A Comparison of Prosody Modification using Instants of Significant Excitation and Mel-Cepstral Vocoder

Bajibabu B, Ronanki Srikanth, Sathya Adithya Thati, Bhiksha Raj, B Yegnanarayana, and Kishore Prahallad

***Abstract*—In this paper, we compare two methods for prosody (duration and pitch) modification. These are prosody modification using instants of Significant Excitation and Mel-Cepstral vocoder. We show that duration modifications are better using Mel-Cepstral vocoder for higher modification factor while pitch modifications are better using instants of Significant Excitations. In the end we show that Mel-Cepstral vocoder provides flexibility for non-uniform prosody manipulation.**

***Index Terms*—Prosody modification, Epoch extraction, Mel-cepstral vocoder.**

I. Introduction

THE objective of prosody modification is to change the pitch contour and durations of the sound units of speech without affecting the shapes of the short-time spectral envelopes . Such a process is useful in text-to-speech synthesis, voice conversion, expressive speech synthesis and speech rate modification [1]. The two main determining factors of a prosody modification method are computational speed and perceptual quality.

There exist several approaches in the literature for modifying prosody [2], [3]. These methods are broadly classified into either the time domain or frequency domain methods. Both domains have their respective pros and cons. The frequency-domain methods gain advantage from the fact that the signal to be modified need not be assumed quasi-periodic, whereas the time-domain methods rely heavily on the assumption of the nature of the signal and are generally more efficient in terms of computational load. Overlap and add (OLA), synchronous overlap and add (SOLA), and pitch synchronous overlap and add (PSOLA) methods are typical time-domain approaches. They operate directly on the time domain waveform to incorporate the desired prosody information. There are frequency domain methods such as the Phase-Vocoder method which operate in frequency domain. Methods like OLA, SOLA are limited to time scale modification where as PSOLA can be used for both time and pitch-scale modification. These methods directly modify the speech signal to achieve the desired prosody modification, which may lead to spectral or phase distortions. Recently a prosody modification method using Instants of Significant Excitation (epochs) was proposed [3]. This method operates on the linear prediction (LP) residual of signal and incorporates desired features by using the knowledge of epochs. Prosody modification in the residual domain is believed to reduce the spectral and phase distortions.

Most of the above methods are non-parametric, as they rely heavily on the speech production model and there is no explicit estimation of the model parameters. Other techniques have been proposed in which the parameters of a speech production model are estimated, and explicitly used in the modification/synthesis stages. The most straightforward of such approaches was the Linear Predictive Vocoder [4], but has now been abandoned because of its inability to provide high-quality modifications. Another such approach using Mel-Cepstral Vocoder represents a more promising approach because the estimated parameters are considered to be highly robust.

In this paper, we compare two methods for prosody modification: (a) prosody modification using epochs and (b) prosody modification using the Mel-Cepstral vocoder. To modify prosody, the former method manipulates the LP residual using the knowledge of epochs. The modified residual is used as the excitation signal. The latter method manipulates the parameters obtained from Mel-Cepstral analysis.

The paper is organized as follows: The underlying principles and a detailed description of both the methods are demonstrated in Section 2 and Section 3. The performance of these methods are demonstrated in Section 4 through a comparison of the methods with natural speech. The resulting benefits of the methods are discussed in Section 5 by various experiments. Finally, Section 6 summarizes the findings, conclusions and scope for future work.

II. Prosody modification using instants of significant excitation (epochs)

In this approach, LP analysis and synthesis method is used to incorporate desired prosody. This method makes use of the properties of the excitation source information for prosody modification. The residual signal in the LP analysis is used as an excitation signal. The successive samples in the LP residual are less correlated compared to the samples in the speech signal, will reduce the spectral and phase distortions. The residual signal is manipulated by using re-sampler either for increasing or decreasing the number of samples required

Bajibabu B, Ronanki Srikanth, Sathya Adithya Thati, B Yegnanarayana and Kishore Prahallad are with the Speech and Vision Lab of International Institute of Information Technology, Hyderabad, 500032, INDIA, e-mail:{bajibabu.b,srikanth.ronanki,sathya.adithya}@research.iiit.ac.in, {yegna,kishore}@iiit.ac.in

Bhiksha Raj is with Carnegie Mellon University, Pittsburgh, USA, e-mail: bhiksha@cs.cmu.edu

for the desired prosody modification. There are four main steps involved in the prosody modification using epochs [3].

1) Deriving the instants of significant excitation (epochs) from the LP residual signal.
2) Deriving a modified (new) epoch sequence according to the desired prosody (pitch and duration).
3) Deriving a modified LP residual signal from the modified epoch sequence.
4) Synthesizing speech using the modified LP residual and the LPCs.

The performance of this method depends upon the accuracy to detect the exact locations of instants of significant excitation.

A. Method to extract instants of significant excitation

A method was proposed in [6] to extract instantant of significant excitation from a speech signal. The method uses the zero-frequency filtered (ZFF) signal derived from speech to obtain the instants of significant excitation of the vocal tract system. Performance of ZFF method is significantly better compared to other methods. The following steps are involved in processing the speech signal to derive the instant of significant excitation.

1) Difference the speech signal $s[n]$ to remove any very low frequency component introduced by the recording device.
$$x[n] = s[n] - s[n-1]. \tag{1}$$
2) Pass the differenced speech signal $x[n]$ through a cascade of two ideal zero-frequency resonators. i.e.
$$y_0[n] = -\sum_{k=1}^{4} a_k y_0[n-k] + x[n], \tag{2}$$
where $a_1 = -4, a_2 = 6, a_3 = -4$ and $a_4 = 1$.
3) Compute the average pitch period using the autocorrelation function for every 30 ms speech segments.
4) Remove the trend in $y_o[n]$ by subtracting the local mean computed over a window obtained from (c) at each sample. The resulting signal $y[n]$ is the zero-frequency filtered signal, given by
$$y[n] = y_0[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_0[n+m]. \tag{3}$$
Here $2N+1$ corresponds to the number of samples in the window used for mean subtraction. The choice of the window size is not critical as long as it is in the range of one to two pitch periods.
5) The instants of positive zero crossings of the filtered signal give the locations of the instants of significant excitation.

B. Prosody modification

Prosody modification involves deriving a new residual signal by incorporating the desired modification in the pitch period and duration for the utterance. This is done by first creating a new sequence of epochs from the original sequence of epochs. For this purpose, all the epochs derived from the original signal are considered, irrespective of whether they belong to a voiced segment or unvoiced segment. The methods for creating the new epoch sequence for the desired modification are discussed in detail in [3].

After obtaining the modified epochs, the next step is to derive the excitation signal of LP residual. For this, the original epochs closest to the modified epochs are determined. The residual samples around the original epoch are placed starting from the corresponding new epoch. Since the value of the desired epoch interval is different from the value of the corresponding original epoch interval, it is necessary to either delete some residual samples or append some new residual samples to fill the new epoch interval. Deletion of required number of residual samples is made in the tail portion of the selected residual samples. Insertion of required number of residual samples is achieved by suitably re-sampling about 10% of the tail portion of the selected residual samples and appending them to the end.

C. Generating the synthetic signal

The modified LP residual is used as an excitation signal for the time varying all-pole filter. The filter coefficients are updated for every X samples, where X is the number of samples corresponding to the frame shift that is used for performing the LP analysis. In this method, a frame shift of 5 ms and a frame size of 20 ms is used for LP analysis. Thus, the samples correspond to 5 ms when the prosody modification does not involve any duration modification. On the other hand, if there is a duration modification by a scale factor β, then the filter coefficients (LPCs) are updated for every X samples corresponding to 5β ms.

III. Prosody modification using Mel-Cepstral vocoder

The prosody modification makes use of both source and system information. Here, the system information is obtained from the Mel-Cepstral coefficients (MCEPs) and the source information from the fundamental frequency ($F0$). In the current state of art, MCEPs are the robust features widely used in speech recognition and synthesis. MCEPs and $F0$ values are manipulated according to the desired prosody. This method involves:

- Extraction of MCEPs and $F0$ values from the given speech signal.
- Modification of MCEPs and $F0$ values according to the desired prosody.
- Synthesize speech using MLSA [8] filter with the modified parameters.

A. Parameters extraction

1) MCEPs extraction: For obtaining MCEPs, several methods have been proposed. By using recursion formulae, MCEPs are calculated from LP coefficients by using the technique of spectral re-sampling [7]. We followed the standard method [5] to extract MCEPs from a given speech signal. In this method, frame size of 25ms and frame shift of 5ms is used. Assuming

$x(n)$ to be a frame of speech, the cepstrum $c(m)$ of a segment $x(n)$ is defined as

$$c(m) = \frac{1}{2\pi j}\oint_C \log X(z) z^{m-1}\,\mathrm{d}z \tag{4}$$

$$\log X(z) = \sum_{m=-\infty}^{\infty} c(m) z^{-m} \tag{5}$$

where $X(z)$ is the z-transform of $x(n)$, and C is a counterclockwise closed contour in the region of convergence of $\log X(z)$ and encircling the origin of the z-plane.

Frequency-transformed cepstrum, so-called mel-cepstrum $\tilde{c}(m)$, is defined as [8]

$$\tilde{c}(m) = \frac{1}{2\pi j}\oint_C \log X(z) \tilde{z}^{m-1}\,\mathrm{d}\tilde{z} \tag{6}$$

$$\log X(z) = \sum_{m=-\infty}^{\infty} \tilde{c}(m) \tilde{z}^{-m} \tag{7}$$

where

$$\tilde{z}^{-1} = \Psi(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, |\alpha| < 1 \tag{8}$$

The phase response $\tilde{\omega}$ of all-pass system $\Psi(e^{j\omega}) = e^{-j\tilde{\omega}}$ is given by

$$\tilde{\omega} = \beta(\omega) = \tan^{-1}\frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha} \tag{9}$$

Thus, evaluating Eqn-(6) and Eqn-(7) on the unit circle of the z̃ plane, we see that $\tilde{c}(m)$ is the inverse Fourier transform of $\log X(e^{j\omega})$ calculated on a wrapped frequency scale $\tilde{\omega}$. The phase response $\tilde{\omega} = \beta(\omega)$ gives a good approximation to auditory frequency scale with an appropriate choice of α. In this fashion, we extract MCEPs for a speech signal.

2) Pitch extraction: There are a number of standard methods that are used to extract $F0$, based on various mathematical principles. Among them the most widely used techniques are auto correlation method in time domain and cepstral analysis method in frequency domain. Here in this paper,we used the cepstral analysis method to extract F0.

B. Prosody modification

1) Duration modification: For duration modification, new MCEPs and F0 sequences are generated to get the desired duration modification factor. The following procedure explains the duration modification by a factor α.

- Multiply the factor by the 100 to get a whole number $X = \alpha * 100$
- Find the gcd: $Y = \gcd(X, 100)$
- $P = \frac{X}{Y}$ and $Q = \frac{100}{Y}$

To increase the duration, after every Q frames, last $P - Q$ frames are repeated whereas to decrease the duration, for every Q frames, $Q - P$ frames are removed. In this way, one can modify the duration by any factor. Modification factors from 0.5 to 2 give good quality and intelligible speech. The above procedure explains the duration modification. In this case, MCEPs and $F0$ value constitute each frame. The logic behind this procedure can be explained intuitively that when a person speaks very fast/slow, he or she has to change the vocal tract configuration very quickly/slowly and in this process, number of frames for each configuration decreases/increases.

TABLE I
RANKING USED FOR JUDGING THE QUALITY AND DISTORTION OF THE SPEECH SIGNAL FOR DIFFERENT MODIFICATION FACTORS.

Rating	Speech Quality	Level of distortion
1	Unsatisfactory	Very annoying and objectionable
2	Poor	Annoying but not objectionable
3	Fair	Perceptible and slightly annoying
4	Good	Just perceptible but not annoying
5	Excellent	Imperceptible

2) Pitch modification: The objective is to generate a new $F0$ sequence based on the desired pitch modification factor. For this sequence, only voiced regions are considered as $F0$ does not make sense in unvoiced and silence regions. If the pitch values are to be modified by a factor β, then $F0$ values are multiplied by the modification factor β to generate the new $F0$ sequence.

C. Generating the synthetic signal

Finally, speech waveform is synthesized directly from the modified MCEPs and $\log F0$ values using the Mel Log Spectrum Approximation (MLSA) filter with binary pulse or noise excitation [8]. In MLSA filter MCEPs are used to generate the filter coefficients, $\log F0$ values are used to generate the excitation signal. A major limitation of the Mel-Cepstral vocoder is that the synthesized speech is buzzy since it uses a simple binary pulse or noise for excitation.

IV. EVALUATION

A perceptual test was conducted to assess the extent to which the transformed speech is perceived as having the intended expressivity. Perceptual evaluation was carried out by conducting subjective tests on 25 research scholars in the age group of 20-30 years. The subjects have sufficient speech knowledge for proper assessment of the speech signals. Four sentences were chosen, two from BDL(male) and two from SLT(female) speakers from the $ARCTIC$ database to perform the test. For each sentence the pitch period was modified by a factor with keeping duration unchanged. Similarly, the duration was modified by a factor with keeping pitch unchanged by using both methods. After the modification, the filenames were encrypted to avoid bias towards a specific method.

The tests were conducted by playing the speech signal through headphones. In the test, the subjects were asked to judge the naturalness, distortion and quality of the speech for various modification factors on a five point scale given in table 1. The mean opinion scores (MOSs) for each of the pitch period modification and duration modification are given in the table 2.

TABLE II
MEAN OPINION SCORES FOR DIFFERENT PITCH PERIOD MODIFICATION FACTORS AND DURATION MODIFICATION FACTORS.

Duration modification factor(α)	Pitch period modification factor(β)	Mean Opinion Score (MOS)	
		Epoch Method	Mel-cepstral Vocoder
0.5	0.5	2.33	3.52
0.5	1	3.88	4.19
1.5	1	3.81	4.22
1.5	2	3.74	3.76
1	0.5	3.63	3.39
1	1.5	4.26	3.96

A. Results

Table I, shows ranking used for judging the quality and distortion of the speech signal. Table II, illustrates the MOS scores for different modification factors.

The Mean Opinion Score (MOS) for each of the pitch period and duration modification factors are given in Table 2. For moderate modification factors, both the methods seem to provide the best possible speech quality. For higher modification factors the Mel-cepstral vocoder provides the better quality than epoch method because it is not operating on signal directly. From the scores it is observed that for all duration modification factors the Mel-cepstral vocoder provides better quality, For all pitch modification factors, epoch method provides better quality. The reason for this is epoch method follows the pitch synchronous prosody modification. The corresponding waveforms can be found on website “http://web.iiit.ac.in/˜ronanki/evaluation.html”

V. NON-UNIFORM DURATION MODIFICATION

The above mentioned methods modify the duration of a speech signal uniformly which may not be the case in a natural conversation. Many speakers speak at different rates in a single utterance which aids them in expressing emotions. Expressive speech contains different speaking rates varying non-uniformly with context to express different kinds of emotions. So, we study the literature on duration analysis of different speaking rates, synthesized speech sounds with different speaking rates by using Mel-Cepstral vocoder.

Studies on duration analysis of different speaking rates showed that the durations of prosodic words were significantly different for three speaking rates (slow, normal, and fast) with systematic increase/decrease in syllable durations when the speaking rate was decreased/increased [9]. For each of the voiced, unvoiced and silence regions, the percentage deviation of duration is computed when speaking rate is changed from normal to fast, or normal to slow. The details of percentage deviation of duration are given in Table III [10]. Negative sign of mean indicates decrease in duration and positive sign indicates increase in duration.

TABLE III
PERCENTAGE DEVIATION IN THE DURATIONS OF SPEECH SEGMENTS FOR DIFFERENT SPEAKING RATES COMPARED TO THE NORMAL SPEECH [10]

Speech Segment	Normal to fast		Normal to slow	
	μ	σ	μ	σ
voiced	-22.71	8.99	37.67	20.88
unvoiced	-26.90	25.68	51.64	50.24
silence	-16.49	10.75	40.79	40.66

TABLE IV
COMPARISON OF UNIFORM AND NON-UNIFORM DURATION MODIFICATION: MEAN OPINION SCORES(MOS)

Duration modification factor(α)	Uniform Modification	Non-Uniform Modification
0.63	3.53	3.58
0.85	4.10	4.12
1.5	4.00	4.22
1.8	3.80	4.00

A few sentences were recorded and then manually labeled them as voiced, unvoiced and silenced regions. Non-uniform duration modification using Mel-cepstral vocoder method as discussed in Section 3 was performed on different segments of speech for different modification factors which were derived from Table III. A perceptual evaluation test was conducted on both the methods and results are tabulated in Table IV. It is observed that for normal to fast case, both uniform and non-uniform modifications yielded closer scores, but in the case of normal to slow, non-uniform modification yielded better scores.

VI. SUMMARY AND CONCLUSIONS

In this study, we compared two methods for prosody modification. The methods are based on state-of-the-art source filter algorithms for prosody modification. In subjective evaluations, the two methods resulted in similar performances for lower modification factors whereas for higher modification factors Mel-Cepstral vocoder performed better in subjective evaluations. But epoch based method has better synthesized speech quality than Mel-Cepstral vocoder. Later in this paper, we studied on how the durations varies for different types of speaking rate.

From the study of different speaking rates, we derived different modification factors for different segments of speech and using those factors, incorporated non-uniform duration modification by Mel-Cepstral vocoder method. Our future plans include development of a robust algorithm that incorporate the prosody as in natural speech. We are also planning to integrate this to Statistical Parametric Synthesis systems as it uses Mel-Cepstral vocoder for speech synthesis.

ACKNOWLEDGEMENT

A part of this work was done as a part of Winter School, Dec 8-21, 2010 held at IIIT-Hyderabad, India. We are thankful to the people who participated in the evaluation.

REFERENCES

[1] D. G. Childers, K. Wu, D. M. Hicks and B. Yegnanarayana, “Voice conversion”, Speech Commun., 8:147-158, June 1989.

[2] E. Moulines and J. Laroche, “Non-parametric techniques for pitch-scale and time-scale modification of speech”, Speech Commun., 16:175-205, June 1995.

[3] K. S. Rao and B. Yegnanarayana, “Prosody modification using instants of significant excitation”, IEEE Trans. Audio, Speech, Language Proc., 14(3):972-980, May 2007.

[4] B. Atal and S. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. J. Acoust. Soc. Am., vol. 50 (2), pp. 637-655, Aug 1971.

[5] S.Imai, "Cepstral analysis and synthesis on the Mel-frequency scale", ICASSP, Boston, 1983.

[6] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals", IEEE Trans. Audio, Speech, Language Proc., 16(8):1602-1613, Nov 2008.

[7] Keiichi Tokuda, Takao Kobayashi, and Santoshi Imai, "Recursive Calculation of Mel-Cepstrum from LP Coefficients",Trans. IEICE, vol. J71-A, pp.128-131, Apr 1994.

[8] S.Imai, K.sumita, C.Furuichi, "Mel log spectral approximation filter for speech synthesis", Trans.IECE, Vol. J66-A, pp.122-129, Feb 1983.

[9] R. H. Kessinger and S. E. Blumstein, "Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies," J. Phonetics, vol. 26, no. 2, pp. 117-128, Apr 1998.

[10] Sri Harish Reddy M. and B. Yegnanarayana, "Incorporation of Excitation Source and Duration Variations in Speech Synthesized at Different Speaking Rates", in Speech prosody 2010, Chicago, USA, May 2010.