Pitch and Duration Modification for Speech Watermarking

Mehmet Celik*, Gaurav Sharma* and A. Murat Tekalp*†
* Electrical and Computer Engineering Department
University of Rochester, Rochester, NY 14627–0126
Email: u.celik@ieee.org, gaurav.sharma@rochester.edu, tekalp@ece.rochester.edu
† The College of Engineering, Koc University, Istanbul, Turkey.

Abstract— **We propose a speech watermarking algorithm based on the modification of the pitch (fundamental frequency) and duration of the quasi-periodic speech segments. Natural variability of these speech features allows watermarking modifications to be imperceptible to the human observer. On the other hand, significance of these features makes the system robust to common signal processing operations and low data-rate source excitation based speech coders. This class of coders is particularly obstructive for conventional audio watermarking algorithms when applied to speech signals. A pitch synchronous overlap and add (PSOLA) algorithm is used for pitch and duration modifications in the watermark embedding phase. Experiments with multiple speech codecs show very good robustness with low data-rate (5-8 kbps) speech coders.**

I. Introduction

In the last decade, digital watermarking has attracted considerable research interest. Numerous applications in various domains (e.g. audio, video) have been proposed and related watermarking techniques have been studied rigorously. Successful algorithms for traitor tracing, authentication, copy prevention, broadcast monitoring and steganography applications have been demonstrated, especially for still images, audio and video signals. Watermarking of speech signals, however, remains a significantly understudied subject.

In general, speech may be considered as a class of one-dimensional signals—and of audio signals, in particular. Thus, watermarking techniques that have been previously proposed for the audio domain may be applied to the speech watermarking problem. Nevertheless, this approach fails to recognize and exploit the domain specific characteristics for improved performance. Similarly, it fails to recognize and counter a set of benign and malicious attacks[1] that are not applicable to general class of audio signals, yet encountered frequently in the speech domain. One such operation is the low data-rate (5-8 kbps) speech compression algorithms that utilize the source excitation model of speech formation [1, Chap.15].

In the literature, only a few watermarking methods have been proposed specifically for the speech domain. In [2], Wu et al. propose a QIM (quantization index modulation) technique that operates on the DFT (discrete Fourier transform) coefficients. The method is tuned for the speech domain by its exponential scaling property, which targets the psychoacoustic masking functions and band-pass characteristics of the CELP [1] coder. In [3], Hatada proposes embedding by modificaiton of the LSP (line spectrum pair) parameters. Likewise, in [4], Wu et al. have utilized the CELP parameters as robust hash values for semi-fragile authentication. These parameters are shown to be robust against common signal processing operations. Note that this method is a signature scheme that transmits the computed hash values through an auxiliary channel, i.e. as a part of file headers. Therefore, it does not require modification of these parameters to embed a watermark. Finally, Hagmuller et al. have applied spread spectrum watermarking with perceptual masking to speech signals for air traffic control applications [5].

In this paper, we propose a robust speech watermarking algorithm. Our method introduces small changes in the pitch (fundamental frequency) and duration of quasi-periodic speech segments in order to embed the watermark payload. Insensitivity of the human perception to small changes in these features and their natural variability ensures the transparency (imperceptibility) of the watermark. On the other hand, stability of such features under low data-rate compression makes the method particularly valuable for semi-fragile authentication applications. A quantization index modulation (QIM) [6] scheme is used to encode payload bits into these features. In the next section, we overview the speech signal characteristics, define these features and explain how they are modified for watermarking purposes. The effectiveness of the algorithm is demonstrated via examples in the subsequent section.

II. Watermarking by Pitch and Duration Modification

A. Speech signal characteristics

Most languages, including English, can be described in terms of a set of distinctive sounds, or *phonemes* [7]. The phonemes may be divided into two broad classes for the purposes of this discussion. The first group comprises of quasi-periodic sounds, such as vowels, diphthongs, semivowels and nasals. These phonemes show periodic signal structures. The second group comprises of the rest of the phonemes, i.e. stops, fricatives, whisper and affricates. These posses no apparent periodicity.

[1] We use the term *attack* to cover both intentional operations aimed at disabling watermark detection and innocuous signal processing operations which a speech signal commonly encounters, such as lossy compression.

0-7803-8874-7/05/$20.00 ©2005 IEEE

The periodicity of the phonemes in the first group is known as the *fundamental frequency* or the *pitch period*. The pitch period of a speech segment is affected by two conditions: *i)* the physical characteristics of the speaker, e.g. gender, and *ii)* the excitement of that speaker. Similarly, the *duration* of these phonemes also vary with the accent, intonation, tempo and excitement of the speaker. For the example shown in Fig. 1(a), the average pitch period and total duration are determined as 7.75 ms (129 Hz) and 140 ms, respectively.

B. Pitch synchronous overlap and add (PSOLA)

PSOLA [8] is a simple and effective method for modifying the pitch and duration of quasi-periodic phonemes. It has been first proposed as a tool for text-to-speech (TTS) systems that form the speech signal by concatenating pre-recorded speech segments. First, a speech signal is parsed for different elementary units(diphones) that start and end with a vowel or silence. During synthesis, various units are concatenated by overlapping the vowels to form words and phrases. In the TTS application, it is often necessary to match the pitch period of two units before concatenation. Moreover, duration of the vowel is modified for better reproduction.

The effect of pitch and duration modification on a speech waveform is seen in Fig. 1. Original signal waveform (see Fig. 1(top)) is 140 ms long and average pitch period is 7.75 ms (129 Hz). The same segment is seen in Fig. 1(middle) after its duration is increased by 10% up to 154 ms. Similarly, it is seen in Fig. 1(bottom) after the pitch period is reduced to 7.2 ms (139 Hz). Neither of these modifications results in any audible artifact or perceptual change.

C. Proposed watermarking system

Our watermarking system comprises of the following steps:

1) Divide speech signal into suitable segments.

2) Modify the average pitch period within each segment by:

2.a) Determining the pitch periods,

2.b) Computing the average pitch period,

$$p_{avg} = \sum_{i=1}^{N} p_i / N. \tag{1}$$

2.c) Computing the new value that reflects the payload bit, b, using dithered QIM [6].

$$p_{avg}^{wm} = Q_b(p_{avg} + n) - n, \tag{2}$$

where n is the pseudo random dither value and Q_b is the chosen quantizer.

2.d) Modifying each pitch period such that

$$p_i^{wm} = p_i + (p_{avg}^{wm} - p_{avg}). \tag{3}$$

3) Concatenate the segments to form the watermarked signal.

At the receiver, the analysis steps outlined above are repeated to determine the modified average pitch value for each segment from which the watermark payload is extracted.

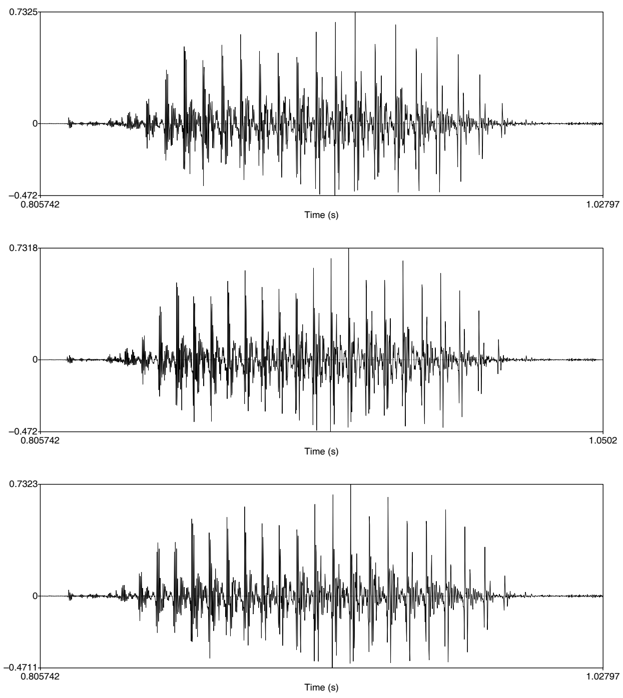

Fig. 1. *a)* Speech segment corresponding to the word *gift* pronounced by a male speaker. The average pitch is 129 Hz and the duration of the quasi-periodic part is 140 ms. *b)* After duration modification with PSOLA. Duration has been increased 10% to 154 ms. *c)* After pitch modification with PSOLA. Pitch values are shifted up 10 Hz to an average of 139 Hz.

In the first step, the algorithm inspects the power of the speech signal in a sliding window and detects the pauses or unvoiced segments. Using these points as separators, speech is divided into continuous words or phrases. At this step, we do not require that the chosen segments correspond to actual words, we only require that the algorithm be repeatable with sufficient accuracy. Fig. 2 shows a speech waveform and its intensity profile. Word/phrase boundaries are seen clearly.

Once speech segments are isolated, pitch periods are determined. The pitch periods are then modified such that the average pitch period of each word/phase reflects a payload bit. As indicated in Eqn. 2, the payload information is embedded by a dithered QIM scheme [6], which is preferred due to its robustness against additive noise and favorable host signal interference cancellation properties. We have experimentally determined that the average pitch period is a robust feature. Therefore, it is not necessary—yet still possible—to impose additional redundancy using projection based methods or spread spectrum techniques.

The proposed method utilizes specific speech signal features associated with speech generation models for the embedding of watermark payload. These are incorporated and preserved in source-model based speech coders that are commonly employed for low data-rate (5-8 kbps) communication of speech. The method is therefore naturally robust against these coders and significantly advantaged in this regard over embedding methods designed for generic audio watermarking. The embedding capacity of the method, though relatively low, is sufficient for meta-data tagging and semi-fragile authentication applications, in which robustness against low data-rate compression is of particularly importance. We also note that the method has not been designed to be robust against a malicious attacker who aims to obstruct detection of the watermark. A systematic modification of the speech features, for instance by re-embedding, would typically disable the watermark.

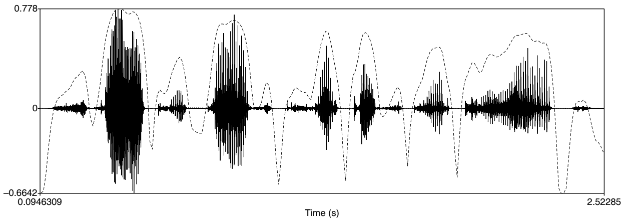

Fig. 2. The speech signal for the sentence *Spherical gifts are difficult to wrap* and its intensity profi le (shown as dotted lines). The sentence is divided into words/phrases at low intensity points, e.g. pauses.

III. Experimental Results

The proposed method has been implemented using: *i)* intensity analysis for segmenting the signal into approximate words/phases; *ii)* auto-correlation method [7] for detecting pitch periods; *iii)* dithered QIM method (see Eqn. 2) for encoding the payload; *iv)* TD-PSOLA method for altering the average pitch period of a segment (word/phrase)[2]. The quantization step size for the QIM method is chosen to be 10 Hz, which satisfies the robustness and transparency requirements simultaneously. Moreover, each segment determined in step *(i)* is checked for periodicity and average power. If a unit does not meet the periodicity and minimum energy requirements, it is excluded from the modification process.

We have tested our system using a database of speech samples provided by NSA for testing speech compression algorithms. The results are illustrated here using a specific 22.5 second sample, in which eight sentences are read by multiple male speakers at various speeds. The speech segment that corresponds to one of the sentences is shown in Fig. 2.

The intensity profile, also seen in Fig. 2, indicates how the sentence is divided into words or continuous phrases. The first isolated segment (see step *(i)* above) is not used for watermarking, because it does not meet the periodicity/energy constraints. Detected pitch periods for this signal before (solid line) and after (dotted line) watermark embedding are shown in Fig. 4. The pitch period is shifted uniformly over a word/phrase when the average pitch value does not lie in the center of the quantization bin that corresponds to the payload bit.

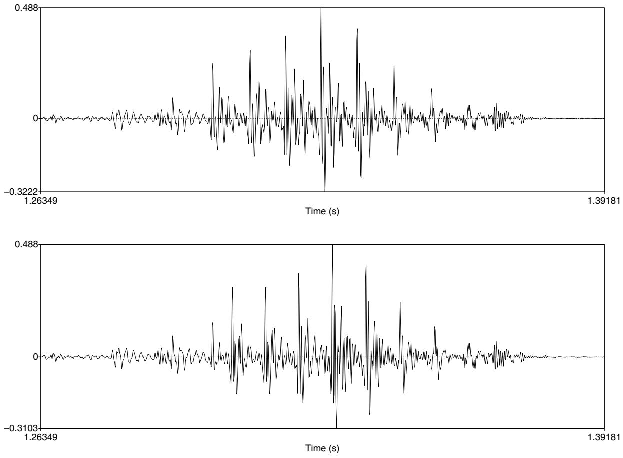

Fig. 3. *a)* Speech segment corresponding to the word *are* for the sentence in Fig. 2. The average pitch is 120 Hz. *b)* Same segment after watermark embedding. Average pitch value has been shifted to 130 Hz.

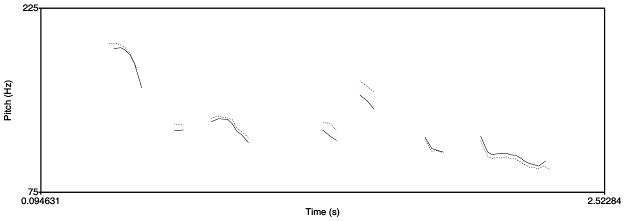

Fig. 4. Detected pitch values for the original (solid) and watermarked (dotted) signals.

We have tested the robustness of our algorithm with low data-rate speech codecs, which are represented by the following examples[3]:

- **GSM-06.10** (Global System for Mobile Communications coder, version 06.10) at 13 kbps. This codec is commonly used in today's second generation (2G) cellular networks that comply with GSM standard.
- **AMR** (Adaptive Multi-Rate coder) at 5.1 kbps. This codec has been standardized for third generation cellular networks (3GPP standard).
- **QCELP** (Qualcomm Code Excited Linear Predictive coder) at 6.8 kbps.This codec has been standardized for third generation cellular networks (3GPP2 standard).

In general, these codecs operate at an SNR of 2-5 dB. As they utilize powerful models for speech formation, the low SNR values do not correspond to a loss in speech quality. Nevertheless, they are very effective at removing audio watermarks that do not adhere to speech production models, especially spread spectrum techniques. In Fig. 5, we see that the performance of a spread spectrum watermark is severely impaired by AMR compression at 5.1 kbps. In this example, an all-pass spread

[2] The implementation uses the toolboxes and the scripting capabilities of the *Praat* [9] phonetics toolkit.

[3] We have used Apple Quicktime for AMR and QCELP codecs and TOAST utility [10] for GSM-06.10 codec.

spectrum watermark is embedded into the speech sample described above. The signal to noise ratio is -27.5 dB and the spreading factor is 2700 samples, which corresponds to an embedding capacity of 3 bits/sec. Compression results in a bit error rate (BER) of $\approx 30\%$.

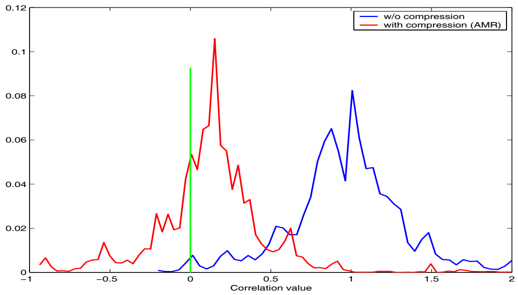

Fig. 5. Effect of compression (AMR 5.1 kbps) on detection performance of spread spectrum watermarking ($SNR \simeq -27.5dB$).

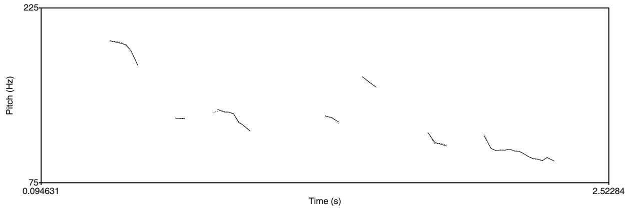

Fig. 6. Detected pitch values for the watermarked (dotted) and compressed (solid) signals. (AMR codec at 5.1 kbps is used.)

The effect of low data-rate speech compression—AMR (Adaptive Multi-Rate) coder at 5.1 kbps—on the pitch periods is shown in Fig. 6. The pitch period curves for the watermarked (dotted) and watermarked-then-compressed (solid) signals are in complete agreement. In this case, all the embedded bits can be recovered without any errors.

The robustness of the scheme against all three coders for the complete test sample mentioned above is summarized in Table. I. The sample is 22.5 sec long. The segmentation of the sample results in 66 units (words/phrases) that are suitable for embedding. The corresponding embedding capacity is 3 bits/sec. The watermark embedding step, however, does effect the results of the parsing process. Sometimes a unit may be divided into two units or two units may be merged to form a single unit, as a result of watermarking modifications. These errors are seen as deletions or insertions in the detected payload. For the given example, one of the 66 units is missed and 4 new units are detected, corresponding to one deletion and four bit insertions. Similar result are obtained for the watermarked-then-compressed samples.

Since the rest of the payload can be transmitted without any errors, these small number of insertions and deletions may be eliminated by introducing error correction coding methods capable of handling insertions and deletions [11]—if required by the application. Overall, the proposed method has a capacity of approximately 3 bits/sec and it is very resilient against low data-rate speech codecs.

TABLE I
ROBUSTNESS AGAINST LOSSY COMPRESSION

	Embed	Detect	GSM-6.10	QCELP	AMR
Correct bits	66	65	65	65	66
Bit errors	0	0	0	0	0
Insertions	0	4	2	2	3
Deletions	0	1	1	1	0

IV. CONCLUSIONS

We propose a digital watermarking method specifically designed for speech signals. The method is based on slightly altering the pitch and duration of the periodic speech segments. As these fundamental speech features remain intact during subsequent signal processing and compression operations, the system is particularly robust against such non-malicious attacks. The modification of average pitch values is determined by using dithered quantization index modulation (QIM), which eliminates host signal interference allowing improved detection. The actual modification of individual pitch values is carried out using a pitch synchronous overlap and add (PSOLA) method that does not cause any perceptual artifacts.

We demonstrate the proposed scheme's robustness against low data-rate (5-8 kbps) source excitation model based speech codecs—such as QCELP and AMR. These coders operate at a SNR of 2-5 dB and they have been particularly challenging for existing methods. The method allows a relatively low embedding capacity (approx. 3 bits/sec), which is suitable for meta-data tagging and authentication applications.

REFERENCES

[1] K. Sayood, *Introduction to data compression*, 2nd ed. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2000.

[2] C. Wu and C. Jay Kuo, "Fragile speech watermarking based on exponential scale quantization for tamper detection," in *Proc. IEEE Intl. Conf. Acoustics Speech and Sig. Proc.*, May 2002, pp. 3305–3308.

[3] M. Hatada, T. Sakai, *et al.*, "A study on digital watermarking based on process of speech production," *IPSJ SIGNotes Computer SECurity*, vol. 017, no. 007, 2002.

[4] C. Wu and C. Jay Kuo, "Comparison of two speech content authentication approaches," in *Proc. SPIE: Security and Watermarking of Multimedia Contents IV*, E. J. Delp and P. W. Wong, Eds., vol. 4675, Jan. 2002, pp. 158–169.

[5] M. Hagmller, H. Hering, *et al.*, "Speech watermarking for air traffic control," in *12th European Signal Processing Conference*, 2004.

[6] B. Chen and G. W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Info. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.

[7] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, New Jersey: Prentice Hall, 1978.

[8] E. Molines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, pp. 453–467, 1990.

[9] P. Boersma and D. Weenik, "Praat: doing phonetics by computer." [Online]. Available: http://www.fon.hum.uva.nl/praat/

[10] "GSM 06.10 coder." [Online]. Available: http://kbs.cs.tu-berlin.de/ jutta/toast.html

[11] M. Davey and D. Mackay, "Reliable communication over channels with insertions, deletions, and substitutions," *IEEE Trans. Info. Theory*, pp. 687–698, Feb. 2001.