High-Rate Data Embedding in Unvoiced Speech

Konrad Hofbauer [1,2]

[1] Eurocontrol Experimental Centre
INO – Innovative Research

91222 Brétigny-sur-Orge, France
konrad.hofbauer@tugraz.at

Gernot Kubin [2]

[2] Graz University of Technology
Institute of Signal Processing
and Speech Communication
8010 Graz, Austria
g.kubin@ieee.org

Abstract

We propose a blind speech watermarking algorithm which allows high-rate embedding of digital side information into speech signals. We exploit the fact that the well-known LPC vocoder works very well for unvoiced speech. Using an auto-correlation based pitch tracking algorithm, a voiced/unvoiced segmentation is carried out. In the unvoiced segments, the linear prediction residual is replaced by a data sequence. This substitution does not cause perceptual degradation as long as the residual's power is matched. The signal is resynthesised using the unmodified LPC filter coefficients. The watermark is decoded by a linear prediction analysis of the received signal and the information is extracted from the sign of the residual. The watermark is nearly imperceptible and provides a channel capacity of up to 2000 bit/s in an 8 kHz-sampled speech signal.

Index Terms: data hiding, speech watermarking, LPC vocoder, LPC residual, voiced-unvoiced segmentation

1. Introduction

Digital watermarking is the embedding of digital data into a video, image or audio signal. The embedding must be done such that the perceptual quality of the host signal does not seriously degrade. In the last decade, watermarking has received considerable attention from various application fields including traitor tracing, authentication, copy prevention, broadcast monitoring, steganography, archiving and legacy system enhancement.

System designs vary depending on the type of the host signal, be it video, image or audio. Our focus lies on *speech watermarking* for the transmission of additional side information over the analogue legacy voice radio communication link between aircraft and air traffic controller [1, 2]. This implies *blind* watermarking where the decoder does not know the original host signal, and an uncoded transmission over an analogue noisy channel.

1.1. Related Work

In its early days, watermarking was seen as a purely *additive* process (Fig. 1a). It was soon recognised, that the audibility can be reduced when the watermark signal is spectrally shaped. The principle is shown in Fig. 1b: the data undergoes adaptive or non-adaptive filtering before it is added to the speech. The prominent class of spread-spectrum watermarking systems which are based on the detection of pseudo-random noise sequences belong to this category. Theoretical considerations however showed that there is an intrinsic limitation on the achievable trade-off between data rate, speech quality and robustness due to the interference between host signal and watermark. The watermark is for perceptual reasons usually around 20 dB below the speech signal, which makes detection a difficult task. Examples of this type of system are [1] and [3], which achieve bit rates in the region of 30 and 200 bit/s. Additive embedding can occur as well in a *transform domain* of the signal (Fig. 1c). This is commonly done in other watermarking domains but is not very popular for watermarking speech.

A new class of watermarking algorithms evolved from the modelling of watermarking as a communication channel with side information [4]. The basic idea is to modify the host signal or its transform domain representation, instead of adding a second signal. The *Quantisation Index Modulation* (QIM) technique would later gain popularity. In principle, QIM relies on requantising the signal with different codebooks depending on the watermark message (Fig. 1d, [5]). The embedding of data in the least significant bits (LSB) follows this principle, but is not very robust to channel noise. Algorithms based on the QIM of line spectrum pair parameters [6], pitch period [7] and LPC residual [8], and with bit rates in the range from 3 to 300 bit/s have been proposed.

A generalisation of QIM is the concept of *modulating* the signal or one of its parameters by the watermark (Fig. 1e). The system presented in [9] estimates and inverts the polarity of speech segments and embeds one watermark-bit per syllable. By contrast, [10] modulates the frequency of selected partials in a sinusoidal speech model and achieves bit rates in the range of 400 bit/s.

It can be concluded that the reported bit rates vary widely, starting at a few bits per second but do not exceed 400 bit/s. This is mostly due to the fact that great care is taken to maintain the original shape or properties of the speech signal in order to minimise perceptual distortion.

A different approach is presented in [11]. In segments where a certain frequency component of an audio signal above 5kHz is of noise-like structure, this component is replaced by a pseudo-random signal, which is modulated by the watermark (Fig. 1g). This method is not directly applicable to low bandwidth radio speech.

We propose a similar approach which is tailored to speech signals, given the way speech is perceived and the lessons learnt from linear predictive coding.

1.2. LPC Vocoder

Linear predictive coding (LPC) of speech signals was first presented by Atal and Schroeder [12] almost 40 years ago and has found its way into many prominent low-rate speech coders [13]. Fig. 2

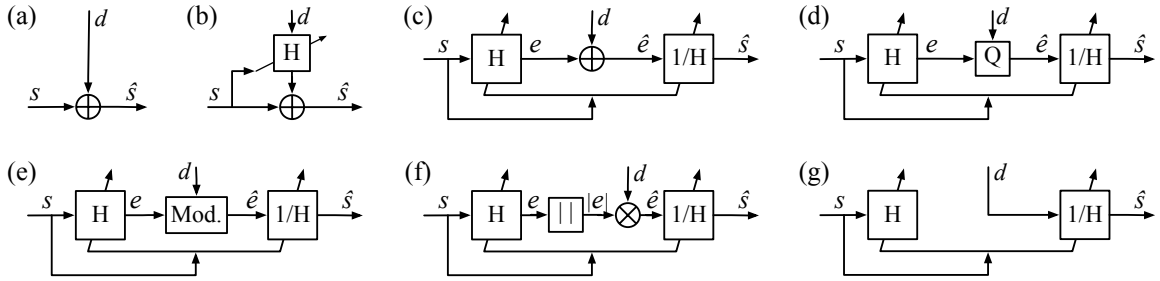

Figure 1: A watermarked signal $\hat{s}$ is generated from the original speech signal s and the data signal d by **(a)** adding the data to s, **(b)** adding adaptively filtered data to s, **(c)** adding the data onto a transform domain signal representation e, **(d)** quantising e according to the data, **(e)** modulating e with the data, and, as proposed in this paper, **(f)** multiplying $|e|$ with the data and **(g)** replacing e by the data.

illustrates the underlying source filter model.

From a recorded speech signal, the coefficients of the filter are obtained by linear prediction from the previous samples. Inverse filtering of the speech signal with this adaptive filter results in the so-called prediction *residual* e.

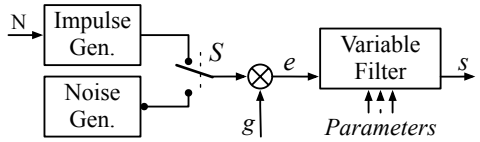

Figure 2: Source filter model of speech production [14]. For voiced segments the excitation e created by the vocal chords is a periodic pulse train with a pitch period N, and a random noise signal for unvoiced segments. Both are amplified by a time-variant gain g. A parametric minimum-phase all-pole filter models the subsequent vocal tract.

In the context of speech coding, it has been shown that the source filter model provides a very accurate representation for the *unvoiced* components of speech [15]. A listening test with 32 subjects showed that resynthesised speech with pure white noise excitation in the unvoiced segments leads to mean opinion scores (MOS) that are almost identical to those of unmodified PCM speech. The conclusion is that the speech quality does not degrade when the LPC residual in unvoiced segments is replaced by white noise of equal power. The proposed watermarking scheme is based on this very idea.

1.3. Proposed Watermarking Scheme

The proposed watermarking system belongs to the family of modulation models (Fig. 1e). The system applies two rather drastic types of modulations. The first is the simple *multiplication* of the speech signal in a transform domain with the data signal (Fig. 1f). The second is the complete *removal* of the original transform domain signal and a replacement by the watermark signal as depicted in Fig. 1g. For the reasons mentioned above, we select the LPC residual of *unvoiced* speech segments as transform domain signal representation. The basic structure of the resulting watermarking system is shown in Fig. 3a. The LPC residual is split up into voiced and unvoiced segments. While the voiced segments pass the system unmodified, the watermarking data is embedded into the unvoiced segments.

2. Voiced/Unvoiced Segmentation

As a basis for the voiced/unvoiced (V/U) segmentation, a pitch estimation is computed with the standard auto-correlation based pitch tracking algorithm as implemented in PRAAT [16]. We use a fundamental frequency range of 60 Hz to 500 Hz and a voicing threshold of 0.35, which is below PRAAT's default value of 0.45 and leads to more voiced and less unvoiced segments. This bias towards voiced components ensures that these perceptually important segments remain unmodified in the later processing steps. Again using PRAAT, the pitch marks in the voiced regions are determined with a cross-correlation value maximisation over adjacent pitch cycles. The pitch marks define the voiced and unvoiced regions, with all non-voiced segments considered as unvoiced and with guard intervals of 0.2 and 0.7 times the mean pitch period before and after the voiced segments (Fig. 4).

3. Analysis/Synthesis Framework

Speech analysis and resynthesis are based on the LPC vocoder approach as discussed in Sec. 1.2. The same linear prediction filter coefficients are used for the analysis and for the inverse synthesis filter. The coefficients are estimated from the 8 kHz speech signal and periodically updated using two strategies.

3.1. Frame-based LPC

A 10th-order LP analysis is updated every 30 samples using the Levinson-Durbin algorithm and a window length of 160 samples. Each frame is divided into two subframes of 15 samples, for which the the LPC coefficients are interpolated using line spectral frequency (LSF) parameters. The residual is computed for each subframe. After potential modification by the watermark embedding, the speech signal is resynthesised using the interpolated LPC parameter sets.

3.2. Sequential Lattice Filter-based LPC

Section 6 will show that frame-based processing requires sample-accurate frame synchronisation between embedder and decoder. In order to circumvent this requirement, we can alternatively update the LPC coefficients after each sample, using adaptive lattice analysis as presented in [17]. We use a 10th-order prediction filter and a forgetting factor of $\beta = 0.97$. The applied one-pole low-pass windowing filter has an effective length of $L = \frac{2N}{(1-\beta)} = 670$ samples. Notably, the sample-accurate interpolation of block-based LPC/LSF parameters would lead to equivalent results.

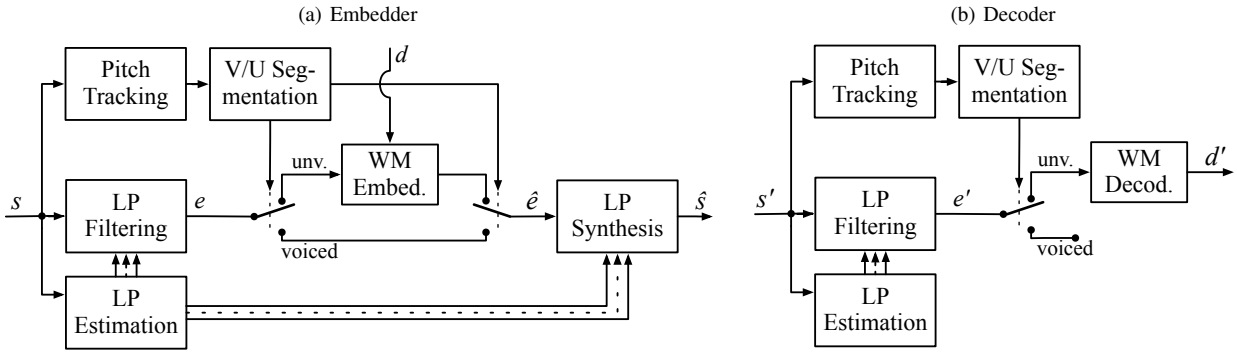

Figure 3: Block diagram of the proposed speech watermarking scheme. Two different data embedding strategies are shown in Fig. 1(f, g).

4. Watermark Embedding

The watermark data is assumed to be a binary coded random signal with signal values $A = \{-1, 1\}$. It is embedded in the LPC residual of the unvoiced segments.

Using the *multiplication* method depicted in Fig. 1f, the absolute value of each sample of the residual is multiplied with a sample of the watermark signal. The power of the residual is thereby accurately maintained, and the information is encoded in the sign of the residual. The symbol rate is 1 bit/sample. With a conservative assumption of 25% of the total speech being unvoiced speech (in contrast to 36% labelled unvoiced in TIMIT [15]), this leads with a sampling rate of $f_s = 8000$ Hz to an embedding rate of 2000 bit/s.

The same rate is achieved with the *replacement* method in which the residual is replaced by the binary watermark signal (Fig. 1g). For each sub-frame, the power of the watermark-signal is matched with the power of the original residual.

The more robust *low-rate* method is based on the same replacement concept. A 15 bit long pseudo-random noise (PN) sequence is used as residual in each subframe . The subframes are BPSK (binary phase shift keying) modulated by the watermark signal. As a consequence, each unvoiced subframe consists of either the PN sequence or the inverse of the same PN sequence. Again, the power of each subframe is matched with the power of the original residual. The resulting embedding rate is 1 bit per 15 samples or 130 bit/s.

5. Watermark Decoding

The watermark decoder obeys the same structure and processing steps as the embedder (Fig. 3b). For both the multiplication and the replacement method, the *sign* of the blindly re-estimated residual of the received speech signal determines the detected watermark bit. In the low-rate method, the sign of the maximum of the cross-correlation between a subframe of the residual and the PN sequence determines the watermark bit.

6. Simulation Results and Discussion

6.1. Simplifying Assumptions

In order to test for limitations inherent to the watermarking scheme, we make a number of simplifying assumptions, whose impact will be the subject of a later, more formal evaluation. First and foremost, the influence of the transmission channel and channel attacks are not addressed at this point in time. Furthermore frame synchronisation between embedder and decoder is assumed, which could be achieved using conventional synchronisation schemes such as pilot sequence embedding. The sequential adaption of LPC coefficients proposed in Sec. 3.2 could circumvent the problem of frame synchronisation, too.

6.2. Voiced/Unvoiced Segmentation

The segmentation is possibly performed differently at the embedder and the decoder side (Fig. 4). In a simulation using the low-rate method and the speech signal described below, 826 out of 3030 subframes or 27% were marked as unvoiced in the embedder. In the watermark decoder, 701 out of 826 subframes or 85% were correctly re-identified, 92 previously voiced subframes were marked as unvoiced, and 125 subframes were no longer marked as unvoiced.

However, in a more advanced implementation, the embedder's segmentation boundaries could be included in the actual watermark data in order to assist the decoder's V/U estimation. For example, each successfully decoded watermark frame could include a pointer to the next subframe containing a watermark. In the case of no errors this would achieve perfect segmentation synchronisation.

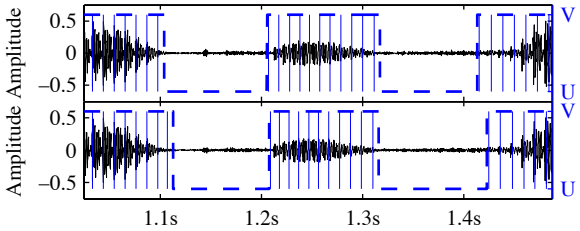

Figure 4: Segment of the original (top) and the watermarked speech signal (bottom). The vertical bars indicate the pitch mark locations, the dashed line the resulting voiced/unvoiced (V/U) segmentation.

6.3. Bit Error Rate (BER)

The following section presents the results of a simulation which uses a sequence of 46000 samples of noisy air traffic control ra-

dio speech recorded with a sampling rate of 8 kHz. As initial experiments showed that all presented methods performed in a comparable way, we focus on the frame-based LPC analysis/synthesis framework for the evaluation.

6.3.1. Low-Rate Method

Using the low-rate method with a bit rate of 130 bit/s, all 815 encoded bits were correctly detected and no bit errors occurred.

6.3.2. Multiplication Method

A bit error rate of about 10% was observed for the multiplication method. These bit errors result from the fact that the adaptive LPC analysis in the embedder and the one in the decoder is performed on slightly different signals and therefore result in slightly different prediction coefficients. For residual values with low amplitude this can lead to a sign change and therefore to a bit error.

6.3.3. Replacement Method

Using the replacement method, a bit error rate of only 0.8% was observed. This is due to the fact that the previously Gaussian residual was replaced by a binary signal (and only the power adapted). The remaining bit errors can be easily overcome with conventional error control coding. Considering the watermark channel as a binary symmetric channel, the channel capacity C in bits per channel use is given by $C = (1-p)\log_2(2-2p) + p\log_2(p)$, where p is the error rate. With 2000 channel uses per second (assuming 25% unvoiced), this results in a watermark data rate of 1865 bit/s, which can be achieved with appropriate coding.

6.4. Perceptual Quality

The quality of the reconstructed speech has not yet been formally evaluated. However, since our signal modifications are very similar to those previously undertaken and formally evaluated [15], it follows that the perceptual degradation is negligible.

Careful subjective listening over headphones showed that a slight difference between the original and the processed speech sound is audible to the expert listener. Especially for noisy speech this difference is not disturbing and does not seem to degrade the perceived speech quality. For demonstration, the unmodified speech signal used in this simulation, as well as the watermarked signals for both the multiplication method and the replacement method, are available on-line [18].

7. Conclusion

We proposed a speech watermarking algorithm which provides—at least under the assumptions made—a bit rate which is 5 to 1000 times larger than in current state-of-the-art speech watermarking systems. The watermark is nearly imperceptible and provides a channel capacity of up to 2000 bit/s. The algorithm is based on the concept of using the speech signal as a carrier of the watermark message. We exploited the fact that the waveform details of the linear prediction residual in unvoiced speech segments is perceptually irrelevant. This allows the embedding of a watermark in the region of the *most* significant bits of the signal. We therefore expect a high robustness to channel noise, which besides a more complete treatment of the synchronisation issue will be the subject of future work.

8. References

[1] M. Hagmüller, H. Hering, A. Kröpfl, and G. Kubin, "Speech watermarking for air traffic control," in *Proc. of the 12th European Signal Processing Conf. (EUSIPCO)*, Vienna, Austria, September 2004.

[2] K. Hofbauer and H. Hering, "Digital signatures for the analogue radio," in *Proc. of the 5th NASA Integrated Communications Navigation and Surveillance Conf. (ICNS)*, Fairfax, USA, May 2005.

[3] Q. Cheng and J. Sorenson, "Spread spectrum signaling for speech watermarking," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[4] I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*. Morgan Kaufmann Publishers, 2001.

[5] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Transactions on Information Theory*, vol. 47, no. 4, 2001.

[6] M. Hatada, T. Sakai, N. Komatsu, and Y. Yamazaki, "Digital watermarking based on process of speech production," in *Proc. of SPIE - Multimedia Syst. and Appl. V*, 2002.

[7] M. Celik, G. Sharma, and A. M. Tekalp, "Pitch and duration modification for speech watermarking," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[8] B. Geiser, P. Jax, and P. Vary, "Artificial bandwidth extension of speech supported by watermark-transmitted side information," in *Proc. of the 9th European Conf. on Speech Communication and Technology EUROSPEECH*, 2005.

[9] S. Sakaguchi, T. Arai, and Y. Murahara, "The effect of polarity inversion of speech on human perception and data hiding as an application," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

[10] L. Girin and S. Marchand, "Watermarking of speech signals using the sinusoidal model and frequency modulation of the partials," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.

[11] R. C. F. Tucker and P. S. J. Brittan, "Method for watermarking data," Feb. 6 2003, U.S. Patent US 2003/0028381 A1.

[12] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals," in *Proc. of the Int. Conf. on Speech Communication and Processing*, 1967, pp. 360–361.

[13] A. S. Spanias, "Speech coding: a tutorial review," *Proc. of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 1994.

[14] P. Vary and R. Martin, *Digital Speech Transmission*. John Wiley and Sons Ltd., 2006.

[15] G. Kubin, B. S. Atal, and W. B. Kleijn, "Performance of noise excitation for unvoiced speech," in *Proc. of the IEEE Workshop on Speech Coding for Telecommunications*, 1993.

[16] P. Boersma and D. Weenink. (2006) PRAAT: doing phonetics by computer [computer program]. Available: http://www.praat.org/

[17] J. I. Makhoul and L. K. Cosell, "Adaptive lattice analysis of speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 654–659, June 1981.

[18] K. Hofbauer and G. Kubin, "Demonstration files," http://www.spsc.tugraz.at/people/hofbauer/interspeech06/.