COHERENT MODIFICATION OF PITCH AND ENERGY FOR EXPRESSIVE PROSODY IMPLANTATION

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

[1]Speech Technologies, IBM Research – 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 expressive TTS and voice transformation systems, implantation of expressive prosody derived from external out-of-domain sources often leads to extreme pitch modification that compromises the naturalness of the synthesized speech.

In this work we investigate and prove a hypothesis that the naturalness loss is in part attributed to a violation of a fundamental relationship between the instantaneous pitch frequency and instantaneous energy of a speech signal.

We propose an enhancement for pitch modification where the instantaneous energy is modified coherently with the pitch frequency and demonstrate the potential of this method in a subjective listening evaluation.

The proposed approach is complementary to and can be combined with spectrum shape transformation methods for achieving the maximal possible quality of pitch modification.

Index Terms— expressive TTS, prosody modification, pitch modification, energy modification, energy modulation

1. INTRODUCTION

Synthesis of speech conveying requested expressive styles on-demand using emotionally neutral voice dataset and voice transformations has generated much interest in the research community and offers great flexibility for commercial applications. For this purpose, expressive target prosody can be generated by a rule-based framework or derived from a relevant, presumably small, expressive data corpus external to the voice dataset used for the synthesis. The prosody target includes basic (phonetic) unit durations and a shape of a pitch contour. Hereafter we focus on the latter part, which greatly affects the perception of the expressive context.

For the sake of generality, considering the step of speech signal generation we refer to the Multi-Form Segment (MFS) synthesis [1] combining different TTS techniques. MFS TTS interleaves parameterized or non-parameterized *template* segments derived from natural speech with *model* segments generated from statistical parametric models. In general, the non-parameterized template segments selected do not exhibit pitch patterns approximating the expressive targets, and PSOLA-based pitch contour modification is needed to match the targets. The model segments and parameterized template segments can be directly generated from the default spectral parameters and the external pitch contour. In any case, the output signal should be synthesized with a pitch contour that exhibits much wider range and much more rapid temporal evolution than a "native" pitch contour that would be inherited from the voice dataset.

There is consensus that extreme modifications of the pitch contour significantly compromise perceptible naturalness and quality of speech signal. A number of research works have demonstrated that a certain degree of quality improvement can be achieved by altering spectral envelope shape in a specific way during pitch modification. Such an approach was proposed in [2] where the pitch modification dependent spectrum transformation is modeled by cepstrum vector codebooks built for low, middle and high F0 ranges. A more generic method was proposed in [3] were speech class dependent LSF parameters transformations are modeled stochastically over a wide range of F0 values. Both the publications demonstrated pitch modification quality improvements in subjective preference evaluations where the original pitch contour was multiplied by a constant factor, i.e. the contour shape was preserved.

Undoubtedly, pitch modification should be accompanied by a proper transformation of spectral envelope shapes for achieving a high speech quality. However there appears to be yet another phenomenon which suggests a relatively simple, but unexplored way for additional pitch modification enhancement.

Multiple research works reported experimental evidence of a salient correlation between instantaneous fundamental frequency and short-time energy of the speech signal. In [4] this phenomenon was observed in professional singers, healthy non-singers and people suffering from voice disorders. In [5] a method was proposed for pitch accent prediction based on the pitch-energy correlation.

Based on this relationship, the quality loss after pitch modification can be partially attributed to an artificially created misbalance between the pitch and energy evolution in time. This explanation is coherent with the subjective impression of the authors gained in informal listening experiments where unnaturally sounding too loud bass segments and too quiet treble segments appeared after implantation of expressive pitch contours to signals generated with a neutral voice.

The pitch-energy relationship suggests that the energy contour must be altered in a specific way during pitch contour modification. The purpose of this work was to explore the potential of this direction alone for the quality improvement in the context of expressive pitch contour implantation.

Note that the spectrum shape transformation, e.g. in LSF representation, does not alter the energy in a controllable manner. Hence the energy contour altering is complementary to the spectrum shape transformation. It is worth noting that we address arbitrary pitch contour modification which is more challenging comparing to the uniform pitch raising or lowering experiments reported in [2] and [3].

The rest of this paper is organized as follows. In Section 2 we demonstrate and quantify the pitch-energy relationship on TTS voice datasets. Section 3 describes our experimental synthesis system including a simple algorithm for coherent pitch and energy contours modification. In Section 4 we describe a subjective evaluation setup and results. Section 5 contains a concluding discussion.

2. EVALUATION AND PARAMETRIZATION OF THE PITCH-ENERGY RELATIONSHIP

We conducted a study in order to validate the pitch-energy relationship and to characterize it quantitatively within a framework used in our speech generation method. To this end we used an MFS TTS voice built from highly expressive Sports News sentences uttered by a female US English speaker and recorded at 22050 Hz sampling rate. As a starting phase of the voice building procedure, the speech signals were analyzed at the frame update rate of 5 ms. The analysis included pitch contour estimation using an algorithm similar to [6] and estimation of harmonic magnitudes using the method proposed in [7]. Within the MFS voice dataset we considered all fully voiced (i.e. comprised of voiced frames only) template sub-phone segments clustered by means of a binary decision tree classification applied to the MRCC spectral parameter vectors [8] representing the segments. Hereafter the clusters are referred to as *leaves*. Harmonic energy level in decibel units was estimated in the full frequency band for each frame comprising these fully voiced segments:

$$EdB = 10 \cdot \log_{10} \sum_{k=1}^{N} A_k^2 \qquad (1)$$

where A_k is the magnitude of the k-th harmonic and N is the number of harmonics for that frame excluding the DC component.

Pruning of leaves containing less than 5 processed voiced frames retained about 1900 leaves containing in total about 1.6 million frames. Within each leaf, the processed frames were represented by points on the (logF0, EdB) plane as shown in the example presented on Figure 1.

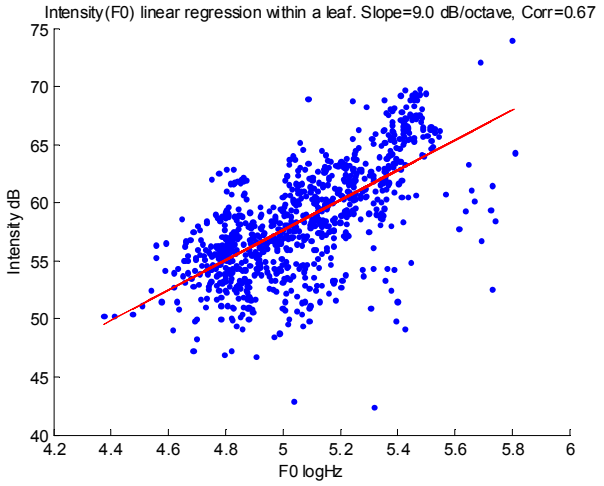

Figure 1. Pitch-energy relationship analysis within a leaf cluster. Frames are depicted by blue points on the F0 logHz – Energy dB plane. The red solid line represents the result of linear regression. Pitch-energy correlation coefficient = 0.67. The line slope = 9 dB/octave.

The cloud of frame-points within a leaf was approximated by the least squares linear regression. (See the red line on Figure 1.) The pitch-energy relationship within a leaf l was represented by the slope S_l of the linear regression expressed in dB per octave. Assuming such a simple linear model, one can say that within the leaf l, altering F0 by x octaves implies $x \cdot S_l$ dB change of the energy on average. Another intra-leaf parameter calculated is the correlation coefficient C_l between the frame-wise sequences of logF0 and EdB values. The analysis reveals that the slope value varies significantly between leaves. Averaging of the intra-leaf parameters weighted by the respective leaf frames count yields the voice level statistics: S_{voice} = 4.5 dB/octave and C_{voice} = 0.52.

The same evaluation performed on a neutral US English dataset built from the voice of another female speaker yielded similar results: S_{voice} = 4.6 dB/octave and C_{voice} = 0.40 although the F0 range was significantly narrower than that observed in the expressive voice dataset.

3. SPEECH GENERATION AND PITCH IMPLANTATION FRAMEWORK

As a baseline we used the experimental MFS system reported in our previous work [9] with a neutral US English female voice dataset. The system was operated in the purely template mode. The encoding scheme applied to the template segments includes a sinusoidal representation [7] followed by the MRCC parameterization of the spectral envelope and phase spectrum [8] and residual noise components represented by random codebook entries and gains.

During synthesis, the above mentioned parameters can be re-sampled along time axis as required for duration modification and then converted to a frame-wise harmonic structures (a.k.a. line spectrum) corresponding to the desired F0 values. A pitch invariant noise component is then added to the harmonic structure. Alternatively, the harmonic structures can be first generated using the original F0 values and segment durations and then re-sampled along frequency and time axes according to the desired F0 and duration values. In any case the final line spectrum is scaled to preserve the original energy.

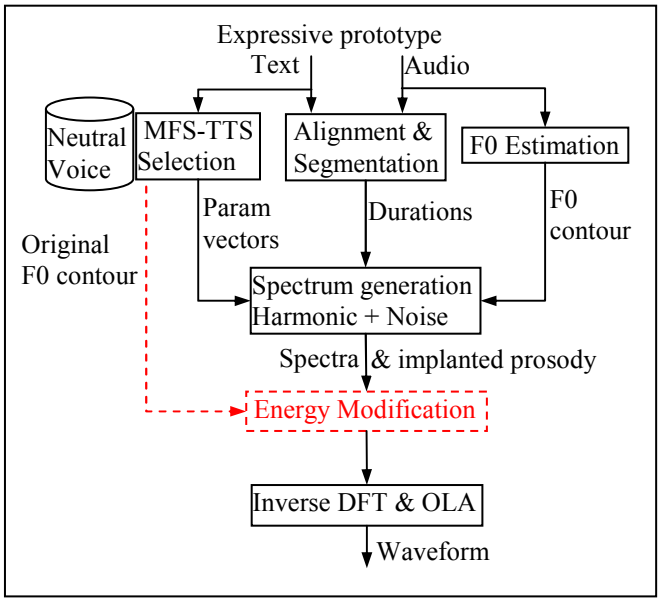

Figure 2. Block diagram of the external prosody implantation simulation framework. The Energy Modification block is included in the baseline system.

For simulation of expressive prosody implantation we used the Sports News recordings of the other female speaker mentioned in Section 2 as a source for input text messages and external prosody. These external recordings, hereafter referred to as *prototypes*, were aligned and segmented using the same phoneset and number of HMM states as the synthesis dataset. Hence one-to-one correspondence can be established at the segment level between a prototype and its TTS version synthesized with the neutral voice.

The simulation process is depicted schematically on Figure 2. The textual contents of a prototype is fed to the baseline TTS system which outputs a sequence of the frame-wise parameter vectors encoding the magnitude and phase spectra and the residual noise component. Then the segmental durations and the pitch contour extracted from the prototype are used to finalize the synthetic signal generation as described in the beginning of this Section.

As shown on the example in Figure 3, typically the external F0 contours significantly differ from the original one (inherited from the selected template segments). It leads to audible artifacts in speech synthesized by the baseline system.

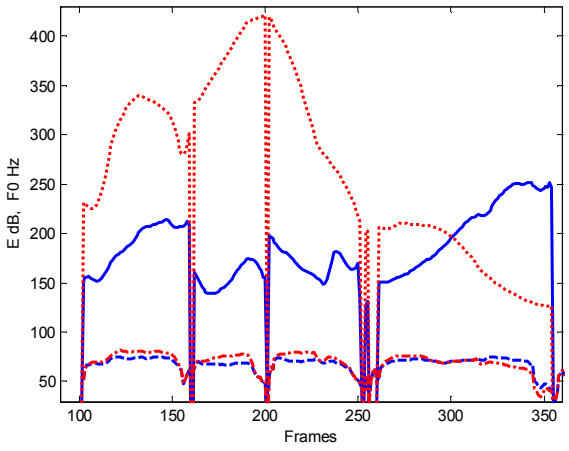

Figure 3. F0 contours: original – blue solid line; modified - red dotted line. Energy contours: original – blue dashed line; altered by the proposed algorithm – red dash-dotted line.

To enhance the baseline system we added the Energy Contour Modification block as shown on the diagram of Figure 2. This block denotes an algorithm that attempts to restore the balance between the energy and F0 at the frame level. It is obvious that the energy $E(k)$ measured at frame k according to the formulae (1) is coherent with original pitch frequency value $F0_{org}(k)$. The pitch-energy relationship suggests that the frame energy should be changed depending on the new value $F0_{out}(k)$ of the fundamental frequency at that frame. We adopt the log-linear representation of the relationship used in our study presented in Section 2. This yields a simple rule for the frame energy change:

$$\Delta E(k) = S_l \cdot \log_2 \frac{F_{out}(k)}{F_{org}(k)} dB \qquad (2)$$

where S_l dB/octave is a leaf dependent coefficient, l is the leaf associated with the segment containing the frame k. S_l can be estimated offline from the voice dataset as explained in Section 2. To speed up the implementation, we ignored the leaf dependent nature of the pitch-energy relationship and used in (2) a constant value S_l = 9 dB/octave selected by trials on few examples.

In order to assure the robustness of the energy contour altering against F0 estimation errors, the sequence of the frame-wise delta energy values (2) is smoothed as follows.

All frames being part of not fully voiced segments are declared nodes with zero ΔE values. For each fully voiced segment *s*, a segmental energy altering value ΔEs is calculated by averaging of the values given by (2) over the frames containing in the segment. The center of the fully voiced segments is declared a node with ΔEs value. Then the $\Delta E(k)$ value for frame *k* is re-calculated by linear interpolation between the nodes surrounding that frame. This process is illustrated by the example shown in Figure 4.

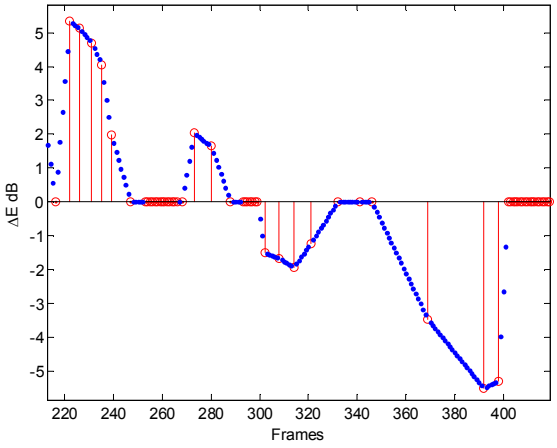

Figure 4. Smoothing of the energy altering values by interpolation. Red stems represent the node frames. Blue dots represent the interpolated energy altering values.

Then a scaling factor *F(k)* is calculated for each frame:

$$F(k) = \min[10^{\log_{10}(\Delta E(k)/20)}, \sqrt{E_{\max}/E_{org}(k)}] \qquad (3)$$

where E_{max} is the maximal original frame energy at the sentence level and $E_{org}(k)$ is the original energy of frame *k*. The limiting operation in (3) allows preserving the maximal frame energy and avoiding the waveform clipping. Finally the line spectrum and the residual noise component associated with frame *k* are multiplied by *F(k)* value. An example of the original and altered energy contours is shown in Figure 3.

4. SUBJECTIVE EVALUATION

12 expressive prototype sentences were re-synthesized in both versions using the framework described in Section 3: without and with the energy contour modification (referred to as "Baseline" and "EnrMod" correspondingly). 10 listeners (including 7 speech researchers) were presented with the randomized stimuli pairs. After listening to the both versions of stimuli, they were instructed to choose between five options: no preference, preference to either version or strong preference to either version.

The results are represented by the bar-chart in Figure 5. The evaluation revealed average preference of 51% including 14% of strong preference for the energy modification version and only 25% for the baseline version.

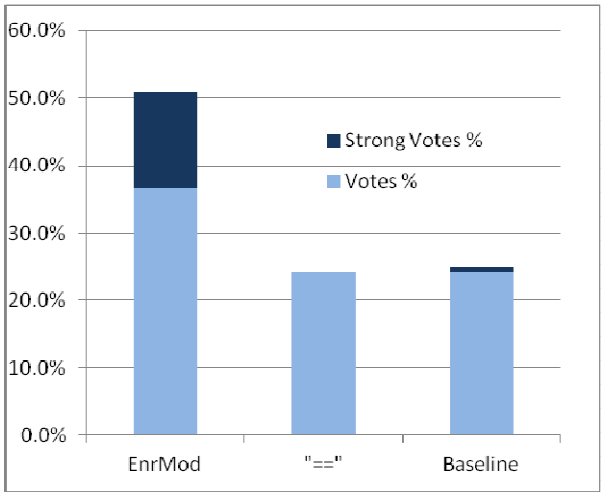

Figure 5. Results of the preference listening test.

5. CONCLUSIONS

Significant loss of the synthesized speech quality after extreme pitch contour modification poses a barrier to expressive speech synthesis. To alleviate this problem we proposed an approach based on the energy-pitch relationship phenomenon. The underlying idea is to restore the natural balance between the short-time energy and instantaneous pitch frequency destroyed by the pitch contour modification. We proposed a method for off-line speech class dependent estimation of the energy altering parameter and a robust method for the energy contour altering controlled by the pitch modification. We demonstrated the potential of the proposed approach in a subjective listening evaluation performed in a challenging setup where external extremely expressive prosody was implanted in speech synthesized with emotionally neutral voice.

In the evaluation we used a simplistic implementation ignoring the class dependent nature of the pitch-energy relationship. It is a reasonable assumption that the full-fledged leaf-adaptive implementation will yield much stronger speech quality enhancement.

The proposed approach can be adapted to TTS systems using non-parameterized speech and PSOLA-based pitch modification. To this end the energy-pitch log-linear regression can be estimated and applied to waveforms presumably in a pitch synchronous manner.

The proposed method of the coherent pitch and energy modification is complementary to spectrum shape transformation methods and can be combined with the latter for achieving the best possible quality of pitch modification.

6. REFERENCES

[1] V. Pollet, and A. Breen, "Synthesis by generation and concatenation of multiform segments", in Proc. Interspeech 2008, Brisbane, Australia, September 2008.

[2] K. Tanaka, and M. Abe, "A new fundamental frequency modification algorithm with transformation of spectrum envelope according to F0", In Proc. ICASSP 1997, Munich, Germany, April 1997.

[3] A. Kain, and Y. Stylianou, "Stochastic modeling of spectral adjustment for high quality pitch modification", in Proc. ICASSP 2000, Istanbul, Turkey, June 2000.

[4] P. Gramming, et al, "Relationship between changes in voice pitch and loudness", Journal of Voice, Vol 2, Issues 2, 1988, pp. 118-126, Elsevier, 1988.

[5] A. Rosenberg, and J. Hirshberg, "On the correlation between energy and pitch accent in read English speech", In Proc. Interspeech 2006, Pittsburgh, Pennsylvania, USA, September 2006.

[6] A. Sorin, T. Ramabadran, D. Chazan, R. Hoory, M. McLaughlin, D. Pearce, F. Wang and Y. Zhang, "The ETSI Extended Distributed Speech Recognition (DSR) standards: client side processing and tonal language recognition evaluation", In Proc. ICASSP 2004, Montreal, Quebec, Canada, May 2004.

[7] 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. ICASSP 2006, Toulouse, France, May 2006.

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

[9] A. Sorin, S. Shechtman and V. Pollet, "Uniform speech parameterization for multi-form segment synthesis", in Proc. Interspeech 2011, Florence, Italy, September 2011.