Noise Robust Speech Watermarking with Bit Synchronisation for the Aeronautical Radio

Konrad Hofbauer[1,2] and Horst Hering[2]

[1] Graz University of Technology, Austria
Signal Processing and Speech Communication Laboratory
`konrad.hofbauer@TUGraz.at`
[2] Eurocontrol Experimental Centre, France
`horst.hering@eurocontrol.int`

Abstract Analogue amplitude modulation radios are used for air/ground voice communication between aircraft pilots and controllers. The identification of the aircraft, so far always transmitted verbally, could be embedded as a watermark in the speech signal and thereby prevent safety-critical misunderstandings. The first part of this paper presents an overview on this watermarking application. The second part proposes a speech watermarking algorithm that embeds data in the linear prediction residual of unvoiced narrowband speech at a rate of up to 2 kbit/s. A bit synchroniser is developed which enables the transmission over analogue channels and which reaches the optimal limit within one to two percentage points in terms of raw bit error rate. Simulations show the robustness of the method for the AWGN channel.

1 Introduction

Tactical air traffic control (ATC) guidance over continental areas currently relies on voice communication between pilots and controllers. Analogue amplitude modulation (AM) radios are and have been used worldwide for this purpose for more than fifty years, and the standards have not been modified in any significant way since. The aeronautical transceivers operate in the very high frequency (VHF) band (118-137 MHz) with double-sideband amplitude modulation (DSB-AM). They are known for their poor signal quality and sensitivity to disturbances along the propagation path such as noise, fading and Doppler effect. This technology is expected to remain in use for the provision of air traffic management (ATM) for many years to come, even if the use of data communications will progressively increase in the medium and long terms [1].

To reduce the complexity of air traffic to a level that can be handled by air traffic controllers, the traffic is organised in volumes of airspace, called sectors. In a particular ATC sector, all pilots and the controller use a single radio channel, which is called the party line. In order for the controller to have a clear picture of the traffic and to be able to give appropriate instructions, the controller must know exactly which pilot he is in contact with at a given point in time. Pilots therefore give their call sign at the start of every voice message.

The identification step is inherently threatened by, among others, poor quality of the audio signal and human error (where the call sign is misunderstood or the wrong call sign is given by accident). The dangers of failed or mistaken identification are obvious: incidents caused by instructions being given to the wrong aircraft. The risk of miscommunication rises where aircraft with similar call signs are present within the same ATC sector. Reducing the risk and thereby increasing the level of safety in ATC had motivated research in this area.

The Aircraft Identification Tag (AIT) concept has been developed in order to reduce call sign ambiguity, to secure identification and to thereby enhance general safety and security in commercial aviation [2]. AIT relies on digital speech watermarking technology to embed identifiers, such as the call sign, into the voice signal before the signal is transmitted to the ground. The embedded tag is a sort-of digital signature of the aircraft. The tag is hidden in the air-ground voice message as watermark. It is meant to be extracted on the ground and for example transformed into a visual signal on the radar screen (Fig. 1). The goal of AIT is to visually animate the aircraft the pilot is communicating with at the time, and in so doing, increase the chances of successful identification.

The following section describes hypothetical operational concepts of AIT. Section 3 proposes a novel speech watermarking method with a special focus on synchronisation. Simulation results which demonstrate the synchronisation performance and the noise robustness of the method are shown in Section 4.

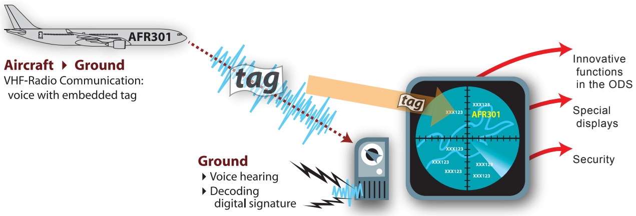


Figure 1. Identification of transmitting aircraft through embedded watermark.

2 Operational Concepts

The operational requirements within the AIT concept specify the necessary performance figures of the watermarking system. Real-time availability of the identification in less than one second implies a certain data rate for the digital transmission. A robust transmission of the message and a verification of the validity of the received data is indispensable. From the user's point of view, the system should not degrade the perceptual quality of the voice transmission. Additionally, it should be autonomous and transparent to the user and not require changes to the well-established procedures in air traffic control and on-board the aircraft.

2.1 AIT Applications

In view of a possible transfer to the industry and a wide-scale implementation, an "AIT Initial Feasibility Study" was performed to provide directions for a full feasibility analysis [3]. This study named three candidate AIT applications.

Identification of Transmitting Aircraft Highlighting the transmitting aircraft on the controllers' human machine interface is the basic-most scenario. For this AIT application the aircraft has to embed its digital signature as a watermark in every transmitted voice message, with the signature being transmitted in less than one second. The embedded data consists of an up to 36 bit long flight or aircraft identifier, such as flight number, aircraft registration number, etc., and does not change within a flight.

In order to be useful for the controller, the embedded tag must be available on the ground within about one second after the start of a voice message. After correlating the AIT signature with available flight plan and radar data, the representing aircraft symbol on the radar screen can be transformed into a visual stimulus until the end of the aircraft voice message. The visual stimulus such as the flashing of the symbol or the animation of the corresponding speed vector as a sinusoid focuses the controller's attention on the speaking aircraft track and thereby supports the understanding of the message.

Uplink of ATC Identification This AIT application consists of the transmission of a non-encrypted ATC domain identifier such as the ATC centre ID, over the voice communications channels from the ground ATC systems to the aircraft. The purpose of this application is to provide a basic authentication of the transmitting ATC station. This allows the pilots to verify that they communicate with the designated sector and have not mistakenly selected a wrong sector frequency. This application is simple and is relatively easy to implement. It is a first step towards more secure voice communication as it identifies 'phantom' controllers who use simple equipment.

Secure Authentication The third AIT application provides pilots and controllers with a real-time confirmation that the received voice message is indeed transmitted by a trusted source. In order to allow proper authentication of the transmitting station or aircraft, the system transmits secured digital signatures in both air-ground and ground-air directions. The digital signatures are exchanged via AIT. The use of secured digital signatures relies on cryptography and therefore on the deployment of cryptographic keys. Key infrastructure and key management for a worldwide aeronautical system is outside the scope of this paper and AIT.

Also within the European Commission FP6 project SAFEE (Security of Aircraft in the Future European Environment [4]) it is aimed to embed a secured authentication in the voice communication using a watermark with an estimated required payload data size of 100 to 150 bits.

2.2 Operational Use and Deployment

The AIT equipment reports the voice message as having positive authentication, negative authentication, or missing authentication.

Positive authentication means that secured digital signatures are embedded in the voice message, that they are correct, and there is no threat. A negative authentication means that secured digital signatures are embedded but that they are incorrect, and that there is a recognised potential threat. This threat could for example result from an attacker playing a recorded voice message including AIT data that does not fit cryptographic keys and time stamps actually in use. A missing authentication may mean that the voice message was too short (less than one second), that the AIT data contains errors that could not be corrected (e.g. due to very poor radio quality), or that there was no AIT data embedded in the voice message (due to equipment failure or lack of equipment).

The system could be put in place gradually over time with subject to deadlines or by select sectors of airspace (ex. the airspace covered by the European Civil Aviation Conference). The different AIT applications could form separate packages and be implemented one after the other. Notably, only a complete and mandatory employment of AIT would deliver the full safety and security benefits.

3 Watermarking Speech

A watermarking algorithm for this particular AIT application faces quite different challenges compared to many other watermarking domains.

The first obvious difference is the host signal domain. A comparably small fraction of watermarking research focuses in particular on speech and its properties. A few QIM-based speech watermarking algorithms have been proposed, which quantise or modulate the line spectrum pair parameters [5], the linear prediction residual [6] or the frequency of partials of a sinusoidal speech representation [7].

The second big difference compared to the classical copyright protection application is that due to the real-time broadcast environment only an ephemeral protection for the moment of the transmission is required, and that an attacker should be unable to produce fake authenticated speech.

Third, again due to the real-time broadcast environment, those types of attacks that would maliciously try to render the watermark unreadable do not apply. Besides attacks such as transformations and collisions, especially the fact that no speech coding is involved removes a big constraint. Embedding can occur in perceptually *irrelevant* speech parameters, which would normally be likely to be removed by a speech coder.

However, the watermark has to be robust against transmission over a radio channel which is time-varying, analogue, narrow-band and noisy. Further on the duration of the watermark must be rather short, as a quick availability of the data is required.

An early AIT prototype which was based on spread-spectrum watermark techniques demonstrated the *feasibility* of the AIT concept [2]. It does not quite fulfil the operational requirements though, which is the motivation for the development of a better performing speech watermarking algorithm, especially when considering the large payload data size that secured authentication requires.

We previously presented a speech watermarking algorithm which exploits the aforementioned differences [8]. However, the system was not tested against the noisy analogue transmission channel. First, perfect synchronisation between the transmitter and receiver was assumed. But, as the watermarking channel is an analogue channel, the digital watermark embedder and decoder are in general *not* synchronised and a synchronisation error is always present. Second, a noiseless transmission channel was assumed. This also does not hold for the desired application since the aeronautical voice radio is in general a noisy transmission channel. The algorithm proposed in this paper is an extension on this work and presents a system which is capable of symbol synchronisation and watermark detection in the presence of channel noise.

3.1 Speech Signal Properties

In general, for a modern watermarking system it is crucial to consider the way the host signal is perceived [9]. Perceptual models are used for this very purpose. It is likewise important to consider the way the host signal is produced, which can provide significant insight into its properties. The following paragraphs outline some speech signal parameters which the proposed watermarking system is based on.

Speech Production and Linear Prediction Linear prediction coding is a powerful and widely used technique for speech processing and coding [10]. Linear prediction (LP) models predict future values of a signal $s(n)$ from a linear combination of the past P signal values $s(n-k)$,

$$\hat{s}(n) = \sum_{k=1}^{P} a_k s(n-k)\,.$$

The prediction coefficients a_k are chosen so that the prediction error (prediction residual)

$$e(n) = s(n) - \hat{s}(n) \tag{1}$$

is minimum in a mean squared error sense. The P-dimensional vector $\mathbf{a}$ of predictor coefficients a_k is given by

$$\mathbf{a} = \mathbf{R}_{\mathbf{ss}}^{-1}\mathbf{r}_{\mathbf{ss}} \tag{2}$$

where $\mathbf{R}_{\mathbf{ss}}$ is the autocorrelation matrix and $\mathbf{r}_{\mathbf{ss}}$ is the autocorrelation vector of the past P input samples. Given the excitation signal $e(n)$ and the predictor coefficients **a** as input, the LP synthesis model output is

$$s(n) = e(n) + \sum_{k=1}^{P} a_k s(n-k)\,. \tag{3}$$

Inversely, given a recorded signal $s(n)$, the LP analysis model computes the residual $e(n)$ using (2) and (1).

The LP model is an all-pole filter model and particularly well suited for speech signals as the poles can model the resonances of the vocal tract. This is applied in the so-called source-filter model of speech production depicted in Fig. 2. In speech signals mostly two types of excitation signals e can be found. In so-called *voiced* speech sounds, such as the vowels, the vocal chords open and close periodically with a certain *pitch* and the excitation signal resembles to a pulse train with a time variant gain g. In *unvoiced* speech sounds, such as the fricatives, the vocal chords are permanently open and create turbulences in the air flow and therefore a white-noise-like excitation signal e, again with a time-variant gain. Speech in regular English language consists of approximately two thirds of voiced and one third of unvoiced segments [11].

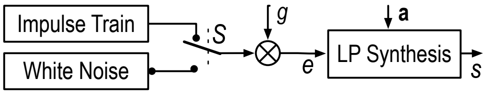


Figure 2. Source filter model of speech production.

Unvoiced Speech Perception and Noise Excitation In *unvoiced* speech it is possible to replace the excitation signal e by a white noise signal of equal power. Previous studies showed that this does not introduce perceptual distortion as long as the time-variant gain g is maintained [11]. This effect is made use of in low-rate speech coding algorithms: In unvoiced segments it is sufficient to code and transmit the LP synthesis model parameters and the running gain—the residual e is substituted on the decoder side by using white noise. This property can also be exploited for embedding a watermark [8]. Our algorithm is also based on this very same principle.

3.2 Watermarking in Unvoiced Speech

Watermark Embedding A block diagram of the watermarking scheme is shown in Fig. 3. The digital discrete-time speech signal with a sampling frequency $f_s = 8\,\text{kHz}$ is split up into voiced and unvoiced segments based on whether a pitch can be found in the signal or not. We use the PRAAT implementation of an autocorrelation-based pitch tracking algorithm [12]. Based on a local cross-correlation value maximisation the individual pitch cycles and the glottal closure instants (the physical counterpart to the aforementioned pulse train) are identified. All regions with pitch marks are considered as voiced regions, whereas all other regions, including pauses, are considered as unvoiced.

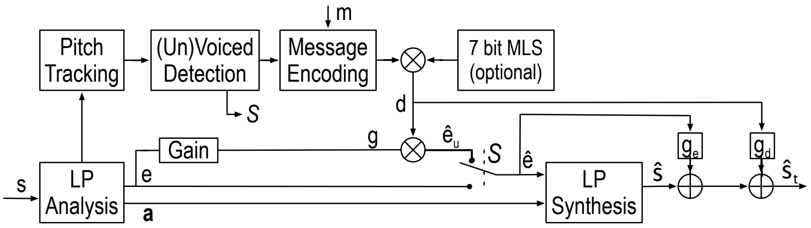


Figure 3. Watermark embedding in unvoiced segments of speech.

In parallel, the linear prediction residual e of the speech signal s is computed. An LP analysis of the order $P = 10$ is performed every 30 samples using (1) and a signal block (window length) of 160 samples. The predictor coefficients $\mathbf{a}$ are interpolated using their line spectral frequency (LSF) representation so that finally an updated set of coefficients is used every $L = 15$ samples. A running gain g is extracted from the residual e by computing the root mean square value within a time window of length L around each sample using a moving average filter on the squared residual $e(n)^2$.

In the unvoiced segments the residual itself is discarded and only its gain g_u is used further on. An entirely new residual is created for the unvoiced segments, which consists of a data signal that carries the watermark information m. The watermark encoding block shown in Fig. 3 outputs a binary-valued signal $d_i \in \{-1, 1\}$: In unvoiced speech segments, d is a sequence of packets of small parts of the watermark message m, in binary encoding. Each packet is prepended by a defined identification sequence which marks the beginning of a packet. In voiced speech segments, d is a uniformly distributed random binary signal. As an optional step in order to increase robustness, the original rate of d of 2000 bit/s is reduced by a factor of seven and each block of seven then identical samples is multiplied with a maximum length sequence (MLS) of equal length. Thus each sample (and bit) of d is spread in time. In both cases, the binary signal d is multiplied by the running gain g in order to form the new residual $\hat{e}_u = gd$ for the unvoiced segments. The unmodified voiced residual e_v and the artificial watermark residual $\hat{e}_u$ are re-joined into a continuous watermarked residual $\hat{e}$ by switching among the two according to the same voiced-unvoiced decision. The speech is resynthesised by LP synthesis according to (3) using the watermarked residual $\hat{e}$ and the predictor coefficients $\mathbf{a}$ that were obtained in the LP analysis.

Two optional measures that impair perceptual quality but, as we will show in Section 4, greatly enhance noise robustness are possible: One is the emphasis of the noise-like and high-frequency components of the signal by adding the residual

$\hat{e}$ itself to the synthesised speech $\hat{s}$ with a constant gain g_e.[3] This could under some circumstances even increase the speech intelligibility. The second option is to create a watermark floor by adding the data signal d with a constant gain g_d to the synthesised speech signal, which is perceived as background noise.[4] The transmitted watermarked speech signal $\hat{s_t}$ can therefore be expressed by $\hat{s}_t = \hat{s} + g_e\hat{e} + g_d d$ with g_e and g_d being zero by default.

Watermark Detection We first assume a digital channel with everything being perfectly synchronised.[5] The basic detection scheme is outlined in Fig. 4. From the received watermark speech signal s' the LP residual e' is computed using LP analysis with the same parameters as in Section 3.2. The estimate d' of the original data signal is given by the sign of the residual, so $d' = \text{sgn}(e')$. If data spreading was used in the embedding, the residual e' is first cross-correlated with the previously applied spreading sequence and then the sign of the correlator output is used as an estimate d'. Using the predefined identification sequence, the embedded data packets, which only occur in unvoiced segments, are located. The data is extracted from these packets, decoded, and results in the message estimate m'.

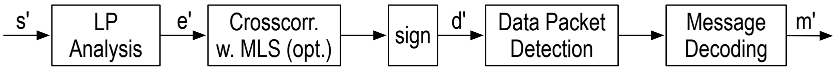


Figure 4. Watermark detection in the residual domain.

3.3 Synchronisation

Synchronisation between the watermark embedder and the watermark detector is a multi-layered problem, certainly so for an analogue radio channel. We address the different aspects of sychronisation from the longest to the shortest time interval.

Synchronisation of the Voiced-Unvoiced Segmentation The previously proposed method contained a voiced-unvoiced detection in the embedder as well as in the detector and requires that the segmentations run perfectly in sync [8]. This is difficult to achieve in practise, as there is a mismatch even with a clean channel due to artifacts introduced by the watermark embedding. One possibility to overcome this issue could be to use an error coding scheme that is capable of handling not only substitution but also insertion and deletion errors [13].

[3] In the simulations the residual is usually not added at all, but if stated so at a level of -20 dB relative to $\hat{s}$.

[4] The watermark floor is applied only where explicitly stated, at a level of -20 dB.

[5] We will deal with synchronisation extensively in Section 3.3.

However, with the packet-based method proposed in Section 3.2 the voiced-unvoiced segmentation in the detector can be completely omitted, since the algorithm simply finds no valid data packets in the voiced regions. The details concerning the packet-based coding of the data as well as its pitfalls and improvements such as packet detection, packet loss recovery, etc. are outside the scope of this paper. Likewise Section 4 focuses on the raw bit error rate (BER) within the unvoiced segments.

Synchronisation of the LP Analysis Frames One might intuitively assume that the block boundaries for the linear prediction analysis in the embedder and detector have to be identical. However, using LP parameters as given above and real speech signals, the predictor coefficients do not change rapidly in between the relatively short update intervals. As a consequence, a synchronisation offset in LP block boundaries is not an issue. Figure 5(a) shows the bit error rate of the proposed watermarking system using a noiseless discrete-time channel. The LP block boundaries in the detector are offset by an integer number of samples compared to the embedder. It can be observed that the bit error rate is not affected.

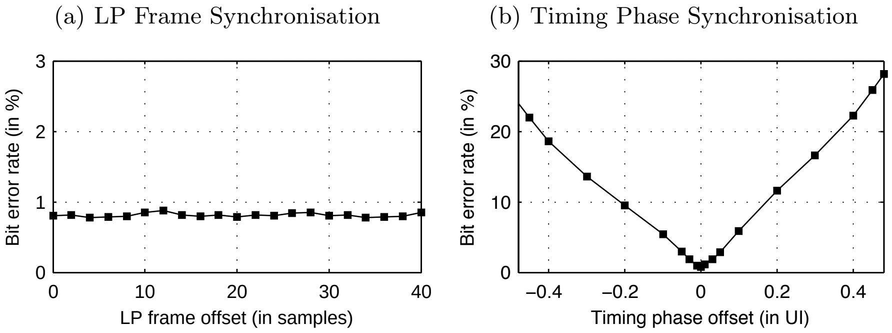


Figure 5. Robustness of baseline system with respect to synchronisation errors.

Data Frame Synchronisation Frame synchronisation is achieved by the periodical embedding of a synchronisation sequence in the digital data stream. The sequence can be detected using, among others, the simple correlation rule, maximum-likelihood estimation or soft-decoding supported methods [14]. This is a well-explored topic in the context of digital communications and is not further treated herein.

Bit Synchronisation and Timing Recovery If the watermarked signal is transmitted over a discrete-time channel, then with the above three measures synchronisation is established. If the channel is however an analogue channel as in the application described in Section 2, the issue of bit or symbol synchronisation arises. The digital clocks in the embedder and detector are in general time-variant, have a slightly different frequency, and have a different timing phase (which is the choice of sampling instant within the symbol interval). It is therefore necessary that the detector clock synchronises itself to the incoming data sequence. Figure 5(b) shows the resulting bit error rate when there is phase shift of a fractional sample in the sampling instants of the watermark embedder and detector. We measure this timing phase error in 'unit intervals' (UI), which is the fraction of a sampling interval $T = \frac{1}{f_s}$.

Although bit synchronisation is a well explored topic, it is still a major challenge in every modern digital communication system due to the strong impact on the detection performance. Prominent methods for non-data aided bit synchronisation are among others the transmission of a clock signal, early-late gate synchronisers, minimum mean-square-error methods, maximum likelihood methods and spectral line methods [15,16].

Spectral Line Bit Synchronisation We present in the following a watermark synchronisation scheme based on the classical spectral line method, due to its simple structure and low complexity.[6] Figure 6 shows the block diagram of the proposed synchroniser. The mathematical derivation of the general spectral line method is given in literature [16] and not repeated herein.

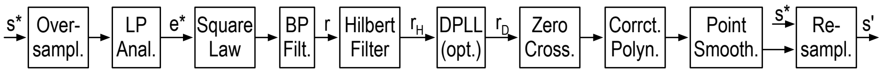


Figure 6. Synchronisation system based on spectral line method.

The received watermarked analogue signal s^* is oversampled by a factor k compared to the original sampling rate f_s.[7] The linear prediction residual e^* of the oversampled speech signal s^* is again computed using LP analysis of the same order $P = 10$ as in the embedder and with intervals of equal length in time. This results in a window length of $k * 160$ samples and an update interval of $k * 15$ samples after interpolation.

We exploit the fact that the (oversampled) residual shows some periodicity with the embedding period $T = \frac{1}{f_s}$ due to the data embedding at these instances. We extract the periodic component r at f_s from the squared residual $(e^*)^2$ with an FIR bandpass filter with a bandwidth of b=480 Hz centred at f_s. The output r of the bandpass filter is a sinusoid with period T, and is phase-shifted by $\frac{\pi}{2}$ with

[6] Whether the proposed structure could be implemented even in analogue circuitry could be subject of further study.

[7] In the later simulations a factor $k = 32$ is used. Values down to $k = 8$ are possible at the cost of accuracy.

an FIR Hilbert filter resulting in the signal r_H. The Hilbert filter can be designed with a large transition region given that r is a bandpass signal.

The zero-crossings of r_H are determined using linear interpolation between the sample values adjacent to the zero-crossings. The positions of the zero-crossings in the positive direction are a first estimate of the positions of the ideal sampling points of the analogue signal s^*. It was found that the LP framework used in the simulations introduces a small but systematic fractional delay which depends on the oversampling factor k and results in a timing offset. We found that this timing offset can be corrected using a third order polynomial $t_\Delta = a_0 + a_1 k^{-1} + a_2 k^{-2} + a_3 k^{-3}$. The coefficients a_i have been experimentally determined to be $a_0 = 0$, $a_1 = 1.5$, $a_2 = -7$ and $a_3 = 16$.

Since the estimated sampling points contain gaps and spurious points, all points whose distance to a neighbour is smaller than 0.75 UI (unit interval) are removed in a first filtering step. In a second step all gaps larger than 1.5 UI are filled with new estimation points which are based on the position of previous points and the observed mean distance between two points. The analogue signal is now re-sampled at these estimated positions. The output is a discrete-time signal with rate f_s, which is synchronised to the watermark embedder and which serves as input to the watermark detector.

Digital Phase-Locked Loop The bit synchronisation can be further improved by the use of a digital phase-locked loop (DPLL), as indicated in Fig. 6. The DPLL still provides a stable output in the case of the synchronisation signal r_H being temporarily corrupt or unavailable. In addition, the use of a DPLL renders the previous point filtering and gap filling steps obsolete.

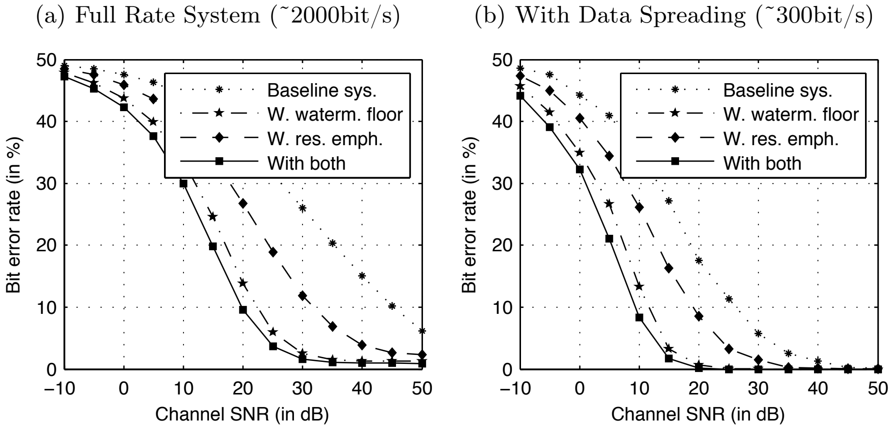
There is a vast literature on the design and performance of both analogue and digital phase-locked loops [17]. Our loop is based on a regular second order all digital phase-locked loop [18]. Inspired by dual-loop gear-shifting DPLLs [19] we use a dual-loop structure to achieve fast locking of the loop. We also dynamically adapt the bandwidth of the second loop in order to increase its robustness against spurious input signals. The exact structure and operation of the implemented loop is presented in an addendum note [20].

4 Simulation Results

The results of simulations of the proposed speech watermarking system using a short sequence of noisy air traffic control radio speech with a sampling rate of $f_s = 8$ kHz are presented hereafter. Out of 45800 samples of the speech signal 12527 samples have been marked unvoiced by the algorithm and used for data embedding, resulting in a raw data rate of 2188 bit/s without spreading or 312 bit/s with spreading. It is important to note that this rate is a variable rate which is dependent on the speech signal. In TIMIT, which is a large English language speech corpus, 36% of the speech is labelled as unvoiced [11].

4.1 Noise Robustness

We first show the robustness of the watermarking system with respect to additive white Gaussian noise (AWGN) on the radio channel assuming perfect synchronisation. Figure 7 shows the raw bit error rate for various signal-to-noise ratios (SNR) of the channel. One bit is embedded per unvoiced sample and the results are given for the cases with and without a watermark floor of $g_d = -20\,\mathrm{dB}$ and for the cases with and without a residual emphasis of $g_e = -20\,\mathrm{dB}$. Both watermark floor and residual emphasis increase the watermark energy in the signal at the expense of perceptual quality.


(a) Full Rate System (˜2000bit/s)

(b) With Data Spreading (˜300bit/s)

Figure 7. Raw bit error rate at different AWGN channel SNR with perfect synchronisation.

4.2 Synchronisation Performance

We already showed in Fig. 5 the adverse effect of a timing phase error on the bit error rate. This timing phase error results from the unsynchronised resampling of the signal. We use a piecewise cubic Hermite interpolating polynomial (PCHIP) to simulate the reconstruction of the continous-time signal and resample the resulting piecewise polynomial structure at equidistant sampling points at intervals of $\frac{1}{f_s}$ or $\frac{1}{kf_s}$ respectively.

In the reconstruction each sample of the embedder output $\hat{s}_t$ serves as a data point in the interpolation. The nodes of these data points would ideally reside on an evenly spaced grid with intervals of $\frac{1}{f_s}$. In order to simulate an unsynchronised system we move these nodes of the data points to different positions according to three parameters:

Timing phase offset: All nodes are shifted by a fraction of the regular grid interval $\frac{1}{f_s}$ (unit interval, UI).

Sampling frequency offset: The distance between all nodes is changed from one unit interval to a slightly different value.

Jitter: The position of each single node is shifted randomly following a Gaussian distribution with variance σ_J^2.

The proposed synchronisation system is capable of estimating the original position of the nodes, which is where the continous-time signal has to be re-sampled for optimal performance. The phase estimation error is the distance between the original and the estimated node position in unit intervals. Its root-mean-square value across the entire signal is shown in Fig. 8 for the above three types of synchronisation errors. The figure also shows the bit error rate for different sampling frequency offsets. The bit error rate as a function of the uncorrected timing phase offset was already shown in Fig. 5(b).

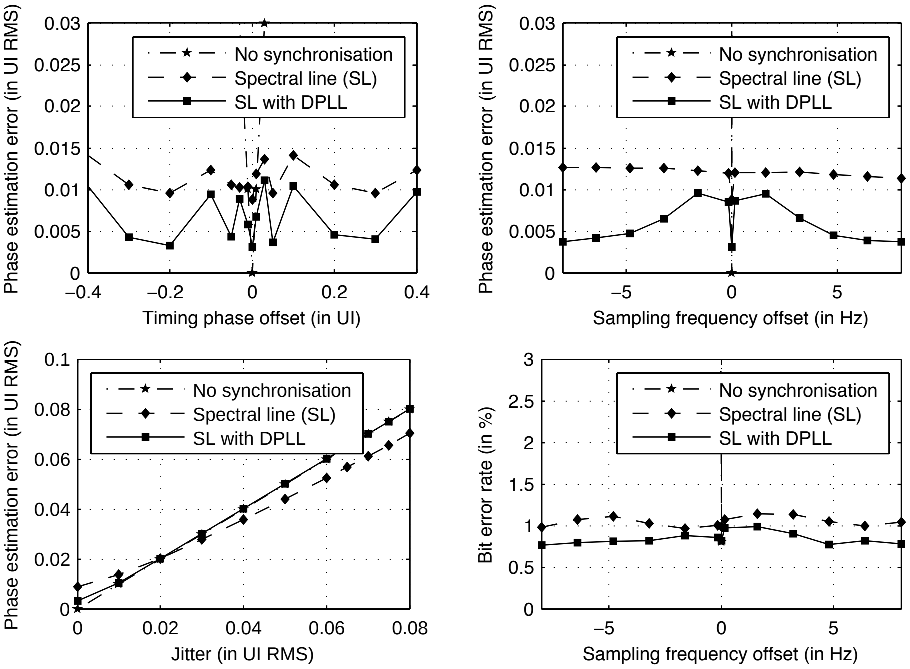


Figure 8. Synchronisation system performance: Phase estimation error and bit error rate for various types of node offsets.

4.3 Overall Performance

Figure 9 shows the raw bit error rate of the overall system including watermark floor, residual emphasis and synchroniser at different channel SNR. Compared to the case where ideal synchronisation is assumed, the raw BER increases by less than two percentage points accross all SNR.

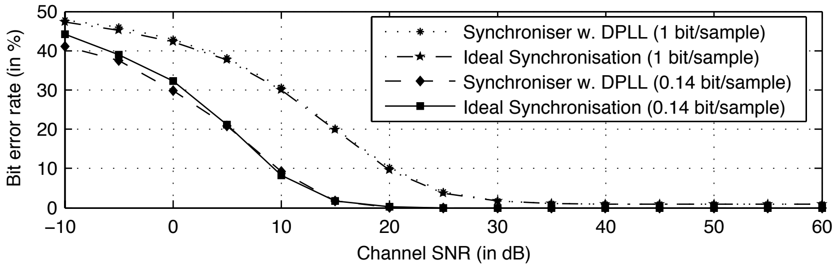


Figure 9. Overall system performance with watermark floor and residual emphasis, with ideal (assumed) synchronisation and with the proposed synchroniser.

The perceptual quality of the watermarked speech has not yet been formally evaluated. The short sample used in the simulations is available online for demonstration [21]. The difference between the original signal and both the baseline system and the system with the residual emphasis is audible but does not seem disturbing, especially for noisy signals. The watermark floor is clearly audible and its level would in practise be adjusted to the expected channel noise level. The watermark floor is then at least partially masked by the channel noise.

5 Conclusion

We presented a speech watermarking algorithm that makes use of the specific properties of speech signals and also exploits perceptual properties of the human auditory system. We showed the robustness of the system against AWGN attacks and also presented a synchronisation scheme for the analogue channel that performs within a range of one to two percentage points of raw bit error rate compared to the theoretical optimum at simulated ideal synchronisation. By limiting ourselves to a certain type of application we are able to allow complete non-robustness to certain types of attacks but can therefore achieve exceptionally high bit rates in comparison to the available channel bandwidth. Further refinement and testing is required to make the method robust against radio channel influences that are beyond AWGN and desynchronisation.

Our method is motivated by, but not limited to the described ATC application. It might also become useful for other legacy system enhancements that require backward-compatibility such as bandwidth extension for wire-line telephone systems, or for broadcast monitoring, archiving, or copyright monitoring for commercial text-to-speech systems. We also hope that the presentation of the potential aeronautical application will inspire further researchers to come up with even better methods for watermarking radio speech.

References

1. van Roosbroek, D.: EATMP communications strategy. Technical Description Vol. 2 (Ed. 6.0), Eurocontrol (2006)
2. Hering, H., Hagmüller, M., Kubin, G.: Safety and security increase for air traffic management through unnoticeable watermark aircraft identification tag transmitted with the VHF voice communication. In: Proceedings of the 22nd Digital Avionics Systems Conference (DASC 2003), Indianapolis, USA (2003)
3. Celiktin, M., Petre, E.: AIT initial feasibility study. Technical report, EUROCONTROL European Air Traffic Management Programme (EATMP) (2006)
4. SAFEE: Security of aircraft in the future european environment. `http://www.safee.reading.ac.uk/` (February 2007)
5. Hatada, M., Sakai, T., Komatsu, N., Yamazaki, Y.: Digital watermarking based on process of speech production. In: Proceedings of SPIE - Multimedia Systems and Applications V. (2002)
6. Geiser, B., Jax, P., Vary, P.: Artificial bandwidth extension of speech supported by watermark-transmitted side information. In: Proceedings of the 9th European Conference on Speech Communication and Technology EUROSPEECH. (2005)
7. Girin, L., Marchand, S.: Watermarking of speech signals using the sinusoidal model and frequency modulation of the partials. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). (2004)
8. Hofbauer, K., Kubin, G.: High-rate data embedding in unvoiced speech. In: Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH), Pittsburgh, PY, USA (September 2006)
9. Cox, I.J., Miller, M.L., Bloom, J.A.: Digital Watermarking. Morgan Kaufmann Publishers (2001)
10. Vary, P., Martin, R.: Digital Speech Transmission. John Wiley and Sons (2006)
11. Kubin, G., Atal, B.S., Kleijn, W.B.: Performance of noise excitation for unvoiced speech. In: Proceedings of the IEEE Workshop on Speech Coding for Telecommunications. (1993)
12. Boersma, P., Weenink, D.: PRAAT: doing phonetics by computer. [progr.] (2006)
13. Coumou, D.J., Sharma, G.: Watermark synchronization for feature-based embedding: Application to speech. In: Multimedia and Expo, 2006 IEEE International Conference on, Toronto, ON, Canada (July 2006) 849–852
14. Cassaro, T.M., Georghiades, C.N.: Frame synchronization for coded systems over AWGN channels. IEEE Transactions on Communications **52**(3) (March 2004)
15. Franks, L.E.: Carrier and bit synchronization in data communication–a tutorial review. IEEE Transactions on Communications **28**(8) (1980)
16. Proakis, J.G., Salehi, M.: Comm. Systems Engineering. Prentice Hall (2001)
17. Gardner, F.M.: Phaselock Techniques. 3rd edn. John Wiley and Sons Ltd. (2005)
18. Shayan, Y.R., Le-Ngoc, T.: All digital phase-locked loop: concepts, design and applications. IEE Proceedings F Radar and Signal Processing **136** (1989) 53–56
19. Kim, B.: Dual-loop DPLL gear-shifting algorithm for fast synchronization. Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on **44**(7) (July 1997) 577–586
20. Hofbauer, K., Hering, H.: An addendum to 'Noise robust speech watermarking with bit synchronisation for the aeronautical radio'. `http://www.spsc.tugraz.at/people/hofbauer/papers/Hofbauer_IH07_Addendum.pdf` (2007)
21. Hofbauer, K.: Audio demonstration files of 'Noise robust speech watermarking with bit synchronisation for the aeronautical radio'. `http://www.spsc.tugraz.at/people/hofbauer/ih07/` (2007)