at multi-form segment joints

Usually the frame alignment is performed between the current voiced frame and the previous voiced frame within the model segments and at segment joints of any kind. Although the boundary-constrained solution described above improves the spectral continuity at multi-form segment joints, the concatenated waveforms still might differ significantly, in particular due to the fact that the model-derived waveforms have minimal phase structure [4] which is not necessarily true for the template waveforms.

Voiced template-model joints

The frame alignment for voiced model segment frames is performed in sinusoidal representation (aka line spectrum) domain by adding linear phase terms to harmonic phases [4]. The term that accounts for the phase alignment maximizes the cross-correlation between the current $c_n$ and the previous $c_{n-1}$ line spectra:

$$n^* = \arg \max_n \|\text{IFFT}(c_n c_{n-1}^*)(n)\|$$ (9)

At the template-model joints, however, we don’t have a line spectrum representation for the previous frame, so we estimate $c_{n-1}$ by applying a DFT to the weighted average pitch cycle $s_p(n)$ within the analysis window:

$$c_{n-1} \approx \text{DFT}(s_p)$$ (10)

5. Experimental setup and results

We conducted subjective listening evaluations in order to assess the leaf representation choice mechanism based on the proposed psychoacoustic scoring and the MFS speech signal generation techniques presented above for a wide range of MTR levels. In a standard MOS evaluation we focused on the assessment of MTR-dependent changes in the MOS score.

$$s_p(n) = \sum_{k=-[N_e/2]}^{[N_e/2]} w(n+kP)s(n+kP)$$ (11)

where $P$ is the pitch period length rounded to the integer number of samples, $N_e$ is a number of summations per $n$-th sample, and $w(n)$ is a windowing function.

Additional continuity improvements at template-model joints are achieved by infusion of low-band harmonic phases from the surrounding template line spectra into the model-derived line spectra (hence overriding the minimal phases).

Voiced model-template and template-template joints

The frame alignment for template segments is performed at joints in time domain. The current template segment is either cropped or expanded at its leftmost edge, and then overlapped-and-added with the previous segment. Instead of selecting the concatenation point based on maximal cross-correlation, a more accurate multi-peak cross-correlation analysis is introduced. Examining the magnitudes and the differences of pitch-cycle distant extrema of the normalized cross-correlation function, we hypothesize whether the segments at the joint are similar and periodic and alter the current segment offset calculation strategy accordingly. In the case of high similarity but low periodicity, we select the peak with the largest magnitude, while in the case of similar and periodic segments (indicated by the cross-correlation peaks with high magnitudes and small differences), we select the peak nearest to zero lag in order to reduce segment duration changes as the result of concatenation. The similar strategy is taken in the low-periodicity and low similarity case. The polarity of the segments is altered based on the sign of the selected peak. The examples of the cross-correlation functions observed in the cases described above are depicted on figure 4.

Figure 4: Examples of cross-correlation functions at segment joints of different types
rather than on absolute speech quality level.

We used an experimental MFS TTS system with MRCC parameterization and adaptive enhancement as reported in our previous work [2]. Male and female US English MFS voices were built from mid-size voice datasets. A number of reference sentences were synthesized for each voice in the full concatenative mode (MTR=0%). Then each sentence was re-generated at MTR levels of 25%, 40% and 60% using the leaf classification method and concatenation improvement techniques introduced above. In each sentence, the same selection of the template segments and phone durations were preserved as in the respective reference. This was done in order to assess the MTR influence on the speech quality in isolation.

An in-house evaluation was conducted using 10 male and 10 female sentences, each generated at MTR levels of 0%, 25% and 40%. Each sentence was evaluated by 20 subjects. All of them were non-native English speakers; about a half of them were speech scientists. All the subjects used hi-fi headphones. The MOS results are shown on the diagram of Figure 5.

One can observe that 25% MTR level did not affect the impression of the listeners while raising MTR to 40% changed the perceived quality only slightly. This observation is valid for both the male and female voices despite of the noticeable quality difference in absolute rates.

A second evaluation was performed using a MOS test publishing tool based on crowdsourcing techniques. A mix of speech experts and anonymous subjects provided by Amazon Mechanical Turk (AMT) platform participated in this evaluation. 20 female sentences, each generated at MTR levels of 0%, 25%, 40%, 60% and 100% (purely statistical mode) were submitted to the evaluation. 100 subjects including 60 native English speakers participated in the test. MOS values and confidence intervals were estimated following the two-way random effects model and outlier-subjects rejection [5]. 37% of the subjects were removed as a result of the outlier rejection. The results are presented graphically on Figure 6. MTR=25% samples received the highest average rate outperforming the reference (MTR=0% or fully concatenative mode) ones while the MTR=40% samples were rated equally with the reference. Even at MTR=60%, a relatively small perceived quality degradation was observed. Still the MTR=60% samples were rated substantially higher than purely statistical parametric speech (MTR=100%).

Although both the evaluations have validated a good performance of the method, the crowdsourcing evaluation appeared to be more favorable to the MFS samples (generated at 25% and 40% MTR levels) than the first in-house evaluation. This fact can be attributed to that at least a part of the external anonymous listeners did not use hi-fi headphones. Hence they could not hear the fine variations in voice character but were more sensitive to discontinuities in the reference samples which are more pertinent to the concatenative synthesis in general. The leaf classification and speech signal generation methods described above enabled generation of the MFS samples in which these discontinuities are reduced significantly while the speech naturalness remains transparent.

6. Conclusions

In the context of multi-form speech synthesis, a novel approach for the acoustic leaf representation choice based on psychoacoustic scoring was presented. The speech signal generation is further refined by utilizing boundary constrained statistical models and an improved time-domain segment alignment. Large scale subjective evaluations showed that the proposed method as channel cue in isolation (that is, not including phonetic and acoustic cues) is able to generate transparent speech at high model-template ratios.

7. Acknowledgements

The authors would like to thank Mr. Milan Legat for assistance with the MOS test publishing tool and processing the crowdsourcing evaluation results.

8. References