

Detection on PSOLA-modified Voices by Seeking out Duplicated Fragments

Yifeng Shen, Jia Jia, Lianhong Cai

Key Laboratory of Pervasive Computing, Ministry of Education

Tsinghua National Laboratory for Information Science and Technology (TNList)

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract—Pitch Synchronous Overlap-Add (PSOLA) refers to a family of signal processing techniques that are widely used for prosodic modification. They can be used to change one person’s voice by altering time-scale and/or pitch-scale characteristics of speech, making the voice unrecognizable or unidentifiable. Well-modified voices may make speaker recognition process, which is critical in digital audio forensic framework, out of work.

Time-domain PSOLA (TD-PSOLA) is the most popular algorithm in PSOLA family. Time- and pitch-scaling form of modification can be applied by TD-PSOLA, and the synthesis quality is extremely high provided that the modifications do not exceed a factor of two. Our paper presents a simple method to find out whether a given speech waveform is modified by TD-PSOLA, and further estimate the modification magnitude on the time/pitch scale. Seeking out duplicated fragments from the time domain of the waveform, we extract the occurrence number of duplicated fragments as well as occurrence frequency in voiced portions of speech. A single feature (duplicated fragments density, DFD) is then calculated, and compared with a threshold (obtained from plenty of former statistic results) to decide whether the speech waveform is modified or not. Experimental results show the effectiveness of our method in detecting modified voices, which are pitches heightened and/or duration lengthened using the TD-PSOLA algorithm.

Keywords—PSOLA, Duplicated Fragments, Speech Processing, Digital Audio Forensic

I. INTRODUCTION

Prosodic modification refers to a class of techniques that can change voice by altering time-scale, pitch-scale or spectral-scale characteristics of the speech waveform. Several models and approaches are available in prosodic modification, like pitch synchronous overlap-add (PSOLA)[2][4][7], sinusoidal model, harmonic plus noise model (HNM), and speech transformation and representation using adaptive interpolation of weighted spectrum (STRAIGHT)[8][9].

Well-modified voices may mask the original voice’s characteristics, such as formant position, formant width, fundamental frequency, duration, subband energy distribution and spectral contour. This makes the speaker recognition process, which is critical in digital audio forensic, out of work.

On the other hand, digital audio forensic becomes more and more important. Both digital audio evidences in law court and online music business are in urgent need for detecting the integrity and authentication of digital audios. Technologies like

the electrical network frequency (ENF) criterion [13][14], butt-splice tamper detection [16] and MP3 files detection [15] have been proposed in the literature. However, to our best knowledge, no methods about detection on prosodic modified voices have been presented.

In this paper, we present a simple method to detection prosodic modified voices by time-domain PSOLA (TD-PSOLA) algorithm, which is the most popular technique in the PSOLA family. TD-PSOLA could modify the voice in the time-scale and pitch-scale simultaneously or separately, changing the voice’s pitch and/or duration. As there may have duplicated frames in the synthetic voice when lengthening the duration or heightening the pitch frequency, we utilize this property to carry out the detection method. Duplicated fragments are defined and extracted from the time domain of the speech waveform. We then calculate the occurrence number of duplicated fragments as well as occurrence frequency in voiced portions of speech. A single feature (duplicated fragments density, DFD) is then calculated, and compared with a Threshold (obtained from plenty of experiment statistic results) to decide whether the speech waveform is modified or not.

The rest of this paper is organized as follows. Section 2 briefly introduces the TD-PSOLA algorithm. In section 3, we define a single feature - duplicated segment frequency, and present a method based on this feature to detect the TD-PSOLA modified voices. Experimental evaluation and detailed results analysis are given in section 4. In the end, section 5 discusses the conclusions and feature work.

II. TD-PSOLA ALGORITHM

The TD-PSOLA algorithm can be used to performing modification on the pitch-scale and/or duration-scale of speech. The synthesis quality is extremely high provided that the modifications do not exceed a factor of two. TD-PSOLA proceeds in three steps:

A. Analysis

The analysis process consists of decomposing the speech waveform $x(n)$ into a stream of short-time signals, which are called as frames. These frames are obtained by multiplying the signal waveform $x(n)$ by a series of analysis windows $h(t)$, as follows:

$$x_m(n) = h_m(t_m - n)x(n) \quad (1)$$

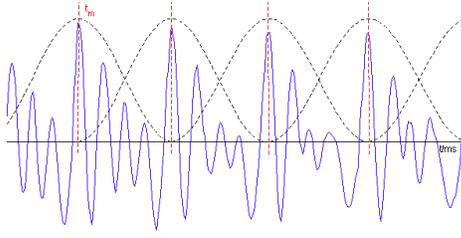


Figure 1. A diagram of pitch-synchronous analysis

$x_m(n)$ is the m -th frame while $h_m(t_m)$ is the m -th analysis window at time t_m . The analysis time instants t_m are set at the exact pitch epoch positions on the voiced portions of speech and a constant rate on the unvoiced portions. The analysis window is generally chosen to be a symmetrical Hanning window. The window length is proportional to the local pitch period. The proportion factor range generally 2~4, causing overlaps between successive analysis windows.

B. Modification

The objective of modification is to transform the stream of frames got in step 1 to a stream of synthesis frames, synchronized on a new set of synthesis time instants t_s . A mapping $t_m \rightarrow t_s$ between the synthesis time instants and analysis time instants is determined according to the desired time-scale and/or pitch-scale modification.

Time-scale modification includes lengthening and shortening manipulations, and it can be performed without reference to pitch. Lengthening is achieved by duplicating certain frames. For instance, when the time-scale modification factor is 1.5, then each 1 of 2 successive frames should be duplicated. Shortening, on the other hand, is achieved by removing certain frames.

Pitch-scale modification is a bit more complex than time-scale modification, as it interacts with duration. In order not changing the number of pitch periods, we perform pitch-scale modification with time-scale compensation. When the pitch-scale modification factor is β (the pitch period is $1/\beta$ times the original), we have to perform time-scale compensation with a factor of β to keep the number of pitch periods. Duplicated frames are added to the stream of synthesis frames when $\beta > 1$. Time-scale and pitch-scale modifications can be manipulated simultaneously or asynchronously.

C. Reconstruction

The reconstruction is done throughout combining the stream of synthesis frames into a synthetic speech waveform. The synthesis formula can be written as:

$$\tilde{x}(n) = \frac{\sum_s a_s \tilde{x}_s(n) h_s(t_s - n)}{\sum_s h_s^2(t_s - n)} \quad (2)$$

Here, $\tilde{x}(n)$ is the synthetic speech; a_s is a time-vary normalization factor compensating for energy modifications, and $h_s(t)$ is synthesis window. TD-PSOLA seeks to minimize

the error between the synthesis frames and the synthetic signal. Least-square method is applied to obtain the synthetic signal. Successive frames are overlap-added with appropriate weights and time-shifts.

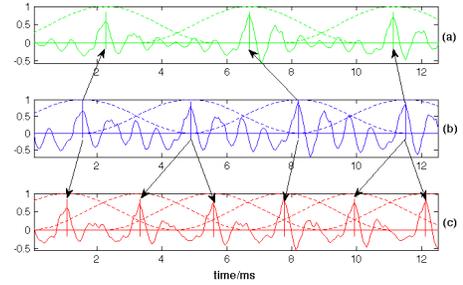


Figure 2. Pitch-scale modification using TD-PSOLA. (a) Modified speech signal with a pitch-scale modification factor 0.75. (b) A segment of voiced speech (original). (c) Modified speech signal with a pitch-scale modification factor 1.5.

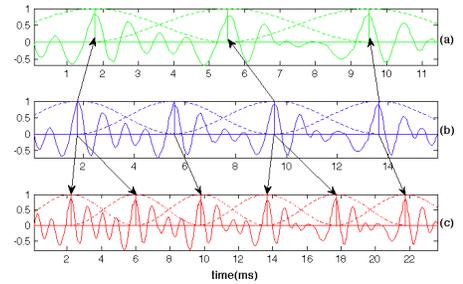


Figure 3. Time-scale modification using TD-PSOLA. (a) Modified speech signal with a time-scale modification factor 0.75. (b) A segment of voiced speech (original). (c) Modified speech signal with a time-scale modification factor 1.5.

III. DETECTION ALGORITHM USING DUPLICATED FRAGMENTS

A. Feature Definition and Extraction

As seen in Fig.2 and Fig.3, when using TD-PSOLA algorithm to lengthen the speech duration or heighten the pitch frequency, duplicate frames appear in the stream of synthetic frames. In fact, let the time-scale modification factor be γ , the pitch-scale modification factor β , duplicate frames will occur when $\gamma \cdot \beta > 1$. However, due to the affection of overlap-addition and synthesis window modulation, the duplicate frames are not “perfect replications” to each other.

To be convenient, we first define some symbols:

N : length of a frame; N_i : length of i -th frame.

P : pitch epoch position of a frame; P_i : pitch epoch position of i -th frame.

Here’s an example of pitch-scale modification procedure in Fig.4. The modification factor is 1.5 that introduces duplicate frames. Fig.4(a) is a segment of an original speech waveform

with 3 pitch epochs. Three analysis windows are applied with a length of 2 times the local pitch period. Fig.4(b) shows the successive three analysis frames extracted from the original waveform. Fig.4(c) is the synthetic speech waveform with the 2nd analysis frame duplicated. So the 2nd and 3rd synthetic frames are duplicate frames. Fig. 4(d) shows the timbre contours in ranges $[P_2 - \frac{N}{4}, P_2 + \frac{N}{4}]$ of 2nd synthetic frame (in red) and $[P_3 - \frac{N}{4}, P_3 + \frac{N}{4}]$ of 3rd (in blue). Fig.4(e) shows the difference ratio of the two timbre contours point by point in Fig.4(d). The difference ratio of two sample points is defined as:

$$\text{diff_ratio}(i) = \frac{|x_1(i) - x_2(i)|}{\max(\max_i\{|x_1(i)|\}, \max_i\{|x_2(i)|\})} \quad (3)$$

Where $x_1(i)$ and $x_2(i)$ stand for the i -th sample points of the previous frame and the latter one. We can conclude from Fig.4(e) that the maximum value of difference ratio is about 5% when the sample points are less than $N/8$ away from the pitch epoch, which means the two timbre contours within the ranges $[P - \frac{N}{8}, P + \frac{N}{8}]$ in each frame are almost “duplicated copies” to each other.

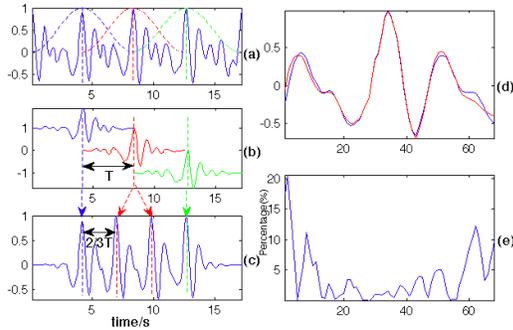


Figure 4. An example of pitch-scale modification procedure.

As is said in section 2, the analysis window may vary 2~4 times the length of local pitch period, we evaluate the window length with the proportion factor 2, 3 and 4. Table 1 lists the results of maximal values of the difference ratios from different sample point range under the determined analysis window lengths. For example, the value in row 2, column 5 means the maximal difference ratio in range $[P - \frac{N}{16}, P + \frac{N}{16}]$ is 2.49% when the analysis window is 2 times length. The value in row 3, column 4 means the maximal difference ratio in range $[P - \frac{N}{8}, P + \frac{N}{8}]$ is 4.03% when the analysis window is 3 times length. We can conclude that the maximal difference ratios are all less than 8% on condition that the range is no border than

$[P - \frac{3N}{16}, P + \frac{3N}{16}]$, which means the timbre contours of the two frames in range $[P - \frac{3N}{16}, P + \frac{3N}{16}]$ are almost identical.

We define such timbre contours around the pitch epoch that are almost identical to each other in two successive frames as duplicated fragments.

We now propose a measurement method to decide whether two successive frames have duplicated fragments. The similarity of two fragments in range $[P - Rad, P + Rad]$ can be calculated as:

$$\text{Similarity}(x_1, x_2) = 1 - \frac{1}{2 \cdot Rad + 1} \sum_{i=P-Rad}^{P+Rad} \text{diff_ratio}(i) \quad (4)$$

Where x_1 and x_2 stand for the corresponding sample points of two successive frames. The two fragments are judged as duplicated when $\text{Similarity}(x_1, x_2) > \text{Threshold}$. As is discussed, we shall keep the fragments not to exceed the range $[P - \frac{3N}{16}, P + \frac{3N}{16}]$, that means,

$$[P - Rad, P + Rad] \subseteq [P - \frac{3N}{16}, P + \frac{3N}{16}], \Rightarrow Rad \leq \frac{3N}{16}.$$

The exact values of the two parameters Rad and $Threshold$ require more experimental statistics and will be discussed later.

TABLE I. AFFECTION OF ANALYSIS WINDOW LENGTH AND FRAGMENT RANGE ON THE MAXIMAL VALUE OF DIFFERENCE RATIO

Proportion factor of Analysis Window	Fragment Range Factor (Fragment Length = Fragment Range Factor * Analysis Window Length)			
	1/4	3/16	1/8	1/16
2	20.55%	7.99%	5.04%	2.49%
3	23.65%	8.18%	4.03%	2.01%
4	14.50%	8.06%	3.37%	1.96%

B. Our Method

We present a simple method to detect TD-PSOLA modified voices by using the defined feature duplicated fragments. The flow chart of our method is shown in Fig.5. Pitch detection and voiced/unvoiced decision algorithm [11] is first applied to the speech waveform of questioned voice. The speech waveform is then segmented into voiced and unvoiced portions. Duplicated fragments are searched frame by frame in the voiced portions of speech. (或者写成: Local peak positions are then marked on the voiced portions of speech. Duplicated fragments are searched over the list of local peak positions. Duplicated fragments 的定义是否写成在峰值周围的段落相似, 而不是目前从 frame 出发的定义???) The total number of duplicated fragments ($num_{dup_fragments}$) is counted with the total length of

voiced portions in seconds ($length_{voiced_portion}$). A single feature named duplicated fragments density (DFD) is then extracted using the following expression:

$$DFD = \frac{num_{dup_fragments}}{length_{voiced_portion}} \quad (5)$$

The single value DFD is then used to discriminate the original voices and TD-PSOLA modified voices.

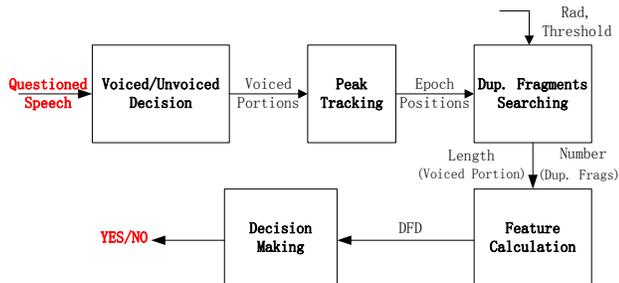


Figure 5. Flow chart of the proposed method.

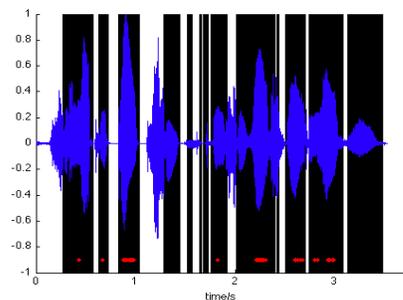


Figure 6. An example of extracting duplicated fragments from a pitch-scale modified voice with factor 1.5. The black areas stand for voiced sections, while the red points are marked as appearances of duplicated fragments.

IV. EXPERIMENTS

In this section, we carry out experiments from different perspectives to evaluate our method. We investigate the two parameters Rad and $Threshold$ to find the optimal choice. Time-scale, pitch-scale modifications with TD-PSOLA algorithm are applied separately and simultaneously over a small data set, and optimal values Rad and $Threshold$ are selected over all the three conditions. We then use these optimal parameters to experiment on a large data set.

A. Experiment Setup

The very famous TIMIT database [12] is chosen for our experiment. It contains 6300 speech clips recorded by different persons all around the world, with each clip's length 5s-10s. The audio format of all speech clips in the database is 16KHz/16bit. We randomly choose 10 persons (5 male, 5 female) from the TIMIT database, with each person 10 speech clips. So we totally have 100 speech clips, 50 male and 50

female, consisting a small data set. Followed experiments are all based on this small data set.

B. Evaluation and Results

1) Pre-estimation of parameters

As is written in literature, human's pitch frequency varies from 40Hz to 600Hz. With the sampling rate of speech in 16KHz, we can get that each pitch period has 27~400 sample points. Further on, as the analysis window length is 2~4 times the pitch period's, we get one frame's length N ranges in 54~1600. As Rad satisfies $Rad \leq 3N/16$, we have $Rad \leq 10$.

We estimate the second parameter $Threshold$ from (4). As the maximal difference ratio is less than 8% in the range

$$\left[P - \frac{3N}{16}, P + \frac{3N}{16} \right],$$

we get that:

$$Similarity(x_1, x_2) > 1 - \frac{1}{2 \cdot Rad + 1} \sum_{i=P-Rad}^{P+Rad} 0.08 = 0.92$$

To make the parameter distinguishable in our detection task, we take $Threshold \geq 0.95$.

2) Time-scaling Detection

The time-scaling factor t is set within a list of values 0.5, 0.7, 0.9, 1.2, 1.4, 1.6, 1.8, and 2.0. Each speech clip in the small data set is modified in time-scale with these 8 different factors, achieving 8 modified speech clips. So it becomes 100 original speech clips and 800 time-scale modified speech clips, with each time-scaling factor 100 clips. Our method is operated on all the 900 speech clips and each speech clip finally comes out with a single feature DFD.

Distributions of DFD values extracted from the 800 modified speech clips and 100 originals are drawn on Fig.7, with $Rad = 10, Threshold = 0.97$. Results show that the DFD values extracted from modified clips with $t < 1$ are smaller than those from original, while the DFD values extracted from modified clips with $t > 1$ are notably larger. The mean values of DFD in each bin are drawn in a green line, which grows rapidly when $t > 1$.

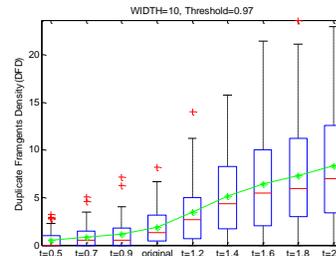


Figure 7. Distributions of DFD with different time-scaling factor t .

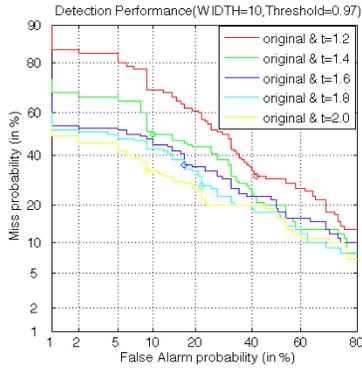


Figure 8. ROC curves of our method's performance in detection time-scale modified voices.

Fig.8 shows five detect error trade-off (DET) curves of our method's performance in detection time-scale modified voices. Each curve stands for the detection performance on a mixed data set, which has 200 speech clips with 100 original and 100 modified with factor t (one of 1.2, 1.4, 1.6, 1.8 and 2.0). Results show that with the growth of factor t , our method's performance is getting better, which means it's more discriminative between modified voices with large time-scale factor and original voices than those with small time-scale factor. The detection cost function is used as the criterion:

$$DCF = C_{fr} * P_{fr} * P_{true} + C_{fa} * P_{fa} * (1 - P_{true}) \quad (6)$$

We set the false reject cost $C_{fr} = 1$, the false alarm cost $C_{fa} = 1$, and the prior true probability $P_{true} = 1/2$. Optimal positions with minimal detection cost are marked out using small circles; it's with 0.33 false reject probability and 0.09 false alarm probability at the optimal position of the yellow curve, which the data set is mixed of 100 original speech clips and 100 with time-scaling factor 2.0.

Table 2:

Values of Threshold	Values of Rad			
	5	6	7	10
0.95	(0.35,0.34)	(0.31,0.37)	(0.36,0.34)	(0.35,0.34)
0.96	(0.29,0.35)	(0.36,0.29)	(0.32,0.33)	(0.37,0.28)
0.97	(0.32,0.28)	(0.35,0.25)	(0.40,0.19)	(0.49,0.09)
0.98	(0.37,0.14)	(0.38,0.11)	(0.38,0.09)	(0.32,0.13)
0.99	(0.39,0.08)	(0.40,0.07)	(0.41,0.04)	(0.45,0.01)

In table 2, we establish the detection performance with different values of parameter pair $(Rad, Threshold)$. Each detection is operated on a mixed data set, which has 600 speech clips consisting of 100 original and 500 modified (t goes over 1.2, 1.4, 1.6, 1.8 and 2.0). Without loss of generality, we set the prior true probability $P_{true} = 1/6$, the false reject cost $C_{fr} = 5$ and the false alarm cost $C_{fa} = 1$. Minimal detection costs are found and the corresponding values (P_{fr}, P_{fa}) are filled in to cells of the table.

3) Pitch-scaling Detection

The pitch-scaling factor p is set within a group of values 0.5, 0.7, 0.9, 1.1, 1.2, 1.3, 1.4, and 1.5. All experiments carried out here are the same as in "Time-scaling Detection" section.

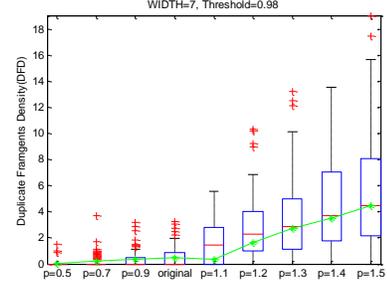


Figure 9. Distributions of DFD with different pitch-scaling factor p .

DFD Distributions extracted from the 900 speech clips are drawn on Fig.9, with $Rad = 7, Threshold = 0.98$. Results show that the DFD values with $p < 1$ are smaller than the originals, while the DFD values with $p > 1$ are notably larger. The mean values of DFD in each bin are lined in green, which grows rapidly when $p > 1$.

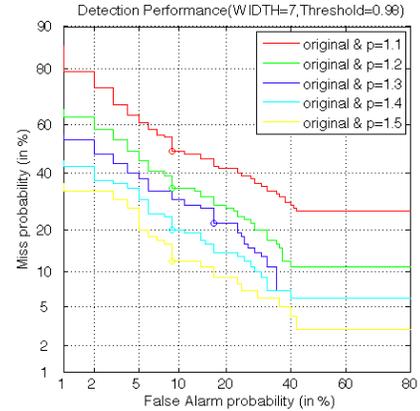


Figure 10. ROC curves of our method's performance in detection pitch-scale modified voices.

Fig.10 shows the DET curves of our method's performance in detection pitch-scale modified voices. Each curve stands for the detection performance on a mixed data set, which has 200 speech clips with 100 original and 100 modified with factor p (one of 1.1, 1.2, 1.3, 1.4 and 1.5). Results show that our method's performance gets better with the growth of p . The detection cost function is also used here with same preset. Optimal positions are marked out using small circles. For example, it's with 0.12 false reject probability and 0.09 false alarm probability at the optimal position of the yellow curve.

Table 3: *****

Values of Threshold	Values of Rad			
	5	6	7	10
0.95	(0.20,0.39)	(0.18,0.39)	(0.17,0.42)	(0.47,0.12)
0.96	(0.30,0.24)	(0.46,0.09)	(0.17,0.37)	(0.13,0.41)

Values of Threshold	Values of Rad			
	5	6	7	10
0.97	(0.37,0.13)	(0.36,0.11)	(0.35,0.12)	(0.33,0.11)
0.98	(0.26,0.16)	(0.28,0.12)	(0.29,0.09)	(0.29,0.08)
0.99	(0.42,0.04)	(0.41,0.06)	(0.42,0.04)	(0.48,0.01)

4) Detection on Simultaneous time- and pitch-scaling

Simultaneous time- and pitch-scale modification is applied on the small data set. The time-scaling factor is within a set 0.5, 0.7, 0.9, 1.1, 1.4, 1.7 and 2.0. The pitch-scaling factor is within a set 0.5, 0.7, 0.9, 1.1, 1.3 and 1.5. Each speech clip in the small data set is then changed into 42 modified speech clips. All the original clips and modified clips go through the presented method and DFD values are extracted.

ACKNOWLEDGMENT

This work is supported by, National Natural, and Science Foundation of China (61003094, 90820304, 90920302). And also funded by Tsinghua National Laboratory for Information Science and Technology (TNList) Cross-discipline Foundation.

V. CONCLUSIONS & FUTURE WORKS

In this paper, we present a simple method to detect TD-PSOLA modified voices, both in time-scale and pitch-scale. A single feature duplicated fragments density is defined and evaluated. Experimental results show that our method is effective when lengthening the duration or heightening the pitch frequency of the voice with a significant modification factor. Future works include research on detecting modified voices on spectral scale, as well as detecting modified voices by various other prosodic modification algorithms and algorithm-unknown applications.

REFERENCES

[1] Introduction to Speech Processing. Richard Gutierrez-Osuna. CSE@TAMU

[2] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA techniques," *Speech Communication*, vol. 11, pp. 175–187, 1992.

[3] Prosody Modification Using Instants of Significant Excitation. K. Sreenivasa Rao, B. Yegnanarayana. *IEEE Transaction on Audio, Speech, and Language Processing*, Vol 14, No.3, May 2006.

[4] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text to speech synthesis using diphones," *Speech Commun.*, vol. 9, pp. 453–467, Dec. 1990.

[5] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Commun.*, vol. 16, pp. 175–205, Feb. 1995.

[6] M. R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. ASSP-29, no. 3, pp. 374–390, Jun. 1981.

[7] S. Roucos and A. Wilgus, "High quality time-scale modification of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, FL, Mar. 1985, pp. 493–496.

[8] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, Munich, Germany, 1997, pp. 1303–1306.

[9] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, pp. 187–207, 1999.

[10] E. Moulines, W. Verhelst. "Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech". *Speech Coding and Synthesis*, Chapter 15, 1995.

[11] S. Gonzalez and M. Brookes. "A pitch estimation filter robust to high levels of noise (PEFAC)", *Proc EUSIPCO*, Aug 2011.

[12] TIMIT Acoustic-Phonetic Continuous Speech Corpus. <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1>.

[13] C. Grigoras, "Digital Audio Recording Analysis: The Electric Network Frequency (ENF) Criterion", *The International Journal of Speech Language and the Law*, vol. 12, no. 1, pp. 63-76, 2005.

[14] A. J. Cooper, "The Electric Network Frequency (ENF) as an aid to Authenticating Forensic Digital Audio Recordings – An automated Approach" *AES 33rd International Conference Audio Forensics Theory and Practice*, Denver, CO, USA, (2008, Jun).

[15] Yang R, Shi YQ, Huang J. Defeating fake-quality MP3. *MM&Sec 2009*: 117–124; (2009).

[16] A.J Cooper. Detecting butt-spliced edits in forensic digital audio recordings. *AES 39th International Conference*, Hillerød, Denmark, 2010 June 17–19.