THE QUEEN'S AWARDS FOR ENTERPRISE 2001

BBC

R&D White Paper

WHP 061

June 2003

DAB: an introduction to the Eureka DAB System and a guide to how it works

C. Gandy

Research & Development
BRITISH BROADCASTING CORPORATION

BBC Research & Development
White Paper WHP 061

DAB: an introduction to the Eureka DAB System and a guide to how it works

C. Gandy

Abstract

This document was originally produced in 1994 as a way of collecting together a lot of information about the then-new Digital Audio Broadcasting (DAB) system being developed as a European project (Eureka 147). At that time, the first public international standard had not been published, and as a result the document was highly sensitive and was released only within the BBC to enable staff coming to work on the project to quickly grasp the fundamentals of the DAB system.

Nearly ten years later, much of the information is still highly relevant to the DAB system now on air in many countries around the globe. Some parts of the DAB system have been modified compared to the 1994 Eureka specification, but usually in a backwards compatible manner. For example, Eu-147 DAB only specified three "RF Modes", but EN300-401 specifies a fourth, additional mode.

Within BBC R&D, this document is still used as an excellent reference by many working on DAB who neither need nor require the intricate details but do require a "broad-brush" understanding of the system. Whilst originally the document was highly confidential, the passage of time has seen all of the information contained move to the public domain, and so it has been suggested that the document should be re-issued to a wider audience. It must be noted, however, that the text is largely unaltered from the 1994 original and therefore may contain errors or omissions when compared to the latest DAB standards. In all such cases, the published international standards are definitive.

Key words: digital radio, Eureka 147, DAB,

White Papers are distributed freely on request.

Authorisation of the Chief Scientist is required for publication.

DAB: an introduction to the Eureka DAB System and a guide to how it works

C. Gandy

PART 1 - THE DAB SYSTEM

PART 2 - HOW IT WORKS

DAB: an introduction to the Eureka DAB System and a guide to how it works

C. Gandy

PREFACE

Because DAB is not yet an established broadcasting system, there are few sources of clear, working information about how the system operates and the nature of signals encountered in the DAB transmission chain. Much of the existing descriptive material is either very general, for instance when it is aimed at gaining international recognition for the system, or it concentrates on specific issues for research purposes, often with extensive use of mathematics.

This document aims to provide an explanation of how the DAB system works in a fair amount of detail but in relatively plain English, largely *without* recourse to mathematics. It is assumed that the reader has a working knowledge of FM, PCM and NICAM 728.

Otherwise, the most detailed documents available at the time of writing are a draft European Telecommunications Standard[1] (ETS), and the (confidential) System Definition produced by the Eureka 147 consortium which has developed the DAB system; but neither of these was written to explain how the system works. Each was written to specify the transmitted signal in a compact document, along the lines of a patent specification, and they contain almost enough information to enable the implementation of DAB hardware. However, to all but the most enlightened of engineers on their first reading they would probably be incomprehensible. This is because extensive use has been made of engineering 'shorthand'; succinct mathematical definitions, and even some computer language. Nevertheless, for their intended purpose, these documents are very well written.

In the future, it is expected that the Eureka consortium will issue 'guidelines for implementation and operation of the DAB system'. That will become the authoritative document, but its preparation is not yet complete.

This document is divided into two distinct parts in order to simplify the section numbering; no cross-references will be made between the two parts.

PART 1 - THE DAB SYSTEM provides an overview of the DAB system, what it offers to broadcasters and listeners, a brief history of its development, and some details of how it can be applied to broadcasting networks.

PART 2 - HOW IT WORKS provides a detailed explanation of how the DAB system works. Although the system is quite complicated, many of its features can be described in a logical progression starting from the main task that it was designed to tackle, that of overcoming the problem of multipath propagation. This approach will be taken here, and along the way, some aspects of receiver implementation will also be discussed.

---

[1] Draft prETS 300 401, Radio broadcast systems; Digital Audio Broadcasting (DAB) to mobile, portable and fixed receivers.

This document cannot be exhaustive and readable at the same time, so its scope will be limited to the use of the DAB system as a means for sound radio broadcasting. Other data-broadcasting applications, such as the transmission of extensive service information (similar to RDS, but greatly enhanced), will be treated only in outline. Some of the more-complicated techniques upon which the system relies will be explained in greater detail in appendices and, for those that have time to read them, these explanations may help to give a clearer understanding of some of the processes carried out in DAB hardware. Such explanations are necessarily limited to how these techniques work and an outline of how they can be implemented in hardware, but not why they are so effective; further reading material will be indicated for those who may wish to pursue this.

A large proportion of the material contained in this document can be found in published text books and ITU-R (formerly CCIR) Reports, and most of that which is specific to the DAB system can be deduced from the published draft ETS. A small proportion is currently considered Eureka proprietary material, but it would probably be impossible to provide a satisfactory explanation of the system without this.

PART 1 - THE DAB SYSTEM

1. OUTLINE DESCRIPTION

DAB, the Digital Audio Broadcasting system, is the development of a European consortium called the Eureka 147 DAB Project. The consortium comprises representatives of European research institutes, broadcasting and electronic manufacturing companies, including the BBC and the EBU.

DAB is a completely new means for broadcasting high-quality sound radio services to mobile, portable and fixed receivers which can use simple antennas. It is designed to operate in any frequency band in the VHF and UHF range for terrestrial, satellite, hybrid (satellite with complementary terrestrial), and cable delivery. The system uses advanced digital techniques to provide ruggedness, sufficient to combat severe multipath propagation to stationary or mobile receivers, yet it is highly efficient in its requirements for RF spectrum and transmitter power. Audio programme information is transmitted as a digital bit-stream, and the system can support a wide range of options for other data, either associated with or independent from the sound programmes.

The DAB signal occupies a bandwidth of about 1.5 MHz and uses a large number of discrete carriers, each independently modulated using QPSK (Quadri-phase Phase-Shift Keying). There are three different transmission modes, applicable to different ranges of radio frequency, and the number of carriers and several other system parameters depend on the mode. Transmission Mode 1 is most appropriate for a large network of terrestrial VHF transmitters, and in this mode the signal uses 1536 modulated carriers at intervals of 1 kHz on a regular frequency comb. Details of how the DAB system works, which will be given in later sections of this document, will be confined initially to Mode 1.

The DAB system is not compatible with existing AM and FM systems, or the NICAM 728 system used for stereo sound with television, but it is anticipated that when domestic receivers equipped with DAB become available they will also be capable of FM reception. This is important because it would be impractical for the BBC to delay the launch of a new DAB service until the stage at which a very large proportion of the UK was served by DAB transmitters. Although some features of DAB are common with NICAM 728, the similarity is little more than superficial. DAB takes advantage of more modern technology, it is vastly more complex, more flexible and the radio signal is much more rugged.

2. WHAT DAB OFFERS TO BROADCASTERS AND LISTENERS

DAB is seen by the EBU as one of the most important developments in broadcasting during the 1990s, and this view is propagating outside Europe to several other interested countries such as Canada. It is likely to replace most existing methods for radio broadcasting, and maybe even contribution links, in the long term.

The system offers solutions to many of the problems which beset FM radio, and can provide:

- **Consistent, high quality reception even in adverse propagation conditions**
  Overcoming the problem of multipath propagation, which greatly upsets FM reception in vehicles. The system is also resistant to continuous and impulsive interference, and simple, omni-directional antennas are sufficient for mobile, portable and fixed receivers.

- **Very high audio quality, approaching the quality of Compact Disc**
  Within the fixed total capacity of the DAB signal, the data rate can be divided between different services with the same, or different audio qualities; the highest quality available exceeds the requirements for broadcasting, and is suitable for contribution links. Flexibility is a key design feature of the system, and the division of the data rate can be changed dynamically (e.g. at programme junctions).

- **Very efficient use of available VHF or UHF radio spectrum**
  A transmitter network could be established over the whole of the UK using only one frequency allocation of about 1.75 MHz bandwidth (including guard-bands). If this carried 6 high-quality stereo services, the spectral efficiency would be more than 6 times greater than FM (which requires about 2.2 MHz total bandwidth for *each* UK stereo network). Also, the distribution of power in the bandwidth occupied by the DAB signal is much more uniform than for most conventional signals, so the potential for causing interference to other systems is greatly reduced.

- **Good coverage for moderate transmitter powers**
  In comparison with FM, considerably greater high-quality coverage can be obtained by DAB using the same transmitter ERP, but unlike FM, there is little 'fringe area'; the boundaries of service areas are much more precise.

- **Push-button controlled receivers which are easy to use**
  The 'all digital' signal requires receivers in which most functions are implemented digitally. The selection of which of the 6 (for example) services will be received is a digital de-multiplexing function which can easily be controlled by push-buttons. The use of a synthesised local-oscillator to select which DAB signal will be received (when several are available) represents no increase in the receiver technology, and removes the need for manual tuning.

- **Additional facilities not possible using analogue FM**
  Extensive service information facilities are available which can greatly outperform RDS. The possible uses of a data channel are limited only by imagination; some examples are traffic messages, paging, and even a Teletext-like service which could carry the programme listings contained in the Radio Times.

Individually, most of these features are consequences of the complete departure from existing analogue modulation methods, and the simultaneous availability of all of them would not be possible at all by analogue means. The first feature listed is probably the most important reason for considering DAB as the future for BBC radio broadcasting, but the others are

gaining in importance as investigations proceed on the practical application of the system in the UK. It must be emphasised that very little experience exists, in the world, let alone in the UK, of planning and implementing DAB transmitter networks.

3. HISTORY OF THE DEVELOPMENT OF THE SYSTEM

Fundamentally, the system has been designed as a flexible, general-purpose, 'integrated services' digital broadcasting system which can transport any kind of data within the overall capacity of the bit-stream; for example, it could be used solely for paging or for transmitting computer data. However, the Eureka 147 Project was initiated through collaboration between IRT[2] and CCETT[3], both of which undertake research on behalf of broadcasting organisations in their respective countries. There is a consensus amongst European broadcasters that sound radio broadcasting has the most pressing need for improvement, so this has been the main thrust of the work in the Project. The DAB bit-stream can also be used for slow-scan television, but other consortia are now researching the wider application of digital techniques to television broadcasting, including HDTV.

From the outset, in 1986, it was recognised by the BBC that DAB could offer the future for radio broadcasting, especially in view of the widespread interest within Europe. Thus, the BBC became a member of the Eureka consortium, and Research Department and Development Group (now combined as Research and Development Department) have made major contributions to the Project.

The embryo of a DAB system was created by the conjunction of two advanced digital techniques, audio bit-rate reduction, pioneered by IRT, and RF transmission using a technique known as COFDM (which will be described later), pioneered by CCETT; but a lot more work was needed to develop this into a usable broadcasting system. The BBC contribution has been diverse, including the third major component in the system; the dynamic, flexible multiplex and system control 'mechanism'. The BBC contribution also involved research into many aspects of the audio, data and RF parts of the system, as well as a major role in determining the final system parameters and drafting the written specification for the system.

As the system has evolved, parameters such as the bandwidth of the RF signal have been changed. Starting at 7 MHz to fill a continental television channel, changes have been made to 3.5 MHz, and then 1.5 MHz in order to fit 4 DAB signals, plus guard bands, into such a television channel; the final specification corresponds to 1.537 MHz bandwidth. During this evolution, extensive field tests of the system were undertaken by Research Department using experimental transmitters in London (Crystal Palace) and Birmingham, and latterly a mini-network of low-power transmitters at existing UHF television transmitter sites in Surrey, followed by a London-wide network. Three successive generations of experimental DAB transmitting and receiving equipment have been produced by the Eureka consortium

[2] Institut für Rundfunktechnik; the research and development institute for the German broadcasters ARD, ZDF, ORF and SRG/SSR.

[3] Centre Commun d'Etudes de Télédiffusion et Télécommunications, the research and development institute for France Telecom and the French broadcaster TDF.

(which includes manufacturers[4] such as Philips, Grundig, Bosch and Thomson), and purchased by the BBC for experimental work.

The development of the DAB system is now approaching completion (in Autumn 1994), and a detailed specification of the transmitted signal has been prepared and issued for public comment as a draft European Telecommunications Standard [1]. Third-generation prototype equipment has been built, conforming to a subset of the specification, and this is in use in the BBC experimental high-power DAB network in the London area.

It is worth keeping in mind that the development of DAB represents the accumulation of a vast amount of wisdom and experience, but also that it has been achieved within a limited time-scale. In many cases, the parameter values used by DAB have been chosen from several options. In some cases, the choices have been made on pragmatic grounds (i.e. if it works, why fix it?), and in other cases because of degrees of subtlety far beyond the scope of this document (and, perhaps, the comprehension of the author!). Much further 'optimisation' is undoubtedly possible but, at this stage, probably undesirable. The first generation of dedicated VLSI chips for domestic receivers is expected to become available during 1994, and following the release of the first series of bulk-manufactured receivers (expected early in 1995), it will be difficult to incorporate any major re-developments.

4. RADIO FREQUENCIES

The DAB system can be used at any radio frequency between about 30 MHz and 3 GHz. The top octave is most suitable for satellite delivery, and an allocation has been reserved internationally for satellite and complementary terrestrial[5] DAB services in the frequency range 1452 to 1492 MHz; in the so-called 'L-Band'. The ultimate future of sound radio broadcasting may indeed lie in satellite delivery, but in the UK, this frequency range is presently used for fixed terrestrial links and it will not become available for DAB on a primary basis until the year 2007. The need for improved radio services is perceived as more urgent than to allow us to wait until then, so the BBC approach for national network services is presently to pursue terrestrial delivery.

Lower frequencies are more appropriate for terrestrial delivery because line-of-sight transmission paths cannot be maintained and longer wavelengths promote diffraction[6] around obstacles. On that basis, the lowest possible frequency should give the greatest coverage for a given transmitter power, but Band I has the drawback of high levels of man-made interference; substantially higher than in most of the higher-frequency bands. The interference rejection properties of the DAB system can render such interference inaudible, but inevitably the coverage obtained for a given transmitter power is reduced.

---

[4] The UK VLSI design company Ensigma is in the process of joining the Project.

[5] In the ITU, this was intended to mean principally satellite delivery, with low-power terrestrial transmitters to fill in areas which are not adequately served, such as those shadowed by groups of tall buildings. However, in some countries, the current interpretation puts greater emphasis on the terrestrial aspect.

[6] Diffraction is what happens when an obstacle blocks the path of a radio wave; the wave is attenuated in the 'shadow region' but the degree of attenuation depends on the size of the obstacle in comparison with the wavelength. 'Optical' shadows are seldom encountered at VHF because of the relatively long wavelengths.

It has been suggested that DAB should eventually replace FM in Band II, but simulcasting would be necessary for a lengthy period whilst the public re-equips with DAB-capable receivers. In view of the current, and expected future, packing density of Band II, a new clear frequency would be needed initially for DAB in a different band; a so-called 'parking band'. The problem with this approach is the large deferred cost of re-engineering the DAB network to Band II because, after perhaps 15 years of simulcasting, it could be serving more than 90% of the UK population.

The more-desirable approach is to establish DAB in a different band, and to leave it there, perhaps giving up some of Band II in the long term when the majority of listeners have been attracted to the new services. In that case, the most realistic possibility for the UK is Band III which, although having been relinquished by the broadcasters, is not yet fully utilised for private mobile radio, and other purposes. In January 1994, the Trade and Technology Minister announced that the UK Government has decided to make available the frequency band 217.5 MHz to 230 MHz for terrestrial DAB. This 12.5 MHz bandwidth should be sufficient for seven DAB signals: one for BBC national networks; one for INR; and five for BBC and independent local radio services.

The use of Band III is being pursued with vigour in the BBC, and planning for a national network is being investigated by Research and Development Department. Theoretical work has demonstrated a trade-off between the number of stations in such a network and their ERPs. Whilst there is a fundamental upper limit to the geographical separation between transmitters, imposed by the DAB system itself, the lower limit is set only by cost. The optimum balance appears to lie at about 70 km separation, with Effective Radiated Powers (ERPs) of around 10 kW, for the bulk of a UK national network. The separation must be reduced if smaller ERPs are used, and another possible combination is 20 km separations with ERPs of about 1 kW.

With Band III, there remains a problem of international frequency co-ordination, because it is still used extensively for television in some of our neighbouring continental countries. Also, in France, the upper frequencies in Band III are reserved for military use. These factors are driving several other countries to pursue L-Band allocations for DAB, for terrestrial use possibly without complementary satellite delivery. The propagation of such short-wavelength L-Band signals is almost line-of-sight, and large numbers of transmitting stations may be needed to provide continuous coverage of large cities. It is notable that a French proposal for using L-Band DAB targets major roads and motorways rather than widespread coverage.

In the UK, the BBC may be obliged to use ERPs somewhat smaller than the optimum, especially near the South and East coasts, in order to achieve international frequency co-ordination. The role of the BBC as the public service broadcaster means that the objective would always be to offer a new service, ultimately, to a large proportion of the UK population, so urban and rural areas would need to be served, as well as motorways. Thus, in some areas, the BBC may be forced to use groups of stations with smaller separations.

5. NETWORK PLANNING

Planning of transmitter networks is generally an iterative process because it is more practical to predict the coverage of a given transmitting station, or a group of stations, than to specify a station given the required coverage. The means for predicting coverage is essentially a mathematical model of radio propagation, and the extensive calculations required to plot coverage maps are nearly always handled by a computer. To achieve accuracy, the model must take into account diffracted and reflected waves as well as the line-of-sight path, if one exists, and this introduces dependence on the radio frequency and characteristics of the signal such as its bandwidth.

In most practical environments other than open rural areas, the propagation scenario can be very complicated and it would be difficult to build up an accurate model on the basis of theory alone, so the results of practical measurements must be introduced. The accuracy and universality of the model improves as more and more measurement data are gathered, analysed and applied.

It is principally for this purpose that the BBC high-power experimental DAB network has been established. The network comprises four Band III transmitters located at existing BBC stations. The programme of measurements (from a vehicle, and in houses) will cover most of the types of environment encountered in London and the surrounding areas. The results will also be applicable to many other areas of the UK, with some notable exceptions such as the valleys in South Wales; temporary stations may be established to extend the measurement work into such areas in the future.

The required end-result is a plan for a transmitter network covering the whole of the UK, from which a phased introduction of DAB services can be planned. The BBC has made a technical announcement[7] of intent to begin DAB national services in September 1995; a formal public announcement of what BBC DAB services will be initiated is anticipated before the end of 1994.

[7] At the Plenary Session of the UK National DAB Forum, on 12th September 1994.

PART 2 - HOW IT WORKS

1. INTRODUCTION

In this section, simplified descriptions will be given of the principles employed in the Eureka DAB system to broadcast sound radio services. The finest details of the system are extensive and many are not amenable to being put into an easily readable form (especially those which concern the arrangement of data), so many of these will be avoided. In practice, such details are only evident in the programming of DSP chips or programmable logic arrays in the prototype equipment, and ultimately in the design of custom LSI devices. Also, matters of hardware implementation, including the design of receivers, cannot be covered here in great detail because many have yet to be decided.

Otherwise, the aim is 'to leave no stone un-turned'; to try to give as full an account of each stage in the transmission chain as is possible, within the constraint of a document of manageable size, in order to provide a clear understanding of the principles involved.

2. THE PROBLEM - MULTIPATH PROPAGATION

Most existing means for radio communication which can use simple omni-directional receiving antennas are affected adversely by multipath propagation, particularly in a changing environment such as when the receiver is located in a moving vehicle. The effect on broadcast FM radio is well known where, in built-up areas, even if there is sufficient mean field strength, mobile reception can be severely impaired by bursts of noise and audio distortion, and sometimes a fluttering effect on the audio signal.

In addition to a direct signal from the transmitter, the receiver is often presented with signals reflected and diffracted by buildings and the terrain. These can combine constructively or destructively in the receiving antenna as the relative lengths of the propagation paths change, or as the wavelength changes. Indeed, the sensitivity to the wavelength, or frequency, is the main reason for the audio distortion with FM signals.

Constructive addition can give up to 6 dB enhancement, for two signals of equal magnitude, but subtraction can cause complete cancellation. This phenomenon is known as fading, and multipath propagation is one of two mechanisms by which it can be caused; the other is blocking of the propagation path by obstacles. In the latter case, the problem can often be overcome simply by increasing transmitter powers, but this is not always true for multipath fading; destructive combination of two signals of equal magnitude causes complete cancellation regardless of the transmitter power. When an FM receiver is presented with insufficient signal power its own front-end noise is demodulated giving a burst of audio noise. In most common environments, reflected and diffracted signals usually have smaller magnitudes than a direct signal, but a direct line-of-sight path might not always be available. Most reflections occur by lossy mechanisms (i.e. *not* large sheets of metal), and reflected signals often have further to travel so they are subject to greater spreading losses.

The receiver performance can be enhanced by using a directional antenna, which is only appropriate to fixed reception, or a diversity system using more than one antenna, which may be applicable to a more-expensive vehicular installation. Indeed, the audio quality available from fixed FM receivers with rooftop directional antennas is very good, but the market for radio consumption has changed from what was anticipated at the start of BBC FM services. The widespread use of portable and mobile receivers now demands a means for delivery which offers the highest audio quality, in most environments within a service area, without the need for a complicated antenna system or one that may need adjustment.

On this basis, there is little that can be done to rescue an FM signal, or any other analogue signal, in the presence of severe fading or interference. Much of the damage is done in the receiving antenna, and there is little scope for 'post-processing' to rectify the situation. However, in broadcasting and many other fields of communication, a solution is being sought in the application of digital techniques, implying a radical departure from the classical methods of broadcasting. This introduces numerous problems for the broadcasters and the receiver manufacturers (none of which is technically insurmountable, nowadays), but it offers the great advantage of 'post-processing' in the form of error correction.

2.1 Error correction

The effect of fading and interference on a digital system is to introduce errors into the received signal, but improvements are possible by virtue of the numerical nature of the bit-stream. Error detection can be achieved in the receiver by sending a small amount of additional data derived from the original data, such as checksums. By sending further additional data, it becomes possible to correct errors.

Such additional data are often referred to as 'redundant' data, but this does not imply that they are not needed; only that they carry no new information. The trivial case would be simply to transmit all of the original data twice with independent checksums, so with the benefit of error detection a complete set of good data could be reconstructed. However, this would make relatively inefficient use of the available bit-rate, as there still remains a probability that some of the same bits could be in error in both of the received versions. Many more-complicated, and ingenious methods have been developed for such Forward Error Correction (FEC; 'forward' implies that action is taken at the transmitter), which make more efficient use of a given amount of redundancy.

Powerful FEC, by whatever method, requires the transmission of substantial amounts of redundant data which increases the demand for radio frequency spectrum. However, the Eureka DAB system achieves greatly improved ruggedness without sacrificing efficiency in its use of radio spectrum by applying in addition what could be called 'advanced' digital techniques. To explain how the DAB system works, we should first look more closely at the effects of multipath propagation, in both the time and frequency domains.

Incidentally, the word 'coding' is used widely in the context of error correction, and digital communications generally. Sometimes the intended meaning is the whole principle of encoding and decoding data, and sometimes it is used instead of 'encoding'. The former meaning will be applied here, and in order to avoid ambiguity, the terms 'encoding' and 'decoding' will be used where they are meant.

2.2 Time and frequency domains

The time and frequency domains provide different viewpoints for the same effects, a principle well known to anyone who has used an oscilloscope and a spectrum analyser to inspect the same signal. The oscilloscope allows inspection of the way that a signal voltage changes as time progresses, almost irrespective of the rate at which it is changing, whereas the spectrum analyser allows inspection of the content of the signal at different frequencies (i.e. rates of change), almost irrespective of time. Often, different aspects of a phenomenon being investigated can be visualised more clearly in one or other of the two domains. For example, the vestigial sideband in a conventional AM television signal can be inspected using a spectrum analyser, but an oscilloscope gives no obvious hint to its existence.

If the effects of a phenomenon are described mathematically in each of the domains, the two descriptions are related by the Fourier transform. For example, the frequency response of a filter is related to its time impulse response by the Fourier transform.

2.3 Time-domain effects

In digital communications, the term 'symbol' is used for the smallest distinguishable unit in time of data transmitted, and this may be just one bit, or more. Conventionally, the transmitted data remain static for the duration of each symbol, and changes occur between symbols at the 'symbol boundaries'. In radio transmission, the data are represented by the modulation of a carrier. For example, the QPSK modulation used in the NICAM 728 system has 4 possible phase states, so each symbol conveys 2 bits.

One effect of multipath propagation is to generate inter-symbol interference. When a direct signal and delayed 'echoes' arrive at the receiving antenna, there will be occasions when their modulation represents data from different symbols, previous as well as present. This is illustrated in Fig. 1 for the case of a single delayed signal, although there may be many in practice.

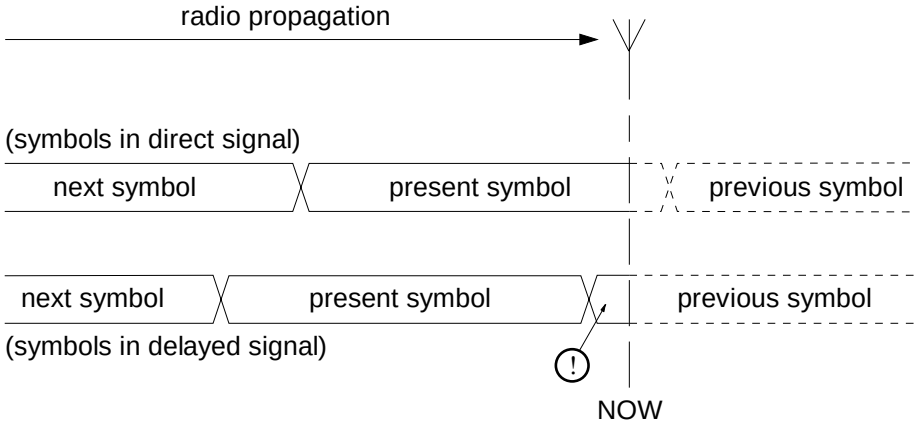


Fig. 1 - simultaneous reception of radio signals carrying two different symbols

Note that this diagram is unconventional in that events occurring at different times (i.e. symbols, in this case) are depicted as travelling from left to right with the passage of time, with the instant of reception (i.e. the present time; 'NOW') stationary. The intention is to illustrate the symbols, as the contents of radio waves, propagating across the page towards the receiving antenna. The alternative of showing a 'snapshot' of static events, with time as a variable increasing from left to right, is more in keeping with the usual method of plotting functions of time, but it may be less intuitive in this case.

When considering the operation of the receiver, it is convenient for this explanation to separate the demodulation and detection functions. The demodulator determines the modulation state of the radio carrier and outputs a signal to the detector. On the basis of this signal, the detector then makes the decision as to which of the expected modulation states was present, and outputs data bits accordingly.

For the example of QPSK modulation, the demodulator measures the phase of the carrier, and the detector decides to which of the four expected phases that measure corresponds. The detector usually introduces a degree of tolerance to small errors in the apparent modulation state, so for QPSK, demodulator signals representing phases within the same quadrant would yield the same output bits. The boundaries between the four quadrants are known as the 'decision boundaries'.

Returning to Fig. 1, when the combined signal is demodulated and the modulation state is detected, the effect of the overlapping different symbols is to cause corruption of the data; that is, inter-symbol interference. However, rather than making a 'snap' decision at one point in time during each symbol, as implied in the figure, in some cases it can be more efficient for the detector to integrate the demodulator output signal over the whole of each symbol, with respect to some timing reference (which may be the direct signal). This makes use of all of the received signal power.

In that case, as long as the delay is less than the symbol duration, the echo will carry the same modulation as the direct signal for some portion of integration period. Undoubtedly this will have some effect because the demodulator output signal represents the phase of whatever is input; in this case the 'vector' sum[8] of the direct and delayed signals. However, a method is available to prevent this from upsetting the operation of the detector; that is, differential modulation which will be described later. Entirely different symbols overlap for the remainder of the integration, so the degree to which the result is corrupted depends on the magnitudes of the echo signals and on their delays.

When a mobile receiver is travelling in a dense urban environment, which imposes one of the most difficult multipath conditions, short-delay echoes are often received in greater numbers than long-delay ones owing to multiple reflections (of a relatively direct signal) from local buildings and terrain. Very long delay echo signals tend to have smaller magnitudes because they have travelled greater distances. Therefore, the majority of the collective echo power arises from short delay echoes, so modulation schemes which use long symbols (or low symbol rates) provide the greatest tolerance to this effect because the proportion of each symbol that is corrupted is minimised.

[8] Strictly speaking, this should be the 'phasor sum' for the simple case where the direct and delayed signals have the same frequency.

2.3.1 Delay spread

Practical measurements can be made of the collective echo power received with different delays. An idealised result is illustrated in Fig. 2 where the transmitted (i.e. direct) signal is shown as a single impulse; in practice, more-complicated signals are used in such measurements to overcome the infinite bandwidth requirement of true impulses, but the results can be presented in a similar fashion.

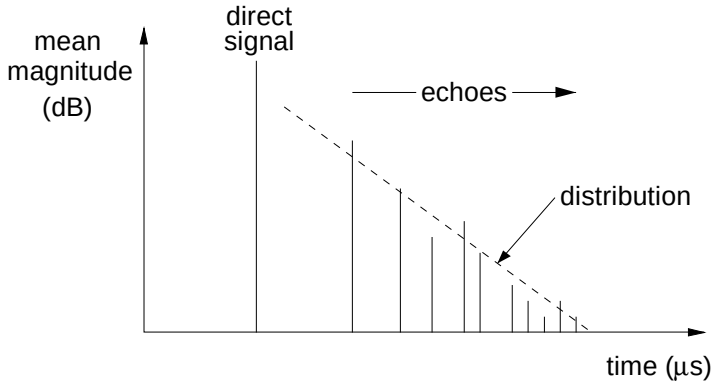


Fig. 2 - example of received echoes for the case of a transmitted impulse

By taking a large number of measurements in a particular type of environment, the statistical distribution can be built up, and in many cases this is found to approximate to an exponential curve. Such a distribution appears as a linear slope when plotted with decibel magnitude scaling, as shown in Fig. 2. The distribution is characterised by a single parameter known as the 'delay spread' [2], which can be interpreted as either the mean or the standard deviation in the case of an exponential distribution[9]; the latter corresponds to the slope of the line.

The incremental effects of greater delay or greater echo power are similar, they both increase the potential for inter-symbol interference, so the delay spread (interpreted as the mean) is a guide to the interference potential of a given type of environment. The average degree of data corruption is proportional to the ratio of the symbol duration to the delay-spread. For the case of outdoor reception from a single terrestrial transmitter, the delay spread can range from less than 0.5 μs to 5 μs, or more; a typical median value (for 50% of locations) is around 1 μs. Echoes with delays outside this range can be encountered, but the percentage of locations for which they contribute significantly to the delay spread is small (less than 1%). Direct reception from a satellite yields values of 0.5 μs, or less.

Thus, it was a pre-requisite for the DAB system that the symbol duration should be much greater than the values of delay-spread encountered in common broadcasting environments. In other words, the symbol duration should have a minimum value of at least 50 μs for terrestrial broadcasting.

---

[9] It is defined as the square-root of the second central moment of the distribution (i.e. the standard deviation), and for an exponential distribution $P_{(\tau)} = (1/T)\ e^{-\tau/T}$, where $P_{(\tau)}$ is the echo power at a delay τ, the delay spread is simply T, which is also the mean of the distribution.

2.4 Frequency-domain effects

If the same multipath effects are observed in the frequency domain, it is found that an uneven frequency response is imposed on the transmission channel, maybe even a comb-filter response in extreme cases, and this changes as the receiving antenna moves. This is intuitively obvious for the case of a direct signal and a single delayed signal arriving at the receiver. Constructive addition occurs at frequencies and receiver locations where the relative delay corresponds to an even number of half wavelengths of the radio signal, but partial or complete cancellation (depending on the relative magnitudes) occurs when the delay corresponds to an odd number.

Consider the propagation of two spectral components of a modulated signal. If their frequencies are very close, the echo delays will subject both components to similar phase shifts, so when they are combined with a direct signal in the receiver, the magnitudes of the components received at the two frequencies will vary in sympathy as the receiver moves. In other words, their fading will be correlated. If the frequency separation is increased, the degree of correlation will be reduced. Ultimately an echo with one particular delay could subject one component to 360° relative phase shift and the other to 180°, for example, and in combination with a direct signal (at 0°), one component would add constructively and the other would cancel.

2.4.1 Flat and selective fading

Thus, the nature of the fading caused by multipath propagation depends on the bandwidth of the signal. If the fading of all components of the signal is correlated, the result is known as 'flat fading', but if the effect on some or all components is not correlated (or negatively correlated, as in the ultimate example above), the result is 'selective fading'. Flat and selective fading are illustrated in Fig. 3.

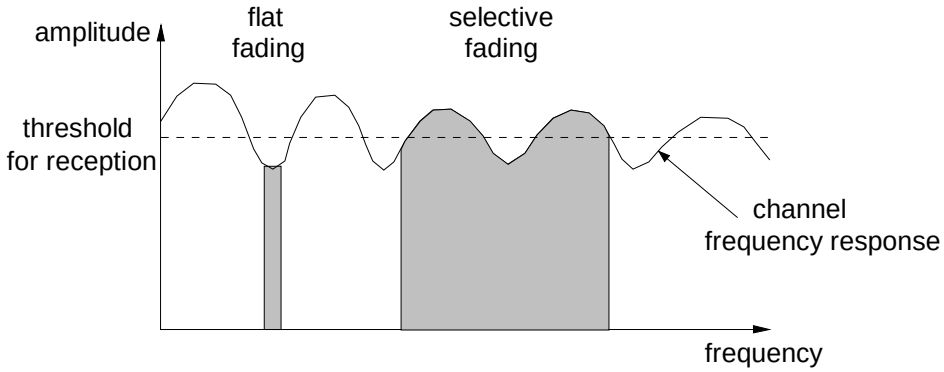


Fig. 3 - flat and selective fading

Relatively narrow-band signals, like FM or NICAM using a single carrier, are more susceptible to flat fading because the whole of the received signal can be severely attenuated. Impairment or errors occur when the resulting signal-to-noise ratio (S/N) in the receiver drops below the threshold for successful demodulation. On the other hand, when a relatively wide-band signal is subject to a typical multipath channel, the effect can be selective fading.

Some proportion of the signal power may always be receivable and, if the S/N is sufficiently large, it may be possible to demodulate the signal successfully. In this case, the received signal is likely to be distorted because of the uneven frequency response of the channel, and this can cause impairment or errors unless steps are taken to alleviate it. However, the total power of the received signal would be subject to smaller excursions than in the case of flat fading.

2.4.2 Correlation bandwidth

For given multipath conditions, the transition from flat to selective fading occurs at a signal bandwidth which provides a sufficiently small degree of correlation, and this is known as the 'correlation bandwidth'. Its precise value depends on the degree of correlation that is sufficiently small for the signal being considered (i.e. it depends on factors such as the power of any FEC applied). The 90% correlation bandwidth has been used in some early DAB development work [3], and in that case the extreme spectral components of a signal with this bandwidth would be subject to fading with 90% correlation.

The variation of the correlation coefficient with bandwidth has a particular distribution for given multipath conditions, and this is related to the distribution of echo power versus delay; it is, after all, no more than a description of the same multipath effects from the viewpoint of the other domain. The two distributions are related by the Fourier transform [1, 2] and the correlation bandwidth is proportional to the reciprocal of the delay spread. It has been found by experiment that the 90% correlation bandwidth is equal to approximately 9% of the reciprocal of the delay spread. For a typical median delay spread of 1 μs, this has a value of about 90 kHz, but possible values can range from less than 20 kHz to more than 1.5 MHz. Values outside this range are only encountered in a small percentage of locations (less than 1%), as is the case for delay spread. Note that the reciprocal relationship means that the worst case for flat fading, a large correlation bandwidth, is associated with a small delay spread; the opposite of the case for inter-symbol interference.

This gives another pre-requisite for the DAB signal; ideally, its bandwidth should be greater than the 90% correlation bandwidths encountered in common broadcasting environments so it will be subject to selective fading for most of the time. In other words, the bandwidth should be at least 1.5 MHz for terrestrial broadcasting.

The bandwidth of a digitally modulated signal is proportional to the symbol-rate. The exact relationship depends on the modulation scheme and factors such as filtering, but generally for a given modulation scheme, doubling the symbol rate doubles the bandwidth. The simple way to increase the bandwidth of the DAB signal without compromising its spectral efficiency would be to make it carry several radio programmes at the same time, by bringing together data representing a number of audio signals and multiplexing them before transmission.

However, this gives rise to a dilemma: if the DAB signal were to use a single carrier, then in order to achieve a wide bandwidth, the symbol rate would need to be high, but this conflicts with the requirement for long symbols.

3. THE SOLUTION - MULTIPLE CARRIERS

A solution is to use not one, but a multiplicity of carriers each at a different radio frequency. By modulating each carrier independently at low symbol rate by a small fraction of the data to be transmitted, individually the carriers will then be relatively resistant to multipath echoes because of their long symbols.

The requirement for mobile reception imposes an upper limit on the symbol duration. The changing characteristics of the transmission channel can have adverse effects on whatever modulation system is used, and generally, the maximum symbol duration is related to the required maximum vehicle speed. This topic will be considered in more detail later (in Sections 3.4.2 and 8.1); presently, it is sufficient to note that:

The maximum symbol duration chosen for the DAB system is 1 ms

In isolation, this allows good reception at vehicle speeds of at least 100 km/hr. Of course, 1 ms goes well beyond the requirement for tolerating multipath echoes and the reason for this will also be explained later (in Section 7.).

By making the bandwidth occupied by the group of carriers greater than all likely values of correlation bandwidth, a 'frequency diversity' advantage is introduced. It is found that the resistance to multipath effects improves as the bandwidth is increased, accommodating more extreme multipath conditions which could otherwise cause flat fading.

The actual bandwidth chosen for the DAB signal is 1.537 MHz

However, this is a compromise to enable four DAB signals to be fitted into a 7 MHz bandwidth continental television channel; a somewhat greater bandwidth might have been chosen on performance grounds alone.

Clearly, the greater the number of individually modulated carriers that can be packed into the given bandwidth, the greater the potential data capacity of the signal, but the upper limit is set by the requirement for independent demodulation without mutual interference. The significant bandwidth occupied by each modulated carrier is determined by the chosen symbol rate and the modulation scheme, and a simple way to avoid mutual interference would be to separate adjacent carriers by frequency guard bands. However, such a simple Frequency-Division Multiplexing (FDM) approach would be wasteful of RF spectrum. Without guard bands, the spectra of adjacent modulated carriers are likely to overlap, and the allowable degree of overlap depends on the method of demodulation, so this establishes a maximum packing density, or a minimum separation between the carrier frequencies. In either case, some form of spectrum analysis technique is needed in the receiver.

The DAB system uses a technique known as Orthogonal Frequency-Division Multiplexing (OFDM) which allows the greatest possible packing density consistent with the use of practicable (mainly digital) processing techniques. This requires the minimum separation of the carrier frequencies to be equal to the reciprocal of the symbol duration, 1 kHz, so the spectra of adjacent modulated carriers will certainly overlap. In that case, the maximum number of modulated carriers would be 1537, but in practice one carrier is not used (this is explained in Appendix 1), so:

The maximum number of modulated carriers in the DAB signal is 1536

In this scheme, the possibility arises that data conveyed by some of the individual carriers will not be received successfully because of selective fading, but in this case the application of FEC is most efficient (and necessary) because the loss of a small number of carriers represents the loss of only a small fraction of the total data. Thus, for the same degree of protection, the amount of redundant data that needs to be transmitted is much smaller than would be required for scheme using a single-carrier. For the DAB system, the abbreviation OFDM is prefixed with a 'C', for coded, to indicate the application of FEC, giving 'COFDM'.

It is worth noting that COFDM is not the only possible solution to the problems of multipath propagation. Spread spectrum techniques have been developed for this purpose, but their spectral efficiency would be considerably lower.

3.1 OFDM generation

OFDM is a method by which closely-packed carriers can be modulated and demodulated without mutual interference (i.e. crosstalk). Generation of an OFDM signal is easily visualised because, in principle, it could be carried out by partly-analogue means using a large bank of synthesised oscillators followed by modulators (i.e. multipliers). This principle is illustrated in Fig. 4 where three, of many, oscillators are shown although it should not be inferred that such a cumbersome arrangement is used in practice.

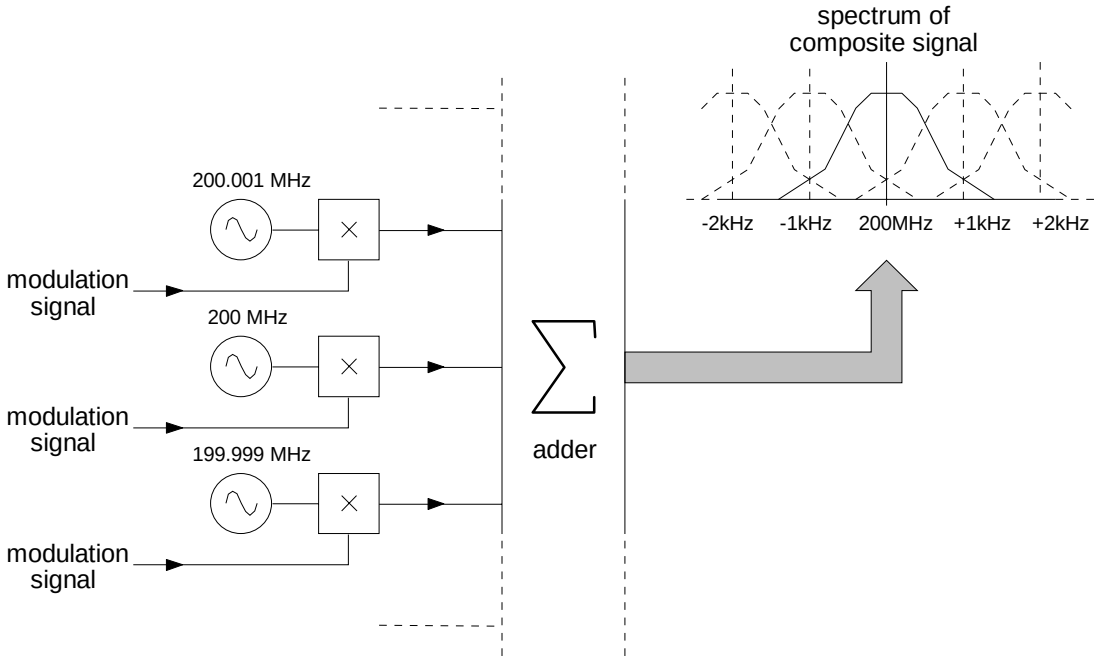


Fig. 4 - generation of an OFDM signal

Each oscillator provides one of the carriers, and the modulation is applied by each multiplier. All of the modulated carrier signals are then added together to make up the composite signal. The addition (or summation) can be considered as taking effect in the frequency domain; all of the frequency components produced by the multiplications are combined without affecting the time-waveform of any component. If the modulation signals are composed of symbols, such that changes only occur at the symbol boundaries then, during each symbol, each modulated carrier is temporarily a sinusoidal wave with a particular phase and/or amplitude representing the modulation.

However 1536 such oscillators could occupy a large room, whereas the current third-generation experimental DAB encoder occupies only part of a 6U rack cabinet. The key to compact digital implementation lies in the recognition that this process parallels the operation of a mathematical process known as the inverse Discrete Fourier Transform (iDFT). The iDFT is a method of calculating the waveform of a signal for which the spectrum is known. The iDFT operates with time-domain and frequency-domain variables which must be expressed as series of discrete samples.

In this case, the array of modulation signals which are to be applied to the carriers during a single symbol provide a specification of the spectrum of the required composite signal for that symbol. The modulation of the carriers is intended to remain static during each symbol, so each modulation signal contains one sampled value per symbol. The array of modulation signals can then be thought of as a series of samples which make up a 'function of frequency'.

The composite OFDM signal will be produced by the iDFT as a series of samples which follow one another in time, so it can be thought of as a 'function of time'.

The array of oscillator signals can be thought of as a function of both frequency and time; the array of different frequencies forms a series with respect to frequency (as with the modulating signals), and if each oscillator provides a sine wave, a sampled version of this forms a series with respect to time.

With this nomenclature, the iDFT (and the contents of Fig. 4) can be expressed as:

$$\text{function of time} = \sum_{\text{lowest frequency}}^{\text{highest frequency}} \text{function of frequency} \times \text{oscillator signal}$$

To produce the first sample of the 'function of time', each of the modulation 'samples' is multiplied by the first time sample of its corresponding oscillator signal, and all of the products are added together. The Greek sigma (**Σ**) indicates the summation, over all values of carrier frequency. For the second sample, the second time samples of the oscillator signals are used, and so on. In this way, any number of samples of the 'function of time' can be produced which represent the composite OFDM signal for one symbol. In practice, this number is minimised in order to constrain the demand for processing, and the fundamental limit is set by the so-called Nyquist criterion; the time-sampling rate must be at least twice the frequency of the highest frequency component represented in the function of time to avoid 'aliasing' distortion.

The period of time over which the whole calculation is performed can be called the 'processing time-window', and it is a requirement for correct operation of the iDFT that the duration of this window and the interval between the oscillator frequencies should have a reciprocal relationship. In this case, the required time-window (i.e. symbol) duration is 1 ms, so the oscillator frequencies are separated by 1 kHz. The required number of carriers, 1536, defines the highest frequency represented in the function of time, so the minimum time-sampling rate is then defined. In reality, the situation is a little more complicated than this brief description, and further details can be found in Appendix 1. The iDFT process is repeated for subsequent symbols, in each case with a new set of values in the array of modulation signals.

It is relatively straightforward to implement this transform in a computer program, or by means of digital hardware of some other form, where all of the sample values are represented by digital numbers. Furthermore, the availability of fast ADCs and DACs means that, nowadays, it is entirely practicable to carry out symbol-by-symbol processing in real time. With modern VLSI technology, the complexity (i.e. the number of carriers) is of relatively minor importance.

3.2 Recovery of modulation signals from an OFDM signal

The recovery of the modulation signals from an OFDM signal (i.e. 'decomposition' of the OFDM signal) is rather less straightforward, but essentially this follows from the generation process by interchanging time and frequency. It was mentioned earlier that integrating over each symbol is an efficient way to determine the modulation state of a carrier, and an extension of this principle provides a useful starting point.

To simplify the explanation, the receiver should initially be visualised as containing a large bank of local-oscillators, mixers (i.e. multipliers) and integrators although, as before, such an arrangement is not used in practice. Each oscillator/mixer combination acts as a demodulator, in the manner of a direct-conversion radio receiver. The incoming signal is fed equally to all of the demodulators, and each of these is followed by an integrator. Each integrator operates over a limited period of time before yielding a result. This is illustrated in Fig. 5.

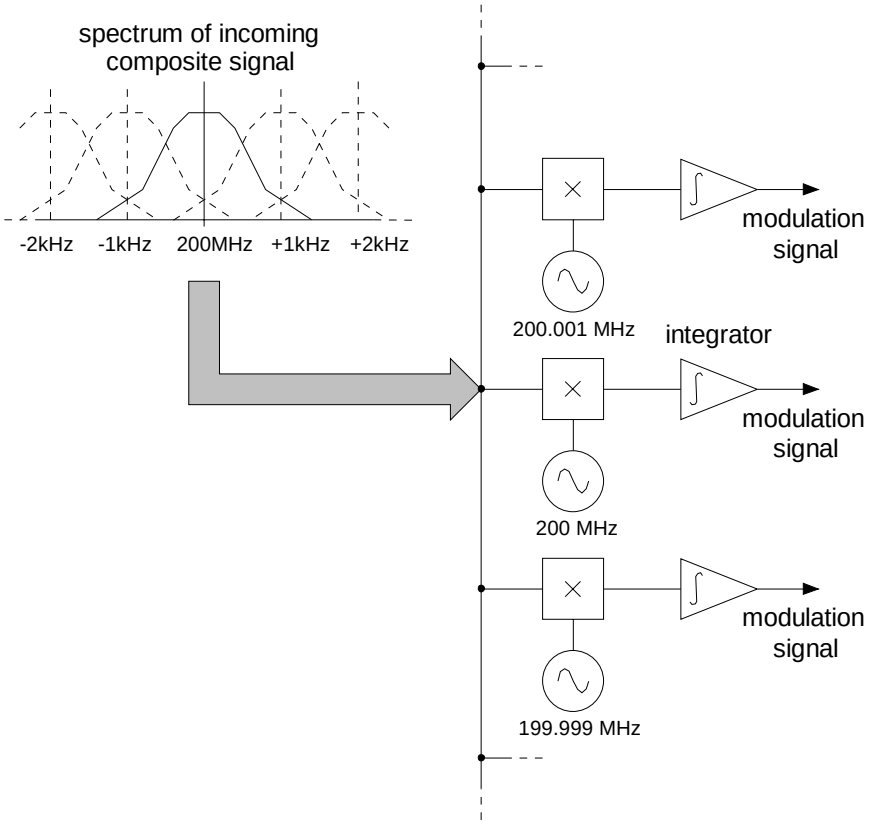


Fig. 5 - decomposition of an OFDM signal

Each modulated carrier is demodulated by the mixer which is fed with a local-oscillator signal of the corresponding frequency but, since the spectra of adjacent modulated carriers are allowed to overlap, each integrator will be presented with interference (or crosstalk) contributions as well as the wanted demodulated signal. However, if the radio frequencies could be chosen so that over the period of integration, the symbol duration, the integrals of all the interference signals amounted to zero, then mutual interference would be cancelled.

This condition is known as 'orthogonality', and is achieved when the carrier frequencies and the local-oscillator frequencies are located on a regular comb where the frequency interval is equal to the reciprocal of the symbol duration[10].

With reference to Fig. 6, take for example the modulated carrier at 200 MHz, with neighbours at ± 1 kHz, ± 2 kHz, etc. either side. The modulation state of each carrier is held constant over each symbol, so each carrier is temporarily a sinusoidal wave with a particular phase and/or amplitude representing the modulation. When the incoming signal is acted upon by the mixer with the 200 MHz local-oscillator, the 200 MHz wave will produce a DC output signal, and contributions from the neighbours will produce 1 kHz, 2 kHz, etc. AC components (the 400 MHz products will be neglected). When the composite signal output by this mixer is integrated over 1 ms, all of the AC components will cancel because they contain whole cycles, but the DC signal will accumulate to produce an output signal representing the modulation state of the 200 MHz carrier alone.

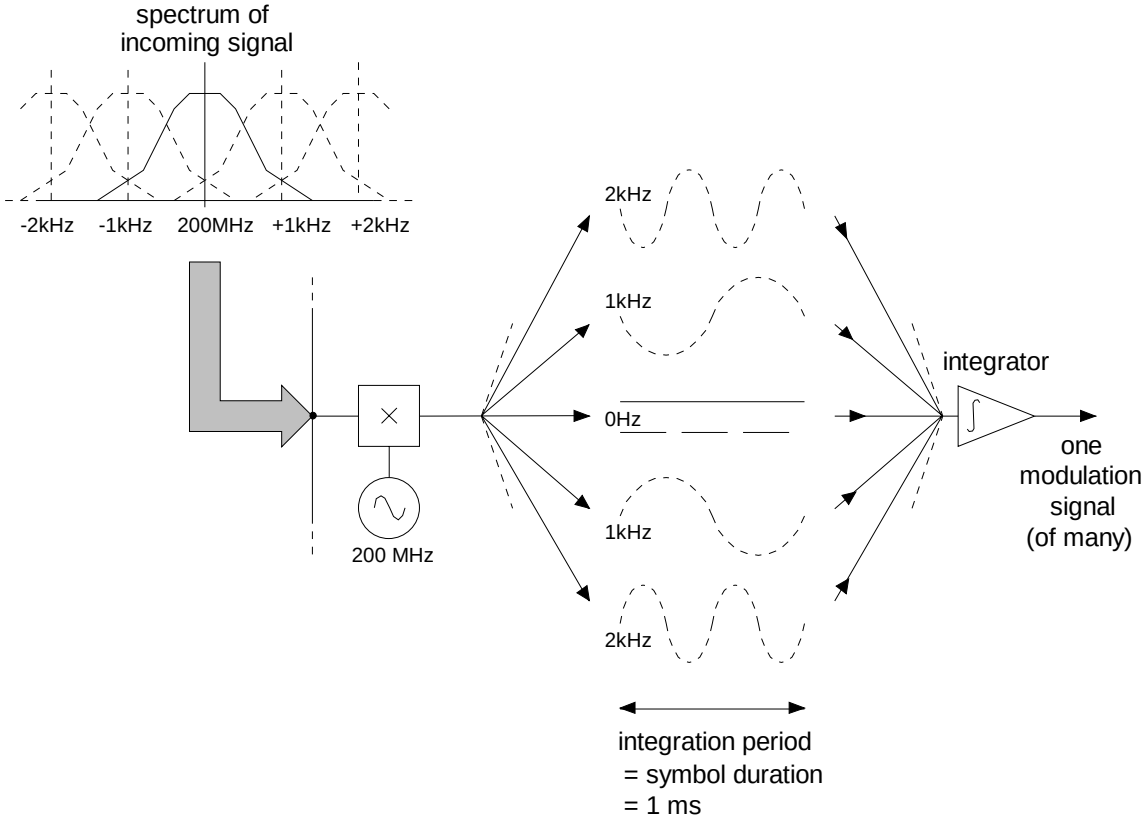


Fig. 6 - demodulation without mutual interference

[10] But not necessarily the symbol rate; consecutive integration periods do not need to be contiguous, there could be pauses between them.

A similar argument applies for each of the other carriers, with input frequencies and local-oscillator frequencies separated by the reciprocal of the symbol duration. The overall effect of this process is to analyse the spectrum of the incoming signal, and to output numerous signals each representing the modulation of one of the carriers.

If analogue processing were relied upon, the acceptable complexity of a domestic receiver would probably limit the maximum number of carriers to as few as 16 but, once again, the solution is to represent the process digitally. It is probably not surprising that the decomposition process has a direct counterpart in mathematics known as the forward DFT, or simply the 'DFT'. The DFT is the digital counterpart of the well-known Fourier Transform which relates the time and frequency domains, but the input and output functions of the DFT are series of discrete samples rather than continuous signals. The DFT is related to the iDFT essentially by interchanging time and frequency.

The incoming OFDM signal can be sampled in time and the series of samples can be thought of as a 'function of time'. As before, each of the modulation signals contains one sample per symbol, so the array of modulation signals can be thought of as a series of samples which make up a 'function of frequency'; that is, a description of the spectrum of the incoming composite signal. The array of oscillator signals can be thought of as a function of both time and frequency, as before.

The action of integrating each output signal corresponds, in discrete terms, to the summation of numerous consecutive discrete samples in time, so the decomposition process (and the contents of Fig. 5) is modelled by the DFT which can be written as:

$$\text{function of frequency} = \sum_{\text{first time sample}}^{\text{last time sample}} \text{function of time} \times \text{oscillator signal}$$

To produce the first sample of the 'function of frequency', that is, the modulation signal from the first (e.g. the lowest-frequency) carrier, each sample of the incoming 'function of time' is multiplied by the first frequency sample of the array of oscillator signals (e.g. the lowest-frequency one), and all of the products are added together. In this case, the **Σ** indicates summation, or integration, over all time samples. To recover the second modulation signal, the second frequency samples of the oscillator signals are used, and so on.

As before, the minimum time-sampling rate is set by the Nyquist criterion and the processing window duration (1 ms) must be equal to the reciprocal of the carrier frequency separation (1 kHz). Thus, a fixed number of samples of the 'function of frequency' can be produced which represent the modulation signals recovered from the individual carriers. The process is then repeated for subsequent symbols to yield the subsequent sets of modulation signals.

It is within the scope of modern VLSI technology to perform this decomposition process within one integrated circuit and, of course, in such a 'fully-digital' system as DAB, it is unnecessary to provide *additional* ADCs and DACs purely for these tasks.

Despite its name, the DAB radio signal is really an *analogue* signal; simply a voltage (or an electromagnetic field) varying with time; the digital aspects are the processes by which it is generated and decomposed.

3.3 OFDM processing by means of an FFT

In practice, an algorithm (i.e. a means for performing a computation which yields the same result) is used to perform the DFT in the receiver, and this is known as the Fast Fourier Transform, or FFT. An inverse FFT is used to generate the OFDM signal in the transmitter. The advantage of the FFT is increased processing speed for a given level of complexity. An FFT operates with complex numbers, in digital form, which represent the amplitudes and phases of its sampled input and output signals; the multiplications to which the last two sections have referred are actually complex multiplications.

A principal difference from the DFT is that the number of time samples must be equal to the number of frequency samples. This means that if *all* of the available samples are used, then the time-sampling rate can only just satisfy the Nyquist criterion. In most practical implementations of an FFT, the number of time or frequency samples is made equal to 2 raised to some power, so a 2048-sample FFT is used to process the 1536-carrier DAB signal. In that case, some 'headroom' is provided against aliasing.

When an FFT processor is presented with a digital representation of a time-domain signal (i.e. of a voltage varying with time), it has the effect of analysing the spectrum of the signal and it outputs numerous baseband signals, each corresponding to a particular range of input frequencies. Each baseband signal represents the amplitude and phase of whatever component of the signal is present in that particular frequency range. Thus, the function of the FFT can also be visualised as that of a bank of band-pass filters, followed by frequency down-converters. The effective filter bandwidths are contiguous and are each equal to the reciprocal of the duration of the processing time-window. The centre frequencies of the pass-bands are integer multiples of this reciprocal. The frequency response of each filter has a $\sin f / f$ shape, where f represents the relative frequency with appropriate scaling. By making the window 1 ms long, each bandwidth becomes 1 kHz and the centre-frequencies fall on a regular 1 kHz comb. The frequency response of each filter exhibits nulls at ± 1 kHz, ± 2 kHz, etc., and this accounts for the cancellation of inter-carrier interference noted earlier.

Therefore, it should be clear that the absolute frequencies of the carriers presented to the FFT processor, and their frequency separation, are intimately related to the symbol duration, and that any divergence from this relationship will cause some loss of orthogonality (i.e. crosstalk between carriers).

All of these features apply equally, but in a reversed sense, to the inverse FFT used in the transmitter. The absolute carrier frequencies and their separation are automatically related to the reciprocal of the symbol duration. Of course, it is possible to specify the spectrum of the OFDM signal such that certain carriers are suppressed (i.e. their amplitudes are set to zero), and this is done for the remainder of the 2048 frequency samples beyond the 1536 that are used for the DAB signal. The spectrum could also be configured such that every other carrier was suppressed, so the relationship between the frequency separation and the reciprocal of the symbol duration need not be 1:1, it could be 2:1 or some other integer ratio. However, a 1:1 relationship provides the greatest possible packing density consistent with the facility for independent demodulation and makes the greatest use of the available processing power.

An alternative, and more-detailed, explanation of the operation of the DFT and the FFT can be found in Appendix 1.

3.4 QPSK modulation and its detection

Because the FFT operates with complex numbers, its use in the generation and decomposition of an OFDM signal allows a choice of method for modulating the carriers. It should not be inferred from Figs. 4, 5 and 6 that the approach is limited to double-sideband suppressed-carrier AM; all of the multiplications are complex. In the DAB system, the chosen modulation method is QPSK so the modulation is carried only by the phases of the individual carriers; their amplitudes are essentially constant and equal. Other methods can be applied to this kind of system with different results, for example, BPSK, 8PSK, 16 QAM, etc. The lower-order methods are more rugged and higher-order methods can offer greater capacity for a given bandwidth.

Phase demodulation is provided by the FFT in the receiver, where the apparent phase of each carrier can be found with a little manipulation of the output complex numbers. This is described in Appendix 1.

A straightforward way to detect the QPSK modulation would be to establish a phase reference in the receiver and to compare the results of demodulation with that reference; the principle known as 'coherent' detection. However, the phase reference would need to be updated frequently to compensate for changing propagation delay as a mobile receiver moves. Such a system would exhibit 'inertia' in conditions where updates were missed because of deep fading, and the result could be erroneous detection for prolonged periods. Nevertheless, this approach may be suitable for static reception (e.g. of digital television).

3.4.1 Differential detection

The DAB system uses a different method whereby the QPSK modulation on each carrier is applied differentially; that is, 2 data bits are signalled by the *change* of phase of a carrier at each symbol boundary, rather than the absolute phase. Detection can be achieved in the receiver by storing each output from the FFT for one symbol and comparing, in some way, the new value with the previous value. This avoids the need for an absolute phase reference, which can simplify the receiver implementation, and correct operation resumes quickly after a deep fade as soon as two consecutive symbols have been received successfully. The disadvantage is impaired performance in the presence of noise and interference (i.e. the previous received symbol could be in error). Up to 3 dB greater S/N is needed to match the performance of coherent detection in the absence of fading.

The data to be transmitted are differentially encoded by treating pairs of bits as complex numbers, and a complex number can be used directly to represent an angle, or phase[11]. Whatever complex number was applied to the modulation of a carrier (i.e. its phase) during the previous symbol is multiplied by the new number to be transmitted, and the result is applied to the modulation for the new symbol. The angle of the result is the sum of the angles represented by the previous modulation and the new number, and this defines the phase for the new symbol. Thus, the 2-bit value of the new bit-pair determines the change of phase.

[11] If a complex number is written as $Re + j\,Im$, where Re and Im are the real and imaginary parts, respectively, the angle that this represents is $\tan^{-1}(Im / Re)$.

Alternatively, the value of a bit-pair can be considered as being conveyed by the ratio of the complex numbers represented by the modulation of a carrier during two consecutive symbols. Since the carrier has constant amplitude, the value of the bit-pair is conveyed by the phase difference. In the receiver, QPSK detection and differential decoding are accomplished simultaneously by dividing the number output by the FFT for the new symbol by that from the previous symbol (i.e. the represented angles are subtracted).

In its simplest form, differential QPSK (i.e. D-QPSK) signals one of the four modulation states by no change of phase at the symbol boundary, another by a 180° change, and the other two by ±90° changes. A 180° change introduces an amplitude discontinuity (i.e. an instantaneous dip in the envelope of the signal), and this can be difficult to preserve accurately when the signal is amplified with less-than-ideal linearity. If this detail is not reproduced accurately in the received signal, the distortion can reduce the ruggedness of the modulation, and this can impair the ability of a receiver to tolerate added noise.

In the DAB system, the chosen method of modulation is '$\pi/4$ offset D-QPSK', where the offset is 45° (i.e. $\pi/4$ in radians). The four modulation states are signalled by ±45° and ±135° changes of the carrier phase at the start of each new symbol, and this avoids 180° phase changes. In practice, this requires only a little further manipulation of the complex numbers. In absolute terms, there are eight possible phase states, four of which are available in any one symbol, and the phase reference (from the previous symbol) rotates by multiples of 45° from symbol to symbol. These features are illustrated in Fig. 7, where the phase of one carrier is shown during three consecutive symbols; the actual phases shown are examples of many possible combinations.

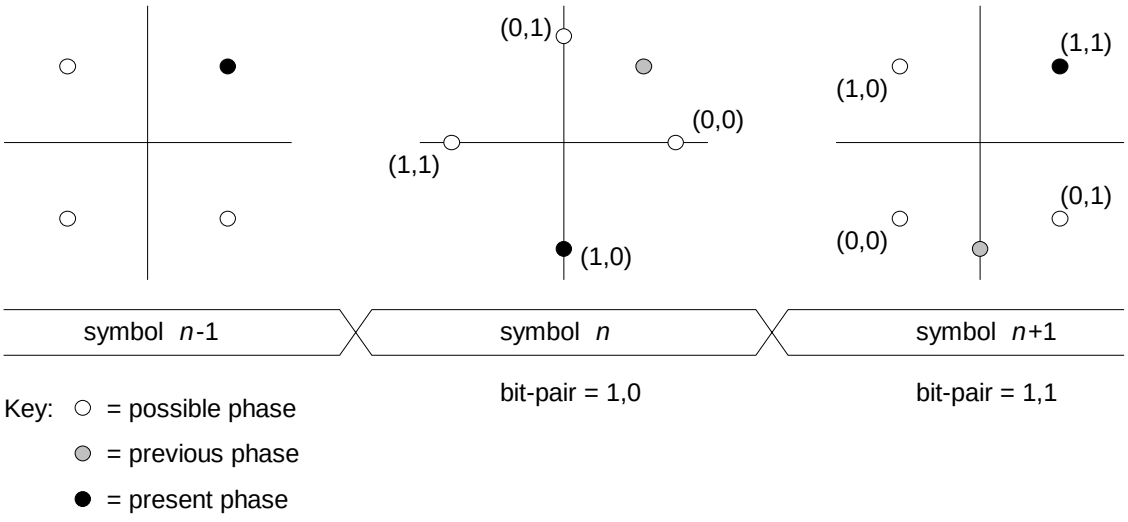


Fig. 7 - $\pi/4$ offset QPSK

3.4.2 Temporal coherence

Any modulation scheme in which information is sent in discrete symbols introduces a requirement for 'temporal coherence' of the transmission channel. In this case, the requirement is that the phase response of the channel must not change significantly from one symbol to the next, otherwise the apparent phase of the received signal will be modified and the ruggedness of the modulation will be impaired.

A simple example of incoherence occurs in the reception of a single modulated carrier which propagates via a single direct path, whilst travelling in a vehicle away from the transmitter. The received signal is subjected to a progressively increasing time delay as the path length increases, which is equivalent to a progressively increasing retardation of its phase. This corresponds to a reduction of the signal frequency; that is, the well-known Doppler frequency shift. From one symbol to the next, the effect is a displacement of the phase from the expected value. The magnitude of the Doppler shift, and the phase displacement, is proportional to both the vehicle speed and the radio frequency.

Of course, in such a simple case, steps could be taken in the receiver to compensate for the Doppler shift, but in typical mobile reception conditions the received DAB signal may contain many multipath contributions propagating over paths with different angles of reflection, giving rise to Doppler shifts of different magnitudes and even different signs (i.e. increased or decreased apparent radio frequencies). Furthermore, contributions arriving via paths of different lengths may be subject to differential fading and this can give rise to additional frequency shifts, usually accompanied by large variations in the magnitude of the resultant. It is not practical to compensate for a large number of these effects simultaneously and adaptively, so the communication system must be able to withstand some degree of temporal incoherence.

3.4.3 Doppler power spectrum

The distribution of signal power versus Doppler 'frequency' (i.e. shift) is known as the Doppler power spectrum, and it has been found that in cluttered environments (e.g. urban) this can contain components at up to twice the frequency that would be expected for the Doppler shift of a direct path alone [3]. The Doppler power spectrum is characterised by a single statistical parameter known as the 'Doppler spread'; a similar type of parameter to the delay spread.

Just as the distribution of echo power versus delay has a counterpart in the frequency domain, as noted in Section 2.4.2, the Doppler power spectrum has a counterpart in the time domain, with which it, also, is related by the Fourier transform. This is the distribution associated with the variation of the correlation of fading with time and, again, this is a description of the same multipath effects from the viewpoint of the other domain. Essentially, in a slowly changing channel, the fading caused by multipath propagation (over a given bandwidth) is correlated to some degree from one symbol to the next over a long period, and the Doppler spread is small. With increased speed of motion, that degree of correlation is maintained for less time, and the Doppler spread is increased. This leads to another reciprocal relationship, between the Doppler spread and the correlation time. On the basis that the phase response of the channel should be correlated from one symbol to the next, at least within the bandwidth occupied by a single QPSK signal, this establishes a maximum symbol duration for a given Doppler spread.

The Doppler power spectrum, and therefore the Doppler spread, is scaled by the speed of the mobile receiver and the radio frequency, so the maximum symbol duration is proportional to the product of the speed and frequency. Alternatively, a given symbol duration establishes a trade-off between the maximum speed and the maximum frequency. This topic will be considered further in Section 8.1.

3.4.4 Soft decision

The arithmetic involved in the computation of the FFT needs to have considerable resolution (e.g. 16 bits) in order to make full use of the orthogonality principle, so the fine detail of the demodulated phase of each carrier can be preserved. The results of the differential detection can retain some of this resolution, so they are multi-valued numbers rather than simple '1' or '0' bits. This principle is known as 'soft decision', and it can be used to enhance the performance of the error correction process which follows (i.e. the values of the numbers can be used to determine with what certainty the data have been received); this is explained in Appendix 2. Note that this has no counterpart in OFDM generation because just two bits are sufficient to specify QPSK modulation.

4. THE BASIC SIGNAL PATH

The group of 1536 carriers is known collectively as an 'ensemble', and the carriers can be viewed as 1536 parallel communication channels, each able to carry a small fraction of the total data. It could be said that 1536 symbols are transmitted simultaneously during each 1 ms symbol, making one 'symbol-block'[12]. QPSK conveys 2 data bits on each carrier during each symbol, so on this basis the gross capacity would be 3.072 Mbit/s. However, after subtraction of some overheads for receiver control and synchronisation, and for the addition of the so-called 'guard interval' (which will be discussed later, in Section 6.9), the bit-rate available for programme services is actually about 2.3 Mbit/s.

At this point, to avoid possible confusion, it is probably worth outlining the signal path through those elements of the DAB transmission chain which have been identified so far. A simplified DAB transmission chain is illustrated in Fig. 8.

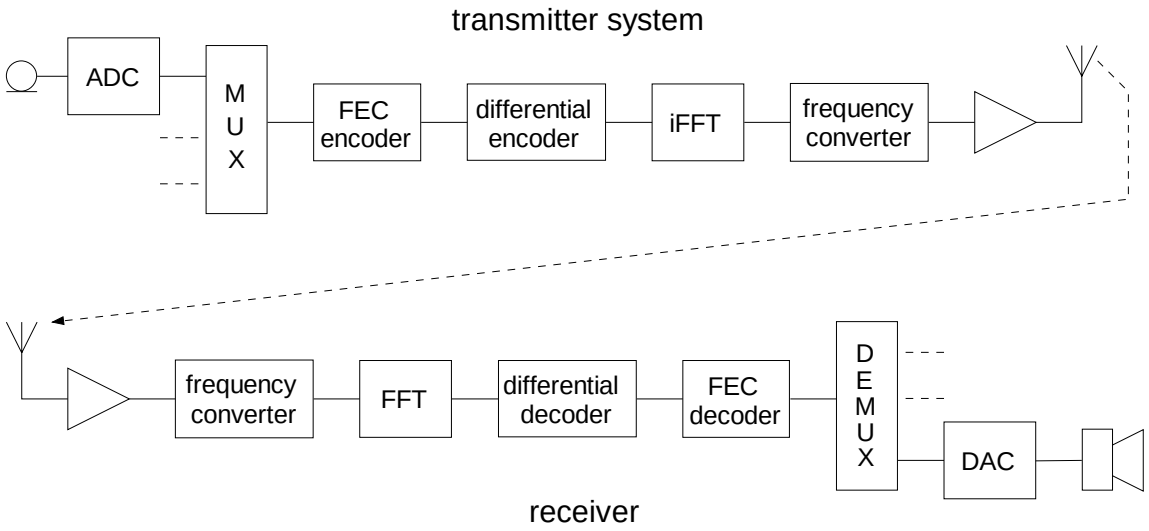


Fig. 8 - a simplified DAB transmission chain

[12] Strictly speaking, this should be one 'symbol-ensemble' because the term 'block' has been superseded in most other cases by 'ensemble'.

The chain of events is as follows:

(a) Audio programme signals are digitised and multiplexed together with ancillary data to produce an ‘un-coded’ bit-stream.

(b) The bit-stream is then encoded for forward error protection by adding redundant bits with appropriate, calculated values.

(c) During each consecutive 1 ms symbol, the ‘coded’ bits are divided into 1536 pairs, and each pair is differentially encoded with respect to its counterpart for the previous symbol.

(d) The 1536 differentially encoded bit-pairs are presented to an inverse FFT where each is used to define the phase of a QPSK carrier; collectively, they specify the spectrum of a 1536-carrier signal.

(e) The inverse FFT synthesises a time-domain signal which has the specified spectrum, and this signal is converted to analogue form, frequency converted then transmitted. This is the OFDM generation process, and it is repeated symbol-by-symbol.

(f) In the receiver, the incoming OFDM signal is frequency-converted to lower frequencies (appropriate to the hardware), digitised and applied to an FFT. Here the spectrum is analysed symbol-by-symbol, and the phases of the 1536 carriers are determined. This corresponds to OFDM decomposition.

(g) The high-resolution digital complex number which represents the phase of each carrier is divided by the value for the previous symbol in order to detect the differential QPSK.

(h) The resulting 1536 differentially decoded numbers are passed to an error-correction decoder where the redundant data and the ‘soft-decision’ detail are used to reconstruct the ‘un-coded’ bit-stream, symbol-by-symbol, as accurately as possible.

(i) The reconstructed bit-stream is de-multiplexed and the audio programme data are converted back to analogue signals, which are reproduced by a loudspeaker.

This is the ‘skeleton’ of the DAB system, but the ‘flesh’ contains several additional processes which make the system workable, and some which enhance its performance still further. The most important of these is source coding.

5. SOURCE CODING

The available gross bit-rate is about 2.3 Mbit/s and, within certain quanta, this can be apportioned to sound-programme data and error protection data as required. However, there is a trade-off between the ruggedness of mobile reception and the programme capacity. The optimum balance for terrestrial radio transmission may be approximately equal amounts of error protection and programme data, in which case the net capacity is around 1.2 Mbit/s.

However, the studio standard for digital audio signals, prescribed by the AES/EBU interface, uses 16-bit linear PCM with 48 kHz sampling rate, so a single full bandwidth (20 Hz to 20 kHz) stereo audio signal requires a bit-rate of some 1.5 Mbit/s. Compact Disc has a similar requirement. Therefore, it is essential that the bit-rate of the sound-programme data must first be reduced, and this is the function of a source encoder. A significant advantage in terms of spectral efficiency is gained when 5 or 6 stereo programmes can be transmitted within a single DAB signal.

The source encoder used in the DAB system can reduce the required bit-rate by a factor of 6, or more. It employs principles that were pioneered by IRT in its 'MASCAM' system[13], and then developed with CCETT and Philips to produce the 'MUSICAM' system[14]. A process based on MUSICAM has been adopted by the Moving Picture Experts Group (MPEG) of the International Standards Organisation (ISO) as a world-wide standard for audio source coding. This system is known as 'ISO/MPEG-1 Audio compression/de-compression' and is described in the ISO standard ISO 11172-3. The system has three different levels of complexity, referred to as 'Layers I, II and III'. An adapted version of the ISO Layer II system is used for DAB, although some proprietary source encoders for DAB are also labelled 'MUSICAM'. However, the adapted version is fully compatible with ISO Layer II decoders regarding the audio signal.

The result of ISO Layer II encoding is a bit-stream with a lower bit-rate from which, at least in principle, the original sound signal can be reconstructed in the receiver. The encoder can operate in stereo or mono mode and the output bit-rate is selectable between 384 kbit/s, for a stereo signal, down to 32 kbit/s for a mono signal[15], with a corresponding reduction in the quality of the re-constructed audio signal.

A value of 256 kbit/s has been judged to provide a high quality stereo broadcast signal [4]. However, a small reduction, to 224 kbit/s is often adequate, and in some cases it may be possible to accept a further reduction to 192 kbit/s, especially if redundancy in the stereo signal is exploited by a process of 'joint-stereo' encoding (i.e. some sounds appearing at the centre of the stereo image need not be sent twice). At 192 kbit/s, it is relatively easy to hear imperfections in critical audio material.

Multiple, cascaded encoding and decoding processes cause additional impairments, so 384 kbit/s is recommended for contribution links if the signal is to be re-coded before DAB transmission.

5.1. Masking and sub-band encoding

The bit-rate reduction is achieved by a combination of techniques which exploit observed properties of human hearing and redundancy in typical audio signals. The first stage is to suppress those components of the audio signal which would be inaudible, and this relies on a principle known as 'masking'.

---

[13] Masking-pattern Adapted Sub-band Coding And Multiplexing.

[14] Masking-pattern Universal Sub-band Integrated Coding and Multiplexing.

[15] The available options are: 192, 160, 128, 112, 96, 80, 64, 56, 48 or 32 kbit/s for a mono signal, or twice these values for a stereo signal.

The sensitivity of the ear to sounds at different frequencies is dominated by the loudest ones, so for example, if a strong audio component is present at 1 kHz, then the threshold of perceptibility for components at similar frequencies is raised, peaking at 1 kHz. A simultaneous component at 900 Hz or 1.1 kHz will only be heard if its amplitude exceeds that threshold. This happens because the response of the brain to the oscillating hair cells in the inner ear provides only limited frequency resolution (i.e. the hair cells have limited 'Q'). Thus, any components of the audio signal having amplitudes below this new 'masking threshold' will not be heard so there is no need to transmit data representing them. This is illustrated graphically in Fig. 9.

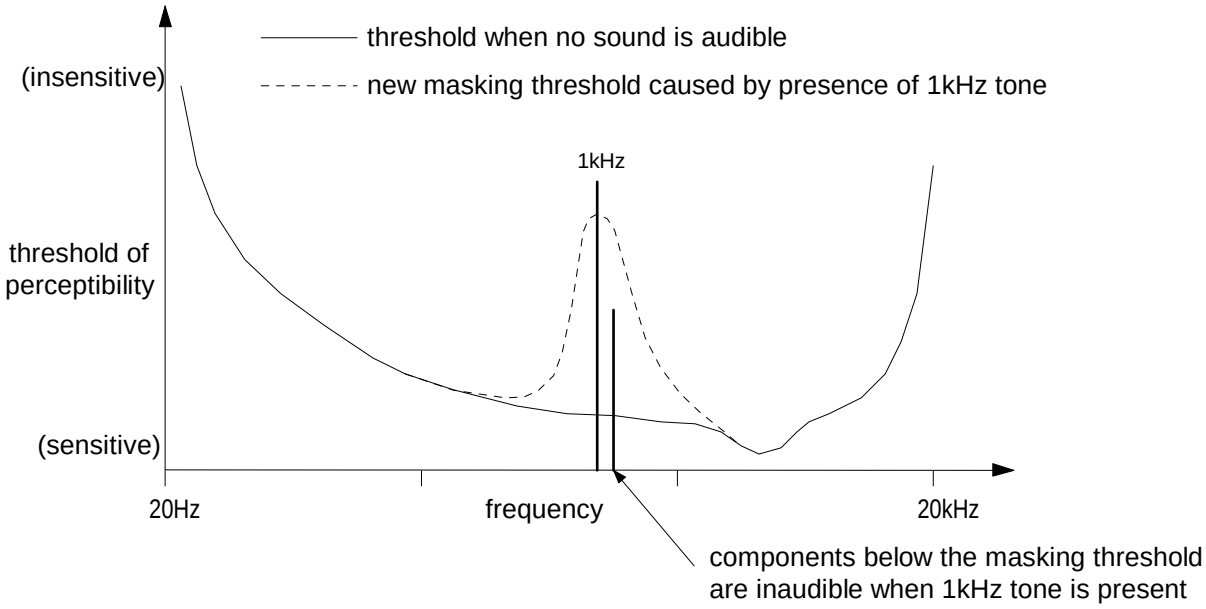


Fig. 9 - the masking principle

The masking principle actually allows bit-rate reduction by two methods:

(a) By *omitting* data which represent inaudible components.

(b) By *re-quantising* the data which are sent with a resolution (i.e. the number of bits; generally less than 16) just sufficient to ensure that quantising noise is effectively masked by other audible sounds.

For a typical audio signal, the masking threshold can exhibit many undulations within the audio band (20 Hz to 20 kHz), but it has been found that processing the signal in sub-bands, each of 750 Hz bandwidth, provides almost sufficient resolution. The incoming data are analysed by a digital 'filter bank' and the selection and re-quantising are carried out on data which represent the contents of 32 sub-bands. The highest-frequency sub-bands, above 20 kHz, are actually not used; their existence is a result of digital processing with a 48 kHz clock frequency. However, determination of the masking thresholds using the sub-band data would not be sufficiently detailed (*viz*. the lowest sub-band covers 5 octaves); instead, the incoming data are analysed using a 1024-sample FFT.

There can be interaction between the sub-bands, and it not always necessary to send data representing the principal contents of all 32; the contents of some can be completely masked by an adjacent strong component. The so-called 'psycho-acoustic model', which is used to determine the masking thresholds and to model the human response to transient sounds, is still being developed and improved.

5.2 Decoding

In the receiver, the ISO decoder applies a simple set of rules to reconstruct, frame by frame from these data, digital representations of the contents of the active sub-bands. These are combined, using a digital synthesis 'filter bank' (more easily thought of as a bank of oscillators), to produce the output 16-bit PCM bit-stream which is then passed to a DAC, the audio amplifiers and the loudspeakers. The decoding process is independent of the detailed structure of the psycho-acoustic model, and this could be revised in the future without the need for receiver modifications. However, it will not be possible to change from Layer II of the ISO standard to Layer III because there are substantial differences between the structures of the decoders. Layer II was chosen as a compromise between cost and performance, but Layer III can offer improved performance, especially at low bit-rates. Nevertheless, Layer II is also being developed by reducing the sampling frequency for low bit-rate applications, and this feature may be incorporated into the DAB system later. A Layer II decoder for a domestic receiver can be built in a single integrated circuit.

Of course, the encoding is a 'non-reversible' process, as some information is deliberately, and irrevocably, omitted. A single encoding/decoding operation at 256 kbit/s yields audio quality which can approach that of Compact Disc, at least in a single-ended test (i.e. not an A/B test), and this is acceptable to most listeners. However, problems can occur in cascaded encoding/decoding operations if the encoders do not perform identical operations, and further information is removed. The tolerance to cascading is improved at greater bit-rates when less information is removed in each encoding process.

5.3 ISO frames

In the encoder, the process of analysing the incoming audio data, determining the masking thresholds, selecting and re-quantising the output data, is carried out repetitively on blocks of data representing 24 ms periods of sound.

For each 24 ms period, a set of scale-factors is derived which represent the coarse amplitudes of the sound components represented in each active sub-band. The data representing the audio waveform for each active sub-band (i.e. the 'sample' data) are then numerically divided by the corresponding scale factor to produce smaller numbers. This provides some additional bit-rate reduction because the smaller numbers can be represented in a bit-stream by fewer active bits; the missing bits are contained in the scale-factors, but these are held constant for the 24 ms frame. This is akin to 'changing gear', relatively slowly, for louder or quieter sounds. The same Near Instantaneous Companding principle is used in NICAM.

A further stage of bit-rate reduction is then achieved by encoding the sample data differentially, that is, by sending data which represent the differences between successive samples.

The data output by the ISO Layer II source encoder are sent in frames of 24 ms duration. During each of these 'ISO frames', capacity is reserved in sequence for the following different categories of data:

(a) Header bits; in a known, unique pattern to facilitate frame synchronisation in the decoder and to indicate in which mode the encoder is operating (e.g. stereo/mono mode, and the output bit-rate).

(b) Bit-allocation data; indicating to which sub-bands the following two categories of data apply and, with reference to a look-up table, the quantisation that has been applied.

(c) Scale-factor data; for each active sub-band.

(d) Sub-band sample data; representing the re-quantised audio signal that was present in each active sub-band for the 24 ms period.

(e) Programme-Associated Data (PAD); a small amount of non-audio data for miscellaneous applications which require coincident timing, such as dynamic range compression, music/speech indication, etc. The effective bit-rate is variable, from a minimum of about 667 bit/s.

The chosen mode and output bit-rate determine the average number of bits per frame. Discrete stereo mode at 224 kbit/s, for example, produces approximately 5376 bits per frame, which is nearly twice the value for mono mode at 112 kbit/s (only one set of PAD are included in stereo mode); both cases provide the same audio quality per channel. Joint stereo mode at 224 kbit/s produces a similar number of bits but it can provide a small improvement in the audio quality, depending on the content of the audio signal.

5.4 Error protection

The different categories of data have different sensitivities to errors, and most categories were listed before, (a) to (d), in order of decreasing sensitivity. Elements of the PAD may have different sensitivities, and their use is optional, but these are normally taken to be highly sensitive. When FEC is applied to these data prior to transmission, the strength of the error correction capability (and, therefore, the amount of redundancy) is varied during each ISO frame to suit these different sensitivities, with the objective of providing a consistent degree of subjective ruggedness. This principle is known as static Unequal Error Protection (UEP) and several 'profiles' of different protection levels are available. The protection is reduced from the chosen maximum level in four distinct steps for categories (a) to (d), and is then restored to the maximum level for the PAD and category (a) of the following frame.

In addition to the UEP, an error-detection word is included immediately after the header to indicate errors in the most sensitive bits of the first three categories. A second error-detection word is provided specifically to indicate errors in the scale-factor data, and this is inserted immediately before the PAD in the previous ISO frame. Both of these words are afforded the maximum protection in the UEP profile. Such error-detection words have values derived from the data they protect, so their insertion before those data implies the use of buffering.

The scale-factor error-detection word is not included in the ISO 11172-3 standard, and its inclusion is part of the 'adaptation' referred to in Section 5.1.

5.5 Concealment

In adverse conditions of fading or interference, which exceed the correction capabilities of the FEC, the ISO bit-stream presented to the decoder can contain errors. Errors in the sub-band samples are relatively benign because their effects are limited to a narrow range of frequencies and the resulting sounds can mimic the intended audio material to some extent. On the other hand, scale-factor errors can give rise to spurious (and obvious) loud sounds in the decoded audio signal. The overall effect at the onset of errors is the appearance of 'grumbling' noises (the effect on a voice could be described as analogous to talking with a mouth full of marbles!). Errors in the bit-allocation data are relatively 'catastrophic' and can give rise to unintelligible noises.

The onset of scale-factor errors can be determined using the specific error-detection word, and this can be used to trigger a concealment strategy in the receiver. Two possible approaches to conceal errors in a single frame are to repeat the scale factors from a previous, error-free frame, or to mute the audio output of the decoder for that frame. Either approach can yield better subjective audio quality than using erroneous data.

If several consecutive frames are in error, the only realistic option is to mute the audio output. Herein lies one of the principal disadvantages of DAB relative to analogue systems like FM; the limited potential for 'graceful degradation' in the presence of severe errors. However, steps have been taken in the choice of the UEP profiles to make muting the very last resort. In a fully developed DAB transmitter network, the aim would be to provide uninterrupted coverage for vehicular receivers on the majority of roads, and some coverage inside buildings for portable and 'hi-fi' receivers; beyond this, for example in basements, it is inevitable that DAB receivers equipped with simple antennas will mute. Whilst the DAB network is being developed, it would be hoped that receivers would be able to switch to FM reception of the same programme if FM reception were adequate.

6. CHANNEL CODING AND MULTIPLEXING

The available gross bit-rate of about 2.3 Mbit/s can provide, for example, five stereo programme services (e.g. Radio 1, Radio 2, etc.) each at 224 kbit/s, leaving about 224 kbit/s for error protection of each service (rate 0.5 coding). Many other combinations are possible.

At the outset, the scope of this document was limited to the transmission of audio signals, but in this context 'service' can also mean a so-called 'general data' service, which may be data for the display of extended text (e.g. the contents of the 'Radio Times'). The partitioning of data into frames representing 24 ms periods of the application is retained but, generally, these are referred to as 'logical frames'. When the service provides an audio signal, a logical frame is equivalent to an ISO frame, adapted as described in Section 5.4. It is helpful to consider each logical frame as a burst of data, because when the data for numerous services are multiplexed together they must be compressed in time, so each logical frame is transmitted in less than 24 ms and other data are transmitted between these bursts.

In the DAB transmission chain, the bit-streams output by the numerous audio source encoders are first subjected to a number of processes, individually, before they are multiplexed together. The multiplexed bit-stream is then subjected to some further treatments before the application of OFDM and generation of the RF signal. The division of these processes, before and after multiplexing, is necessary to maintain flexibility in the DAB system; to allow bit-rates of individual audio channels to be changed independently. Several of these processes are known collectively as channel coding because they pre-condition the bit-stream, or the signal, in order to extract the best possible performance from the radio transmission channel.

6.1 Energy dispersal

Energy dispersal is used to break up strings of similar bit patterns to ensure an even distribution of power in the transmitted RF signal with respect to time and frequency (i.e. from one carrier to the next). The data from each source encoder are first applied to a scrambler, where a Pseudo-Random Binary Sequence (PRBS) is added modulo-2 to the bit-stream. Modulo-2 addition is the function carried out by an exclusive-OR gate. The same PRBS is available in the receiver and the sequence is timed to start afresh at the beginning of each logical frame, so the scrambling can easily be removed, again by modulo-2 addition.

The PRBS is generated by a 9-bit shift register with feedback from two of the taps combined by an exclusive-OR gate. The output of this gate is also taken as the output of the generator, and this is applied, bit-by-bit, to the bit-stream using a second exclusive-OR gate. At the start of the PRBS, all stages of the shift register are set to a value of 1.

Further scrambling can be applied at this stage for conditional access (e.g. to secure subscription radio services).

6.2 Convolutional encoding

The main FEC is applied to the scrambled data for each audio channel by a convolutional encoder. Different audio channels are encoded independently, so different amounts of redundant data can be added for different degrees of error protection. The average amount of redundancy applied to a channel (known as the ‘coding rate’) is selectable; a typical example is ‘rate 0.5’, where 50% of the transmitted bits convey unique data and the redundant data consume the remaining 50% of the bit-rate.

The instantaneous coding rate is varied during each logical frame for UEP according to a look-up table which is also available in the receiver. Information about which entries in the table should be used is signalled in the Multiplex Configuration Information (MCI), which will be described later (in Section 6.6). The MCI is transmitted with uniform rate 0.33 (i.e. powerful) coding. General data can also be transmitted with uniform coding.

Convolutional encoding is explained in Appendix 2, along with its counterpart in the DAB receiver, namely Viterbi decoding.

6.3 Time interleaving

The Viterbi decoder, which is the preferred means for applying the error correction in the receiver, offers outstanding performance in conditions when the transmission channel produces a random stream of errors. However, its performance can be impaired by bursts of errors lasting longer than a critical duration, and ultimately it will output erroneous data. This can occur when the receiver is mobile because of occasional flat fading or interference, for example. Also, the operation of the Viterbi decoder involves 'memory', so the effect of a serious burst of errors can be extended in time. Therefore, it is desirable to disperse any cluster of erroneous bits in the receiver before they are presented to the Viterbi decoder, and this is achieved by a process of time interleaving which requires action at both ends of the transmission chain.

After convolutional encoding, each logical frame contains a number of bits which depends on the source-coding mode and bit-rate, and the average convolutional coding rate. Apart from infrequent events when the system is re-configured, the average rates remain constant so consecutive frames contain the same number of bits. The bits in each logical frame are dispersed in time, or 'interleaved', by delaying their transmission by different amounts, and the same set of different delays is applied to the corresponding bits in each frame. 16 different magnitudes of delay are used and each is a multiple of 24 ms (i.e. the delays range from 0 to 15×24 ms), so a delayed bit is effectively transferred to a later frame but its position within the frame is maintained. This is a continuous process so there is always one frame of interleaved data ready for transmission.

In the receiver, the incoming bits are delayed by complementary amounts to restore the original sequence; this process can be called 'dis-interleaving'. The bit-stream is then passed to the Viterbi decoder and any burst of errors introduced between the interleaver and dis-interleaver is dispersed in time by the application of these complementary delays, improving the likelihood of effective error correction. This does not combat fading when the receiver is static, although it may help to reduce the impact of bursts of interference.

Time interleaving introduces a constant delay of at least 360 ms into the DAB transmission chain. Additional, smaller delays are incurred elsewhere in the chain in encoding, multiplexing and decoding processes, and there may be some other requirements for buffering (see Section 6.6). For the third-generation experimental equipment, the total delay has been measured as approximately 700 ms. Therefore, such procedures as off-air cueing at OBs will need to be adapted or eliminated when using DAB signals. If a satellite link is used for distribution to terrestrial transmitters, the total delay could approach 1 second.

Further details of the time interleaving process can be found in Appendix 3.

6.4 Multiplexing

The scrambled, coded and time-interleaved bit-streams for all of the different services are then brought together in a time-division multiplex known as the 'Main Service Channel' (MSC) which consumes most of the capacity of the DAB signal. Other ancillary channels carry data for synchronisation and other 'house-keeping' functions. The MSC has a fixed total capacity of 2.304 Mbit/s, which is related to the fundamental timing and modulation parameters and cannot be varied; padding bits are inserted to consume unused data capacity.

The MSC is organised in frames of 55296 consecutive bits, and these are known as 'Common Interleaved Frames' (CIFs) because they contain time-interleaved data from a number of different sources. Each CIF is divided into a number of time-slots in which logical frames of data for the individual services are transmitted, each as a burst of bits which represents a 24 ms period of the application (e.g. audio from one source). The repetitive bursts for each service provide what is known as a 'sub-channel'.

Each CIF is also broken down into 864 'Capacity Units' (CU), where one CU contains 64 bits and is the smallest addressable division of a CIF. All of the possible combinations of source-coding bit-rate and convolutional coding rate are arranged to yield whole numbers of CUs per CIF. For example, 224 kbit/s source-coding with rate 0.5 convolutional coding produces a 'gross' bit-rate of 448 kbit/s, so this is the required capacity of the sub-channel. This bit-rate corresponds to 10752 bits per 24 ms, so the sub-channel requires 168 CU per CIF. There is sufficient capacity for five audio signals coded in this way (840 CU), plus a remaining 24 CU (i.e. 1536 bits per 24 ms = 64 kbit/s) which can be used for other, perhaps non-audio, applications. In that case, a total of six sub-channels would be formed, but many other combinations are possible. A single CU only ever contains bits for one service.

The number of sub-channels, their capacities and the sequence in which they appear in each CIF, that is, the order in which the different sources are addressed, is known as the multiplex configuration. This is usually held static from one CIF to another but it can be changed at a frame boundary.

The sequence of bits representing each time-interleaved logical frame is kept together in the multiplexing process. This feature simplifies the design of receivers because the data for a selected service will always appear in the MSC at predictable times (constant times, if the multiplex configuration is not changed), so a domestic stereo receiver will only need to process these data, plus a small amount of administrative data, and not the whole capacity of the MSC.

In transmission Mode 1 (the different modes will be discussed later), the CIFs are concatenated in groups of four before transmission, and are compressed in time to allow the inclusion of synchronisation data and other vital information. The resulting transmission frames (sometimes called 'OFDM frames') are made exactly 96 ms long, of which the first 1.297 ms is reserved for a synchronisation 'null symbol'. The remaining 94.703 ms are divided into 76 symbols, each with a *total* duration of 1.246 ms; the meaning of this, and the departure from the 1 ms value, given before, will be explained shortly (in Section 6.9). The first four symbols are reserved for synchronisation and other data, and the remaining 72 symbols, 89.712 ms, are used for transmission of the four CIFs. The 221184 bits total requirement of the four CIFs is provided by 2-bits/symbol (QPSK) modulation of the 1536 carriers for 72 consecutive symbols (i.e. 72 symbol-blocks).

Note that one symbol-block accommodates 48 CU, so it may carry data for more than one service.

The data for each CIF are transmitted in 18 consecutive symbol-blocks, over a period of 22.428 ms, and the attendant time-compression is accomplished by buffering. For the earlier example of a 224 kbit/s sub-channel (448 kbit/s gross), the 10752 bits representing one logical frame are transmitted as a burst during 4 symbol-blocks (leaving 24 CU unused; available for another service), and therefore over a duration of 4.984 ms.

Fig. 10 illustrates the division in time of a single Mode 1 transmission frame and a single CIF; the passage of time is drawn from left to right across the page. In this example, the CIF is shown as filled with components from 6 sub-channels (corresponding to the previous example). In other cases, unused CUs would be filled with padding bits.

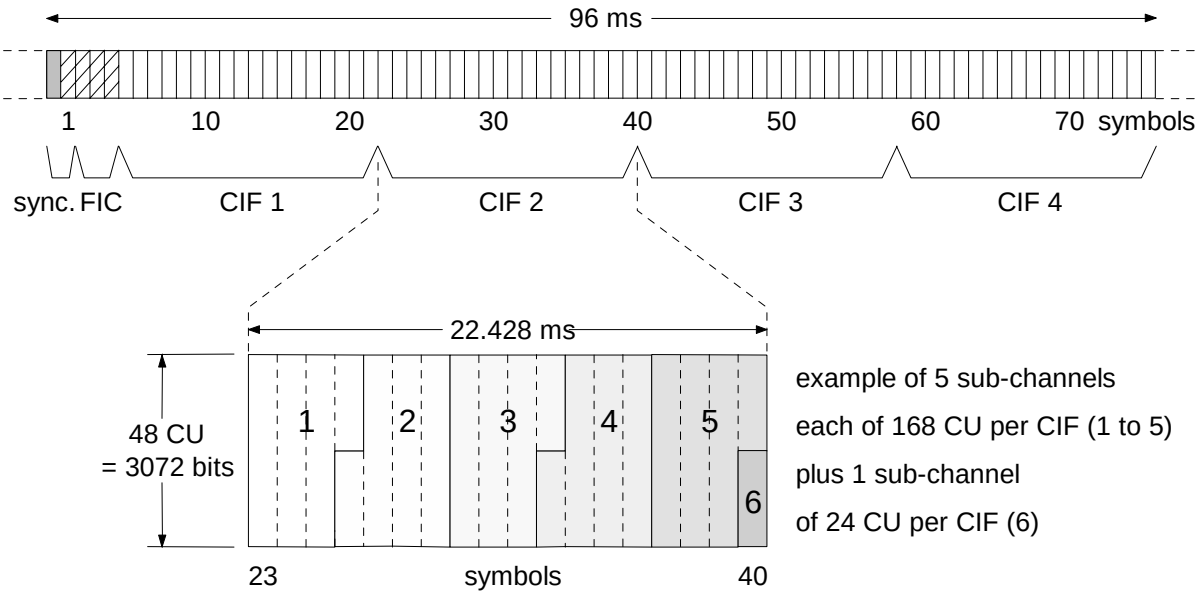


Fig. 10 - the DAB transmission frame and a Common Interleaved Frame

It should now be clear that although the 'COFDM' abbreviation contains the term 'Frequency-Division Multiplex', the multiplexing of digital audio signals in the DAB system is fundamentally with respect to time and not frequency. The term 'frequency division' relates to the way that the time-multiplexed data are distributed amongst multiple carriers.

6.5 Synchronisation channel

The first symbol-block in each transmission frame is reserved for the Phase Reference symbol. This and the 1.297 ms null symbol at the beginning of the frame make up what is known as the 'synchronisation channel', which provides facilities for AFC, time and phase synchronisation in DAB receivers. The operation of these features is rather complicated, so Appendix 4 is devoted to its description.

6.6 Fast information channel

The remaining 3 symbol-blocks at the beginning of the transmission frame are used to carry the Multiplex Configuration Information (MCI) and other vital data which are needed to set up the receiver circuitry before audio signals can be decoded. These data are scrambled and convolutionally encoded at a static uniform rate, but they are not time interleaved so they do not suffer the inherent delay. They provide what is known as the 'Fast Information Channel' (FIC), which is a practical necessity to make a receiver respond rapidly to the user when it is initially switched on.

The MCI provides the receiver with a succinct description of the contents of the four CIFs which follow it. The configuration can be changed dynamically to alter the division of the available capacity to individual programme services. For example, a selectable news programme could be *added* to the multiplex at hourly intervals, without disrupting continuing programmes (e.g. sports); this might require other programmes to relinquish some bit-rate. The MCI must effectively 'look ahead' to provide information about such changes in advance, so this requires buffering of the four CIFs. If this is not integrated with other processes which require buffering, an additional propagation delay of 96 ms will be introduced. In practice, the system used for signalling configuration changes is rather more complicated, and involves sending a 'count-down' signal which forewarns the receiver several transmission frames in advance. Several processes in the receiver require identification of individual logical frames and this is provided by running frame count (modulo 5000) which is transmitted in the MCI.

The absence of time interleaving makes the FIC data less rugged, so powerful error protection is applied. The data contain specific error-detection words, and rate 0.33 convolutional coding is applied corresponding to 67% redundant data. Also, most of these data are expected to change infrequently so the digital equivalent of a low-pass filter can be applied in the receiver.

The FIC can also carry data for non-audio applications such as 'Service Information' (SI, offering features like RDS), paging, traffic messages and conditional access. The capacity available for such applications is variable and depends on the complexity of the multiplex configuration in use, and hence the amount of detail which must be carried in the MCI. It is beyond the scope of this document to consider SI further, but some useful information is available on this topic and examples are listed in the Bibliography.

The total gross capacity of the 3 symbol-blocks is 9216 bits, but with rate 0.33 coding the total net capacity is 3072 bits. They occur once per 96 ms transmission frame, so the net bit-rate of the FIC is 32 kbit/s This capacity is actually divided up into 12 Fast Information Blocks (FIB), each of 256 bits, and 3 FIBs are associated with each of the CIFs carried in the MSC for that transmission frame. 16 bits of each FIB are devoted to an error-detection word.

6.7 Frequency interleaving

Apart from the Phase Reference symbol-block, 75 symbol-blocks of data, or 230400 bits, have now been identified which make up each transmission frame. In principle, these can be formed into a serial bit-stream to be applied to the 1536 carriers by D-QPSK and OFDM during 75 consecutive symbols, preceded by the null and Phase Reference symbols. By dividing the bit-stream into 75 blocks of 3072 consecutive bits, and associating pairs of bits within a block, 1536 bit-pairs can be made available during each symbol to be mapped onto the carriers.

If the mapping was a simple one-to-one relationship (e.g. the first two incoming bits were associated as a bit-pair, and this always modulated the lowest-frequency carrier, etc.), static selective fading or relatively narrow-band interference could impair one or more of the sub-channels selectively; they could suffer most of the errors. Even in the case of mobile reception, long strings of bits would be subject to relatively correlated fading events and this could give rise to bursts of errors. As noted before, this is the least favourable condition for successful error correction by a Viterbi decoder.

Instead, the mapping is based on a static pseudo-random series, and when the bit-pairs are recovered in the receiver, fading and interference events which are correlated amongst groups of adjacent carriers are dispersed within the bit-stream (i.e. they are dispersed in time). Subsequent bit-pairs are then affected by events which are uncorrelated in the frequency domain, and the dispersal of clusters of bit-errors improves the performance of the Viterbi decoder. This process is known as frequency interleaving, about which further details can be found in Appendix 3.

In practice, the bit-pairs are formed by partitioning the incoming bit-stream into blocks of 3072 consecutive bits and associating the 1st bit with the 1537th bit; the 2nd bit with the 1538th bit; and so on. This adds a further element of dispersal.

The data carried in the Phase Reference symbol are not subjected to frequency interleaving but are available as a further 1536 bit-pairs mapped to particular carriers (see Appendix 4).

6.8 Modulation and OFDM generation

The mapped bit-pairs are differentially encoded with respect to their counterparts for the previous symbol (as described in Section 3.4.1) and the 'difference' bit-pairs are, in principle, held in a register. This register is effectively the 'modulator', insofar as it contains an array of two-bit complex numbers which define the modulation states of the carriers, symbol by symbol. The way that these numbers are applied affects only the phases of the carriers in the transmitted signal; the amplitudes are all held constant and nominally equal.

This array is presented to an inverse FFT which synthesises the digital representation of a time-domain signal containing the 1536 carriers with appropriate phases; that is, the DAB signal. This is fed to a digital-to-analogue converter, producing a baseband analogue signal which can then be up-converted to the transmission frequency, amplified and transmitted in the normal way. The actual course of events is not quite as simple as this 'parallel' explanation, because the time-domain signal is produced as a succession of digital samples, 2048 per symbol, and this is explained in Appendix 1.

6.9 Addition of the guard interval

Even though relatively long symbols have been achieved by using OFDM, the degree of resistance to delayed echo signals (appearing in one symbol, but carrying the modulation of the previous symbol), would be limited. In the Eureka DAB system, this limitation is overcome most effectively by introducing a 'guard interval' between consecutive symbols.

The operation of an FFT demands that processing must be carried out in a clearly defined time-window. In this case, repetitive processes are carried out in time-windows with the same durations as the symbols. Hitherto, the symbol duration has been given as 1 ms, and the frequency separation of the carriers as the reciprocal, 1 kHz. In that case, the baseband time-domain signal contains sinusoidal components at 1 kHz, 2 kHz, etc., and these all contain whole numbers of cycles over the duration of one symbol, with phases dictated by the modulation data. It follows that if identical modulation data were presented to the inverse FFT for two consecutive symbols, all of these components and, indeed, the composite time-domain signal, would exhibit a seamless join at the boundary between the two symbols.

This is illustrated in Fig. 11; the composite ‘waveform’ will be discussed later (in Section 9).

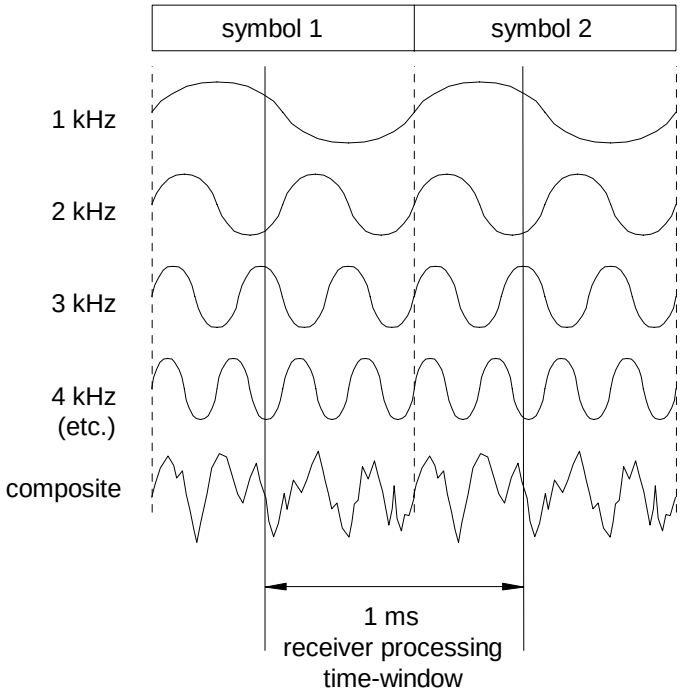


Fig. 11 - components of two identical consecutive symbols

Another FFT must be used in the receiver to separate the modulation from the multiple carriers, and this, also, must operate in time-windows having 1 ms duration.

Now if this two-symbol signal was presented to the FFT in the receiver, the exact point at which processing was started would not matter as long as the waveform was continuous throughout the 1 ms of processing. In other words, orthogonality would be preserved. Moving the starting point changes the apparent phases of components of the signal, but this would be overcome by the differential demodulation.

If all symbols were transmitted twice in this manner, this would allow the receiver a large tolerance in the synchronisation of its timing, and it would also provide immunity, rather than just resistance, to some delayed echoes, as will be explained shortly.

However, it would be very wasteful to transmit twice as many symbols as were used in the receiver, and the practical compromise is to transmit about one-and-a-quarter. In the transmitter, the ‘active’ symbol, over which original data are sent, is indeed 1 ms, but a 246 μs interval is allowed between consecutive symbols. In this 246 μs interval, a portion of the DAB signal ‘waveform’ is repeated by storing part of the sampled time-domain signal in a register, and reading it out for a second time. By choosing the repeated portion to be from the end of the active symbol, and by sending it immediately before the active symbol, a seamless joint is effected and the concatenated waveform still satisfies the requirements for orthogonality in any 1 ms within the 1.246 ms ‘total symbol’ duration. The period over which the signal is repeated is known as the ‘guard interval’. The alternative arrangement is equally applicable, of repeating a portion from the beginning at the end of the active symbol.

The point in each symbol at which the receiver processing time-window begins can be varied. By choosing this receiver 'timing reference' to coincide with the beginning of the wanted total symbol in the significant echo with the greatest delay, all of the contributions to the received signal, direct and echo, then carry the same, wanted, symbol. This is illustrated in Fig. 12 (with the same conditions as for Fig. 1 in Section 2.3); the guard intervals are denoted by the shaded portions of the symbols.

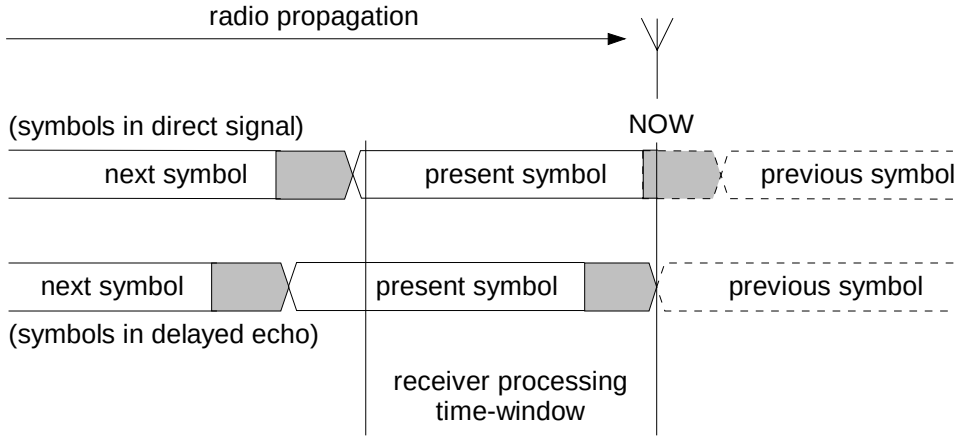


Fig. 12 - operation of the guard interval

The timing reference is derived from the Phase Reference symbol (see Appendix 4) by calculating the impulse response of the transmission channel. The reference is common to all carriers and is updated at the beginning of each transmission frame. The actual method of its derivation is a matter of receiver implementation, but the aim is to vary it adaptively to make the greatest constructive use of echo signals in a changing environment. In the current third-generation receivers it appears to be based on a simple calculation; the reference is adjusted so that the window starts at the end of the guard interval for the earliest arriving signal for which the magnitude exceeds a certain threshold. This threshold may be a fixed number of dB below the total signal power, which is sensed by AGC circuitry. In that case, the receiver can make constructive use of all echo signals with delays (relative to the earliest signal) less than the duration of the guard interval.

When consideration is given to signals at radio, rather than baseband frequencies, the initial flaw appears to be that the various echo contributions will arrive at the receiver with different phases, which could upset the QPSK demodulation process, or even cause cancellation. The first problem is overcome by the differential modulation; as long as the echoic environment is not changing very rapidly, it does not matter what the absolute phase of the resultant is, only the phase changes from symbol to symbol. The second problem can be realised, but this is no different from the normal effect of multipath propagation noted in Section 2.4, which the error correction and interleaving are designed to combat. Strictly speaking, it should be stated that the receiver 'can make constructive use of all *receivable* echo signals with delays less than the duration of the guard interval'; it cannot make use of severely attenuated carriers. It follows that such a guard interval can only be applied to a system which uses multiple carriers (e.g. OFDM) *and* powerful error correction together.

Those components of echo signals which are usable can add to the available signal power, improving the carrier-to-noise ratio and reducing the probability of transmission errors. In a typical mobile reception scenario, the various contributions, direct (if present) and echo, would be expected to exhibit some degree of differential fading (e.g. an echo is reinforced when the direct signal is attenuated). In this case, the ability to make use of whatever contributions are available, almost regardless of their delays, is a major advantage of the DAB system and this can provide an element of 'space diversity'. Because the receiver timing reference is common to all carriers, the guard interval is applied even when it is not needed; for example, when differential fading gives rise to only one significant contribution (e.g. direct or reflected) at a particular carrier frequency.

When components of the received signal are delayed by amounts greater than the guard interval duration, they cause inter-symbol interference but the degree of interference depends on the ratio of the delay to the symbol duration, as was noted in Section 2.3. With increasing delay, there is a gradual transition from a constructive to a destructive effect, and it is found that signals with additional delays of up to about 5% of the active symbol duration can still provide useful contributions to the total signal power. On this basis, the criterion for 'usefulness' is a relative delay of about 1.2 times the guard interval duration.

Occasionally, particular propagation conditions can give rise to 'pre-echoes'; that is, significant signal contributions arriving with *smaller* delays than the signal from which the timing reference has been calculated. This is illustrated in Fig. 13, where it can be seen that the guard interval of the following symbol appears in the receiver processing window.

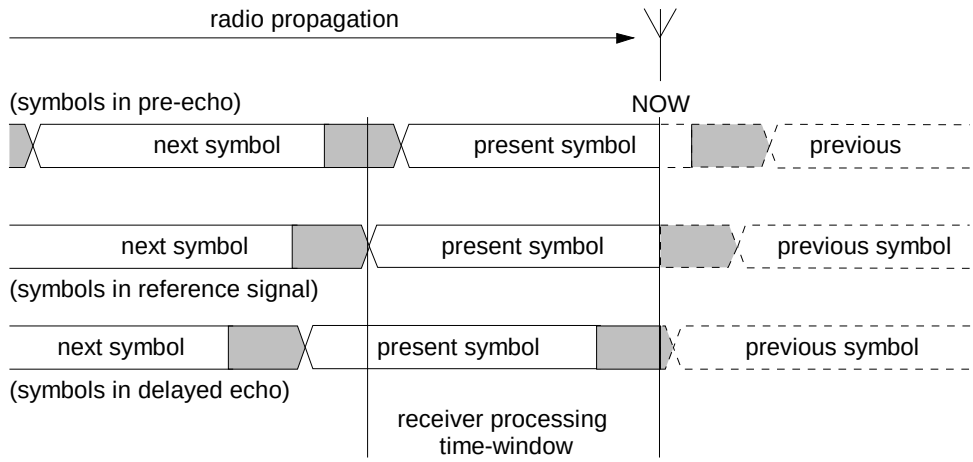


Fig. 13 - incidence of a pre-echo

This could occur at the instant when the direct signal path changes from being blocked (e.g. by a building) to un-blocked. The reference is updated once every transmission frame, so if such a pre-echo appears 'dynamically', shortly after a Phase Reference symbol has been received, appropriate action cannot be taken for up to 96 ms, in Mode 1 (and even then, the control loop may be damped). The effect of pre-echoes is the same as for echoes with delays greater than the guard interval. Ideally, the method for deriving the timing reference should prevent the incidence of static pre-echoes, but this is not always the case for the third-generation experimental receivers. It would be expected that improvements could be made in the future.

The introduction of 246 μs pauses does not affect the operation of the inverse FFT in the transmitter. The duration of the processing window is still 1 ms, so the separation of the carrier frequencies remains at 1 kHz, the reciprocal. However, the rate at which the phases of the carriers are changed (i.e. the modulation frequency) is reduced to the reciprocal of the *total* symbol duration; that is, about 803 Hz. This has an impact on the fine detail of the spectrum of the DAB radio signal, which will be discussed later (in Section 9). Similar comments apply to the FFT process in the receiver.

Of course, there is a price to pay for devoting about one fifth of the total symbol to what is, in effect, redundancy. The theoretical gross capacity is reduced from 3.072 Mbit/s, for the case of consecutive 1 ms symbols, to 2.466 Mbit/s, for the case of consecutive 1.246 ms symbols (ignoring the null symbol). Consequently, the spectral density is reduced from the theoretical maximum for QPSK, of 2 bit/s per Hz, to 1.604 bit/s per Hz. The actual capacity of the MSC is 2.304 Mbit/s, so if the time and bit-rate needed for the synchronisation channel and the FIC are considered as overheads, the effective spectral density is about 1.5 bit/s per Hz. Sometimes, error protection data are also considered as overheads.

It is arguable that some, or all, of the time given up for the guard intervals could, alternatively, have been made available for additional error protection. It is possible that this would have enhanced the performance of the DAB system in some conditions of short-delay multipath propagation, but at least one major advantage would have been lost; that is the potential for operating DAB transmitters in single-frequency networks.

7. SINGLE-FREQUENCY NETWORKS

Earlier, in Section 3, it was noted that the 1 ms symbol duration goes well beyond the requirement for tolerating multipath echoes. Indeed, 1 ms permits a guard interval as long as 246 μs whilst keeping the loss of capacity to manageable proportions, and this facilitates constructive use of delayed signals which have travelled up to 87 km further than the direct signal (applying the factor of 1.2 noted in Section 6.9). Essentially, there is no difference between a long-delay echo signal and an identical DAB signal radiated from a second transmitter on the same frequencies, so this feature makes possible the concept of a Single-Frequency Network (SFN) of DAB transmitters.

When signals from two transmitters can be received at the same place at similar signal strengths, the combination of the path length difference and any deviation from co-timing of the radiated symbols must result in a delay difference of less than 1.2×246 μs. Also, their radio frequencies need to be within about ± 10 Hz of one another. Any departure from these conditions will cause interference.

When more than two transmitters are used, normal propagation loss is relied upon to attenuate signals from distant transmitters which are delayed (relative to the signal from the closest transmitter) by more than ($1.2 \times$) the guard interval duration. There is some potential for problems in weather conditions which promote abnormal propagation (e.g. 'ducting'), but these conditions would be expected for only a small percentage of time. The effect of such 'network-generated' interference is minimised when planning a complete network by careful choice of the transmitter sites and ERPs.

It is conceivable that complete cancellation of the signal could occur at points on a line equidistant from two transmitters. This would be most likely if the only propagation paths were line-of-sight but, in practice, multipath propagation and blocking of one or other of the direct paths is quite likely, and these reduce the likelihood of complete cancellation. The breadth of the potential 'mush area' (i.e. the 'thickness' of the line) is small, about 195 m (*viz.* c / 1.537 MHz, where c is the velocity of propagation), beyond which the addition of the two signals will cause selective rather than flat fading. Where flat fading is possible, peaks and nulls will occur alternately at intervals across the line of a quarter of the wavelength of the RF signal. For example, in Band III the nulls would be separated by about 0.65 m, half the wavelength at 230 MHz. For a static receiver, the precise locations of such nulls would be expected to vary with changing weather conditions; that is, the effective path lengths would change owing to refraction in the atmosphere, and transmitter masts moving in the wind. In the case of a mobile receiver, temporal incoherence could be expected to randomise the locations and, perhaps, introduce an element of selective fading. Also, the time interleaving would reduce their impact above a certain vehicle speed.

The great advantage of the SFN principle is that the coverage of a national network is not related to the amount of radio spectrum available; DAB could provide 5 or 6 stereo services over the whole of the UK using just the one DAB channel. This is quite unlike FM, for which a national network carrying a *single* stereo service requires about 2.2 MHz of spectrum in the UK (3.3 MHz on the Continent) because adjacent transmitters must operate on different channels to avoid co-channel interference. In principle, the coverage of a DAB SFN is limited only by the cost of the transmitting stations; it can be extended outwards, or smaller and smaller gaps can be filled, simply by adding more stations to the network. It is even conceivable that the broadcast signal could be amplified and re-radiated in listeners' premises using domestic 'active deflectors', if they could be engineered to prevent self-oscillation!

Within a national SFN, there is no need to re-tune a mobile receiver whilst travelling, and in areas of overlapping transmitter coverage, the SFN provides an added benefit of increased spatial diversity. Of course, matters such as distribution of the DAB signal to large numbers of stations and synchronisation of their radiated signals are not trivial, but the potential advantages outweigh the problems.

8. TRANSMISSION MODES

So far, the discussion has been limited to Transmission Mode 1, but there are three possible modes in which parameters such as the number of carriers and the symbol duration are changed to adapt the DAB signal to applications other than terrestrial networks.

8.1 Why they are needed

Different applications, terrestrial and satellite, call for different radio frequencies, for reasons of spectrum availability and for practicality (e.g. the size of antennas on spacecraft). The frequency range considered extends from 30 MHz up to 3 GHz, but this is not possible using Mode 1 alone.

It has already been noted (in Section 3.4) that the requirement for coherence from symbol to symbol establishes a relationship between the maximum speed of a mobile receiver, the symbol duration, and the maximum radio frequency of the DAB signal. If it is desirable to maintain satisfactory reception at a maximum vehicle speed of at least 100 km/hr (62 m.p.h.), the choice of 1 ms symbol duration imposes a limitation on the maximum radio frequency. The principal cause of incoherence is the Doppler shift, and its effects will now be examined in greater detail.

When a receiver is mobile, the apparent frequency of the received radio signal is modified by the Doppler shift. The frequencies of all elements of a single signal are multiplied by amounts which are proportional to their transmitted frequencies and the vehicle speed. For example, at an effective vehicle speed of 100 km/hr. (i.e. towards or away from the transmitter), the Doppler shift expressed as a ratio is 0.093 parts-per-million, so all frequencies are multiplied by $1 \pm (0.093 \text{ ppm})$. The absolute magnitude of the effect (i.e. measured in Hz) is greatest when radio frequencies are considered, but the frequency separation of the carriers and even the symbol rate are all multiplied in the same ratio. In all practical cases, the modification of the carrier frequency separation is so small that it has negligible effect.

When a single (e.g. direct) DAB signal is received, the magnitude of the Doppler shift would be expected to change relatively slowly with changes in the vehicle speed or the angle of approach to the incoming radio wave. AFC in the receiver can compensate for slow changes of the radio frequency and the symbol rate, as described in Appendix 3. This scenario also applies to the case of static reception from a moving satellite over a single propagation path.

The potential problem arises when two or more contributions to the received signal arrive from different directions (e.g. echoes, or signals from different stations in an SFN). They may be subject to different Doppler shifts, perhaps even up *and* down in frequency, and these cannot be counteracted completely by agility in the receiver. The local-oscillator frequency is adjusted once per transmission frame according to the characteristics of the dominant received contribution, if there is one, or some aggregate of a group of contributions. It follows that some of the contributions can contain frequency errors.

Two of the features of the DAB system have limited tolerance to a frequency error introduced between the transmitter and the receiver:

(a) OFDM - for orthogonality to be maintained, the modulated carriers which are presented to the FFT in the receiver (following appropriate frequency down-conversion) must be centred on frequencies which are multiples of the reciprocal of the processing window duration (i.e. 1 kHz). If either the carrier frequencies are all shifted, or their frequency separation is altered (applying different shifts to different carriers), the result is crosstalk between the carriers, leading to erroneous data. The important point, noted in Section 3.1, is that waveforms which contain a whole number of cycles in 1 ms give zero results when integrated over 1 ms; a frequency error changes that number of cycles, perhaps fractionally.

(b) Differential phase modulation - a static frequency error is equivalent to a progressively increasing phase error; frequency is, by definition, the rate of change of phase. This corresponds to a static phase error from one symbol to the next, which modifies the differential phase modulation on each carrier. In the absence of noise and phase errors, after demodulation and differential decoding, the apparent carrier phases

should be mid-way between the adjacent pair of decision boundaries (at ±45°), giving the greatest margin against incorrect decision. This margin is reduced when the phase error is applied, impairing the tolerance to added noise and interference.

The second effect is dominant at low S/N ratios. For a given Doppler shift, the magnitudes of these effects can both be reduced by increasing the frequency separation of the carriers and by reducing the symbol duration.

The potential damage caused by these effects (and also RF interference) can be quantified with reference to the performance of the DAB system in the presence of Gaussian noise[16] in the transmission channel. The result of excessive noise or interference is the same: errors in the recovered bit-stream. In the absence of interference or fading, with rate 0.5 coding the third-generation experimental receiver requires a S/N of about 6 dB in order to output audio signals continuously, without muting. If interference is present as well as noise, the minimum S/N requirement is increased, and the magnitude of this increase (i.e. the impairment) is a guide to the amount of damage the interference is causing, by whatever mechanism. The impact of most types of interference depends on the FEC code rate.

A set of reference simulation results have been produced by the Eureka consortium, and they imply that the maximum radio frequency at which Mode 1 can be used is 375 MHz, consistent with a vehicle speed of approximately 100 km/hr. and causing 1 dB impairment of the S/N performance 'in the most critical multipath condition, occurring infrequently in practice'. It has not been possible to verify this by BBC measurements. At a radio frequency of 375 MHz, motion at 100 km/hr. would give a maximum Doppler shift of ± 35 Hz. In worst case conditions in an SFN, it is conceivable that two equal power signals could be received with positive and negative frequency shifts of 35 Hz each (i.e. separated by 70 Hz), but practical measurements have yielded impairments greater than 7 dB for this case. The reference figures must correspond to some 'less-catastrophic' scenario with a smaller Doppler spread or unequal powers.

8.2 Formulation of the three modes

The way to overcome the maximum frequency limitation is to formulate an alternative parameter set in which the carrier frequency separation is greater than 1 kHz and the symbol duration is less than 1 ms; in fact, a total of three sets have been formulated. However, some parameters need to be held constant in order to maintain the fundamental advantages of the DAB system and to simplify the design of receivers which should respond equally to all three modes. The signal bandwidth is held at 1.537 MHz, so the number of carriers is reduced, and a CIF always represents a 24 ms period of the audio signals which contribute to it.

Some other parameters are interdependent. In the additional 'transmission modes', the reciprocal relationship between the symbol duration and the carrier frequency separation is retained in order to maintain orthogonality and spectral efficiency. Shorter symbols impose greater demands for absolute timing accuracy, so the duration of a transmission frame is reduced to a smaller multiple of 24 ms in order to update the receiver's timing reference more frequently. Consequently, the numbers of symbols per frame are different. The duration of the guard interval is kept at a similar fraction of the total symbol duration.

[16] As is generated by receiver front-end amplifiers, and radiated by the earth, the sky, etc.

The features of all three modes are summarised below, in Table 1. All of the durations are whole multiples of 1/2048 ms so the table contains some approximations. The resulting maximum radio frequencies correspond to 1 dB impairment of the S/N performance at the point of failure for a vehicle speed of approximately 100 km/hr. (or 4 dB impairment at 200 km/hr.) 'in the most critical multipath condition, occurring infrequently in practice', according to the Eureka reference simulation results.

Parameter	**Mode 1**	**Mode 2**	**Mode 3**
number of carriers	1536	384	192
carrier frequency separation	1 kHz	4 kHz	8 kHz
maximum radio frequency	375 MHz	1.5 GHz	3 GHz
transmission frame duration	96 ms	24 ms	24 ms
number of symbols/frame	76	76	153
total symbol duration	1.246 ms	312 μs	156 μs
guard interval duration	246 μs	62 μs	31 μs
'active' symbol duration	1 ms	250 μs	125 μs
null symbol duration	1.296 ms	324 μs	168 μs

Table 1 - characteristics of the three transmission modes

Mode 1 - as described, is intended for terrestrial transmission, particularly using SFNs. The maximum frequency limitation is unlikely to be problematic because relatively line-of-sight propagation makes higher frequencies less suitable for large networks (*viz.* in view of the relatively large number of transmitters that would be required).

Mode 2 - is intended principally for terrestrial transmission using individual transmitters (i.e. local radio). The guard interval is sufficiently long to ensure immunity from multipath propagation, but is not really suitable for SFN applications (at least, not using omni-directional transmitting antennas). It has been suggested that this mode could also be used for hybrid satellite/terrestrial transmission in the L-Band (with emphasis on the 'terrestrial' aspect).

Mode 3 - is intended for cable delivery and satellite-and-complementary-terrestrial transmission. The relatively large carrier frequency separation reduces the demands on local oscillators for short-term frequency stability[17]. The short guard interval should be adequate for direct satellite reception, which would be expected to give rise to less multipath propagation.

For Modes 2 and 3, several other matters such as the allocation of bit-pairs to carriers, the frequency interleaving, and even the time interleaving are modified relative to Mode 1. These changes are consequences of the different numbers of symbols per transmission frame.

[17] Frequency variations which are too rapid to be counteracted by AFC, but which could otherwise impair symbol-rate coherence and cause errors. Expressed as phase-modulation noise in the baseband 30 Hz to 3 kHz, the critical phase deviation is approximately 0.03 radian RMS for the onset of errors using Mode 1. Proportionately greater amounts can be tolerated by Modes 2 and 3; approximately 0.12 and 0.24 radian, respectively.

In many respects, the relationships between parameters for Mode 2 and their counterparts for Mode 1 involve either multiplication or division by 4, and the relationships between Mode 3 and Mode 2 contain a further factor of 2. It seems odd that the Eureka reference simulation results appear to take no particular account of SFN operation in Mode 1. In view of what has been said here on this topic, it might be expected that the quoted maximum frequency would be somewhat smaller than a quarter of that for Mode 2.

It should be emphasised that these maximum frequencies are not precise, and it should not be inferred that the system will not work at greater frequencies or greater vehicle speeds, only that the impairment of the failure S/N ratio can be greater than 1 dB. Some potential applications for DAB call for considerably greater speeds (e.g. the French TGV rail system uses speeds in excess of 270 km/hr.) and in such cases the maximum frequencies for 1 dB impairment would be reduced further.

9. THE RF SIGNAL

9.1 Frequency domain characteristics

The long-term spectrum of a single QPSK signal (i.e. over many symbols with random selection amongst the four modulation states), has a power distribution following $(\sin f / f)^2$, where f is the separation from the centre-frequency with appropriate scaling. This has a peak at the centre-frequency and nulls at frequencies where f corresponds to a multiple of the symbol frequency, f_S. The half-power points occur at about 0.44 of the separation between the central peak and the first null, and the first sidelobes peak at about -13 dB. Subsequent sidelobes decay at 6 dB/octave. This is illustrated in Fig. 14.

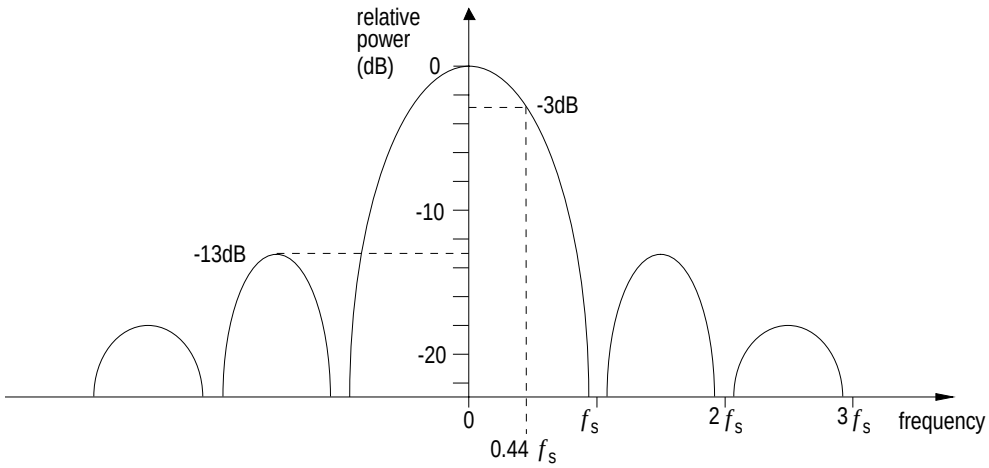


Fig. 14 - the spectrum of a single QPSK signal

Relative power is shown in Fig. 14 with decibel scaling, as is conventional for the display of a spectrum analyser, so the extremities of the nulls cannot be shown. Sometimes this spectrum is drawn showing the distribution of voltage with frequency, and conventionally the first (and other odd-order) sidelobes are shown having negative value (i.e. they are drawn hanging below the horizontal axis); the meaning of this is not intuitive when points along the horizontal axis correspond to different frequencies.

In the Mode 1 DAB ensemble, the centre-frequencies of the QPSK signals are separated by 1 kHz, but the effective symbol frequency is approximately 803 symbols per second; the reciprocal of the 1.246 ms total symbol duration. Thus, the peaks in the long-term spectrum of any one QPSK signal do not coincide with the nulls in its neighbours' spectra. This is contrary to the impression given by illustrations in some items of open literature, where the guard interval is neglected and the total symbol duration is taken as 1 ms.

However, this does not imply any departure from the conditions required for orthogonality. Insofar as it affects the operation of the FFT in the receiver, the 'short-term' spectrum of the signal, during any 1 ms processing window, consists of 1536 lines, each corresponding to an un-modulated sinusoidal wave. The changes (i.e. modulation events) occur in between processing windows, but during a single FFT process the signal is treated as though it were static. This is covered in slightly greater detail in Appendix 1. The short-term spectrum cannot be viewed using a conventional spectrum analyser, although it may be possible to display it using a specialised FFT analyser.

The overlapping spectra of six adjacent QPSK signals are illustrated below in Fig. 15; one as a plain line and five as dotted lines. This represents a small portion of the ensemble at the high-frequency edge; only three of the main lobes are shown.

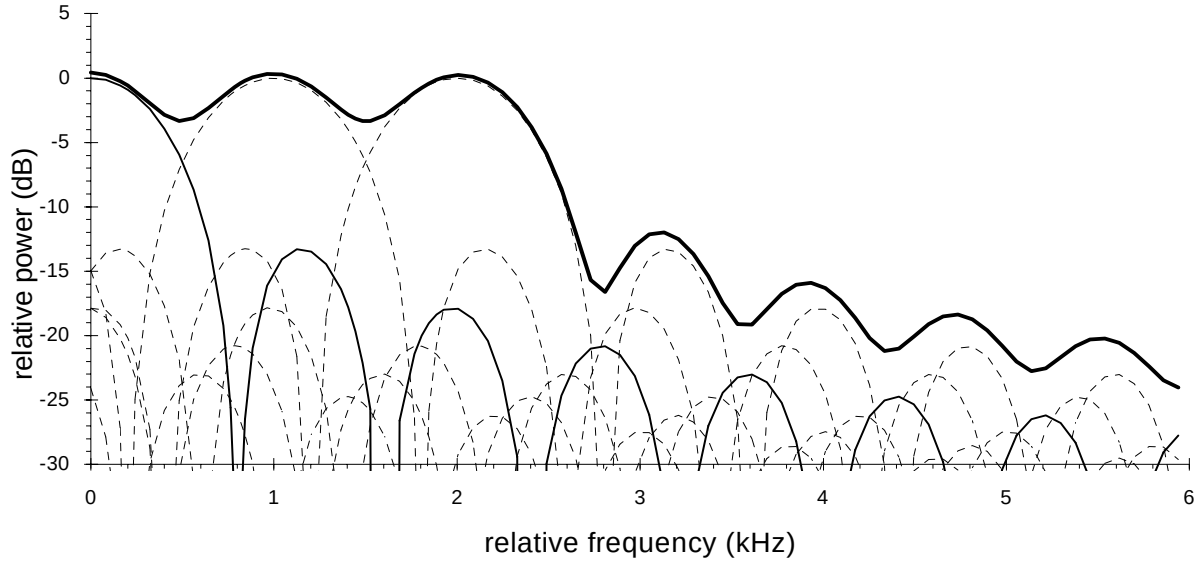


Fig. 15 - the spectrum at the high-frequency edge of the ensemble

Mid-way between two adjacent QPSK signals there are equal power contributions from the two main lobes, each at about 0.22 of the power of the central peak (i.e. at about -6.5 dB), assuming simple power addition. There are also small contributions from the first sidelobes of the two neighbouring signals, each at about -13 dB, and smaller contributions from other sidelobes of their neighbours. With the assumption of random modulation (i.e. that the various components are truly independent), the spectrum of the total DAB signal is given by the sum of the powers of the contributions, and this is indicated by the bold line in Fig. 15; the accuracy of this approximation would increase with time. Thus the spectrum of the ensemble is essentially flat-topped with a peak-to-peak ripple of about 3.8 dB. This can be observed using a spectrum analyser when the resolution bandwidth is set to less than 1 kHz; it usually helps to use display averaging to build up the approximation over many symbols.

At the edges of the ensemble, the overlapping $(\sin f / f)^2$ decays of the nearest QPSK spectra contribute to sidelobes which decay with increasing frequency separation from the ensemble. The first apparent sidelobes peak at about -12.5 dB with respect to the peaks of the ripple, or about -44 dB with respect to the total power in the ensemble (since a power ratio of 1536 corresponds to 31.8 dB). It is impossible to observe a single sidelobe *and* the total power at the same time using a conventional spectrum analyser. Between the half-power points, the total bandwidth of the DAB signal is approximately 1.537 MHz, and the relatively even distribution of power, in comparison with many types of single-carrier signal, reduces the potential for interference to other, smaller-bandwidth radio systems.

The ensemble actually contains the place for a 1537th carrier at its centre frequency, but this carrier is not deliberately generated if it is present; it is an artefact of the particular implementation of FFT processing that is used to generate the OFDM signal. This is explained more fully in Appendix 1.

9.2 Time domain characteristics

Viewed in the time domain, the DAB signal has characteristics similar to band-limited white noise; that is, over a long period, components at all frequencies in the signal bandwidth are represented so no clear waveform is discernible. This is illustrated by Fig. 16, where a single symbol is identified; the signal drawn here is a baseband DAB signal, prior to frequency conversion to RF. With truly random modulation of the different carriers, apart from the repetition during the guard interval, the signal voltage is essentially random within certain bounds.

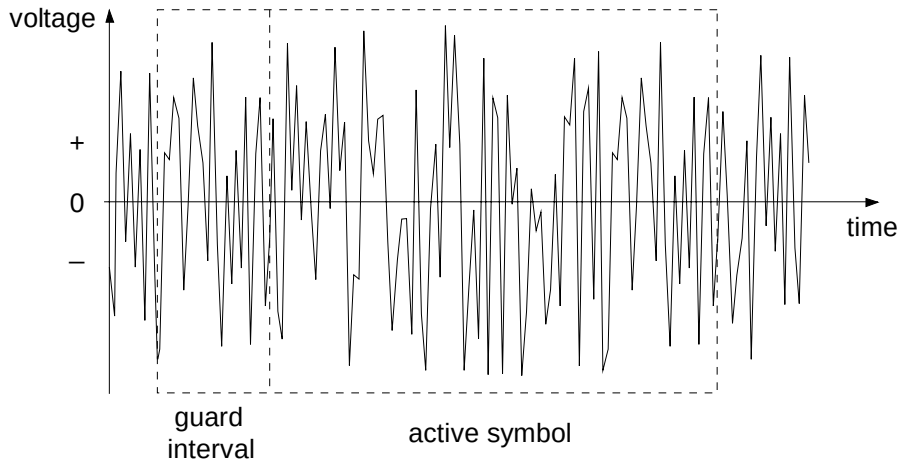


Fig. 16 - the DAB time-domain signal

If the signal was sampled many times, and the probability of encountering a particular voltage was plotted with variation of that voltage, the result would be a graph of the 'probability density function' of the signal voltage. This would have a shape similar to the bell-shaped Gaussian distribution for white noise, illustrated in Fig. 17. If the voltage gain of the system was kept constant, the distribution would converge towards a constant shape as the number of samples increased.

The signal voltage can have only one value at any instant in time (i.e. per sample). In the absence of a bias (e.g. if the signal is AC-coupled), the mean voltage is zero and the probability of encountering zero voltage is greatest; this is manifested by the abundance of zero-crossings in the signal 'waveform'. The probability of encountering positive and negative voltages diminishes as the magnitude of the voltage considered increases.

The instantaneous power which can be developed by such a signal into a load is proportional to the square of the voltage, and over a period of time there will be many contributions to the signal power arising from samples having different voltages. The average signal power is the average of these contributions (i.e. the total power divided by the number of samples considered). Because the shape of the distribution is constant, the probability of obtaining samples with any particular voltage is constant, so over a given (long) duration, the expected number of 1 Volt samples, for example, is constant. Therefore, the average power is constant even though the instantaneous voltage is changing continuously.

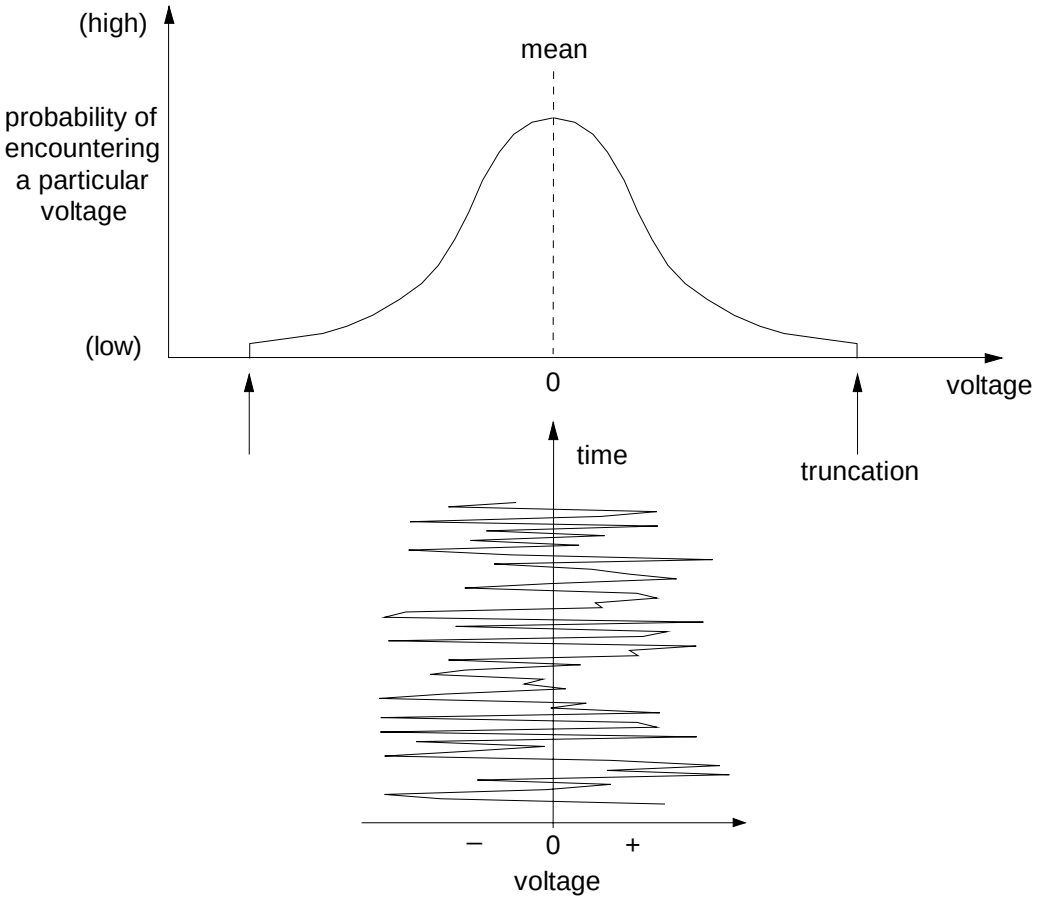


Fig. 17 - probability density function of the signal voltage

The average power of a DAB signal is also equal to the sum of the powers of the individual carriers and, since their amplitudes are nominally equal, this corresponds to a value about 32 dB (i.e. $10 \log_{10}$ of 1536) greater than the power of any single carrier. The affect of increasing the signal voltage (or power), by amplification, is to change the width of the probability density function whilst retaining the same bell shape, so greater voltages are expected with given probabilities.

An average-power meter (e.g. one which senses the heating of a load) can be used to measure the relative powers of DAB signals and to assess the absolute power of a DAB signal, provided that the voltage gain of the signal path to the meter is held constant. However, care should be exercised when using any power indicating device which employs a diode detector; its accuracy can depend on the time constants in the circuitry surrounding the diode.

Occasionally, large voltage peaks will occur. If each carrier is considered as a voltage vector rotating at its own frequency (different from all the others), then during some symbols the set of modulation phases can be such that a large number of the vectors momentarily fall into line. In principle, the maximum possible voltage would correspond to the addition of all 1536 carrier voltages, implying a peak voltage about 64 dB (i.e. 20 $\log_{10}$ of 1536) greater than the voltage contributed by a single carrier. It follows that the peak *instantaneous* power of the DAB signal could be 64 dB greater than the power of a single carrier, or 32 dB greater than the average power of the DAB signal (i.e. a 'peak-to-mean' ratio of 32 dB). In practice, the maximum voltage is limited by the generating hardware, which includes a DAC which has limited dynamic range, followed by analogue amplifiers. An example of this peak voltage limitation is shown in Fig. 17, where the distribution is 'truncated' rather than continuous over all voltages. For hypothetical Gaussian noise, the peak voltage tends to infinity but the probability of its occurrence tends to zero.

Digital generation of the DAB signal introduces quantisation and, whatever the resolution (i.e. the number of bits), this limits the dynamic range of the resulting signal at both extremes. Clipping and the addition of quantisation noise both cause some distortion of the signal. The effect of clipping can be interpreted as some loss of orthogonality, but the general effect of both types of distortion is to impose additional demands on the error correction process in the receiver. However, by appropriate choice of the average working points of the non-linear devices (i.e. by adjusting voltage gains in the signal path), the incidence of severe distortion in the generating equipment can be made infrequent. Nevertheless, the occurrence of occasional large-magnitude peaks does impose demands on the power amplifiers used to transmit DAB signals, and this is the topic of the next section. The actual peak-to-mean ratio or 'crest factor' can be determined by plotting the probability density function (e.g. using a counter with an adjustable voltage threshold), and this single number (10 dB, for example) is a useful guide to the requirements for power amplification.

In general, the use of a spectrum analyser for anything but *inspection* of the ensemble should be regarded with caution. Clearly, 'analysis' implies that only part of the signal is being displayed at one time, so any conclusions drawn about the total power should take into account the ratio of the analyser's resolution bandwidth to the 1.537 MHz bandwidth of the ensemble. The only way that such an instrument can portray the whole of the signal is to set its resolution bandwidth to greater than 1.537 MHz, whereupon the ensemble appears as a single peak. Even then, however, there may be some doubt as to how the instrument responds to such a signal with a large peak-to-mean ratio.

9.3 Power amplification

When a DAB signal is amplified, any non-linearity leads to the generation of Intermodulation Products (IPs). Distortion of the signal implies the generation of an 'error' signal; that is, the difference between the actual amplified signal and the desired un-distorted signal. This error signal manifests itself as the IPs, but they are not confined to the bandwidth of the DAB signal.

The power in the IPs depends on the transfer function of the amplifier and its operating point; improved linearity and increased back-off both reduce the IP power. Harmonics and other spurious signals may be generated in a DAB transmitter, but there is no fundamental reason why these cannot be removed by filtering. However, the IPs produced by the final power amplifier are intrinsic to the nature of the DAB signal.

Obtaining linearity by Class A operation is probably impractical for amplifiers producing 1 kW or more, in view of the very low electrical efficiency that is achieved (this can be less than 5%), and operation with large degrees of back-off (e.g. 10 to 20 dB) would require expensive amplifiers with large power ratings. There is scope for the application of linearisation techniques, as in the case of television transmitters, and it has been demonstrated that a simple pre-corrector can offer a significant improvement for an amplifier which is backed-off by 10 dB or more. However, practical pre-correctors for DAB amplification with minimal back-off are still being studied.

The occasional large-magnitude peaks in the DAB signal voltage can be clipped by amplifier saturation. The peak-to-mean ratio of the input signal can exceed 10 dB, so at least 10 dB output back-off would be needed to accommodate all of these peaks without distortion. In practice, occasional distortion of only the greatest magnitude peaks is found not lead to significant impairment of the signal, but this still requires some 6 dB back-off for a typical Class A/B amplifier.

IPs which are generated within the bandwidth of the ensemble cause interference to the DAB signal itself. They modify the instantaneous phases of the QPSK carriers (by phasor addition), and this reduces the integrity of the signal. However, if the mean total power of these IPs is kept at least 10 dB below the mean total power of the DAB carriers, in any given bandwidth ($\leq$ 1.537 MHz), serious impairment is avoided. This might require 3 to 6 dB output back-off for a typical Class A/B power amplifier.

IPs are generated at all possible beat frequencies between the carrier frequencies and their harmonics. Which combinations are significant depends on the transfer function of the amplifier, but a cubic component is usually predominant (especially for a push-pull amplifier) and this gives rise to so-called third-order IPs. In that case, the frequencies of the IPs are of the form $2 \times f_a - f_b$ or $f_a + f_b - f_c$, where f_a, f_b and f_c are carrier frequencies. These IPs all lie on the same regular 1 kHz comb as the carriers in the ensemble and they cover three times the bandwidth; that is, they stretch over $\pm$ 2.3055 MHz either side of the ensemble centre-frequency (or 1.537 MHz either side of the ensemble bandwidth).

IPs are not generated at nearby frequencies by an even-order component (e.g. square-law). Fifth, and greater, order IPs cover greater bandwidths, but their power is usually insignificant when the amplifier is operated away from saturation, as is the case for DAB at the moment.

IPs which fall outside the bandwidth of the ensemble, if radiated, could cause interference to other transmissions which occupy adjacent channels (e.g. DAB and other systems). Their power can be reduced by inserting a band-pass filter between the output port of the power amplifier and the antenna system. The required frequency response for this filter depends on the powers of the IPs generated and the maximum allowable out-of-band emissions. Such out-of-band third-order IPs appear on a spectrum analyser with decibel scaling (having small resolution bandwidth; say 10 kHz) as downward-curving skirts either side of the ensemble. This is illustrated in Fig. 18 on the next page.

In Mode 1, each skirt contains 1536 individual IPs having overlapping spectra, and each IP is made up from one or more components. The instantaneous power of an individual IP will be somewhat random, but the trend of the mean power, averaged over many symbols, will follow approximately the number of components which can be generated at that comb frequency. This reduces linearly with increasing separation from the edges of the ensemble, but with decibel scaling the spectrum appears curved.

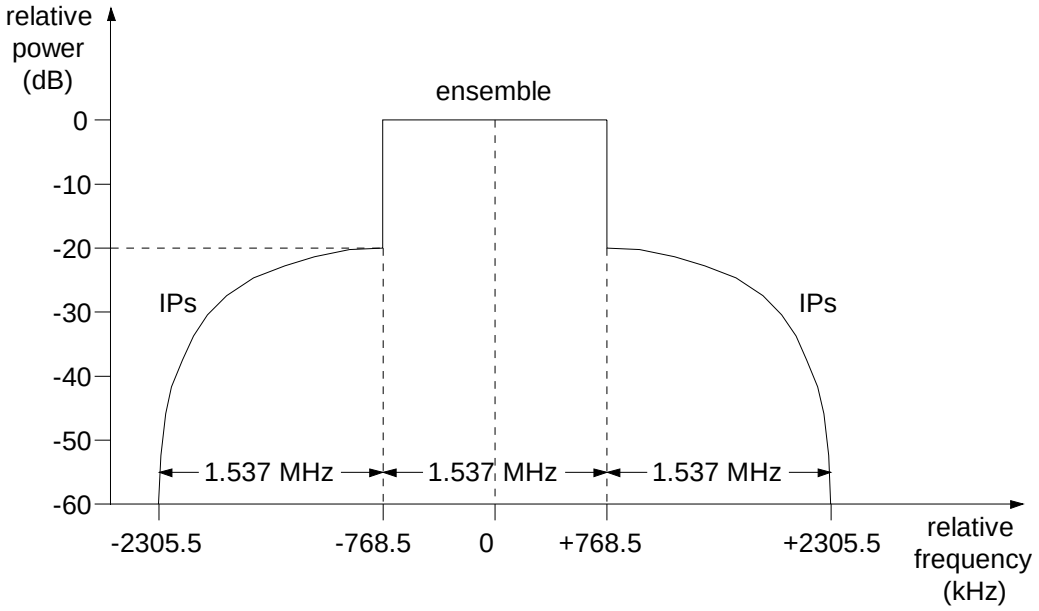


Fig. 18 - envelope of the spectrum of a DAB signal with third-order IPs

A simple way to quantify the levels of these out-of-band IPs is to compare the mean power of a particular one with the mean power of one of the wanted DAB carriers. The favoured method is to consider the IP separated by about 200 kHz from the edge of the ensemble. The initial slope of the skirt has little effect on the accuracy of this approach (<1 dB), and this avoids the influence of the decaying $(\sin f / f)^2$ spectra close in to the ensemble. The result is commonly given as the 'relative IP level' which, for a typical amplifier operated with 6 dB output back-off, would be some -20 to -30 dB; it is shown as -20 dB in Fig. 18. In practice, a spectrum analyser is used for this measurement so the result is actually the relative power of a group of IPs and a group of carriers, in the same bandwidth (e.g. 10 kHz), and this is sufficiently accurate for most purposes.

It should be noted that Fig. 18 is a drawing of the idealised spectrum for a resolution bandwidth of several kHz. In practice, the signal-to-noise ratio at the output of a DAB transmitter is limited by the hardware, and the dynamic range which can easily be displayed is often less than 50 dB.

The general conclusion is that until suitable linearisation techniques have been developed, power amplifiers for DAB signals will need saturated output power ratings at least 6 dB greater than the required DAB output power.

10. CONCLUSIONS

This document has given a general introduction to the purposes, benefits and operation of the Eureka 147 DAB system. Beyond this text, much additional information can be found in the documents listed in the Bibliography, and when the Eureka 147 guidelines document becomes available, that will become the authoritative reference.

It is hoped that this document will leave readers with the impression that although the DAB system is apparently very complex, this complexity is manageable, and all necessary for the system to achieve its demonstrable outstanding performance. Those that have attended one of the several public DAB demonstrations, given by Research Department and Engineering Information Department, will probably recall just how impressive this performance is.

It is inevitable that the future of broadcasting lies in the domain of digital techniques, and it is most likely that the future of BBC national-network radio services lies in the use of the Eureka DAB system.

11. ACKNOWLEDGEMENTS

Thanks must be recorded to the many colleagues at Research and Development Department who have assisted the author with helpful discussions about the techniques involved in the DAB system, particularly: M. C. D. Maddocks, J. H. Stott, C. R. Nokes, A. P. Robinson, H. Lau and P. Shelswell. The assistance of representatives of partners in the Eureka 147 consortium is also acknowledged, particularly A. Müller of Daimler Benz. J. P. Chambers, formerly of BBC Research Department, was the original source of Fig. A3.1 (in Appendix 3) illustrating the time interleaving process.

C. Gandy, BBC Research and Development Department, 29th September 1994.

12. REFERENCES

[1] ETSI. Final draft prETS 300 401, Radio broadcast systems;
Digital Audio Broadcasting (DAB) to mobile, portable and fixed receivers.
Sophia-Antiplois, September 1994.

[2] EBU. 1988. Advanced digital techniques for UHF satellite sound broadcasting; collected papers on concepts for sound broadcasting into the 21st century.
EBU Technical Centre, August 1988. pp. 32 - 34.

[3] CCIR. Report 955-2: Satellite Sound Broadcasting with portable receivers and receivers in automobiles. CCIR, 1990.

[4] ITU-R. Draft Recommendation BS 1115 (formerly Draft Recommendation 10/52): Low bit-rate audio coding. Input document to ITU-R Study Group 10 meeting, Geneva, February/March 1994.

13. BIBLIOGRAPHY

The following texts are non-confidential and provide much useful background information, although extensive use is made of mathematical notation in some cases. The most recent are given first:

1. RATLIFF, P. A. 1994. Eureka 147 Digital Audio Broadcasting - the system for mobile, portable and fixed receivers. Proc. of Second International Symposium on DAB, March 1994.

2. RILEY, J. L. 1994. DAB: Multiplex and system support features. BBC Research and Development Department Report No. BBC RD 1994/9.

3. BELL, C. P. and WILLIAMS, W. F. 1993. Coverage aspects of a single frequency network designed for digital audio broadcasting. BBC Research Department Report No. BBC RD 1993/3.

4. MADDOCKS, M. C. D. and PULLEN, I. R. 1993. Digital audio broadcasting: Comparison of coverage at different frequencies and with different bandwidths. BBC Research Department Report No. BBC RD 1993/11.

5. MADDOCKS, M. C. D. 1993. An introduction to digital modulation and OFDM techniques. BBC Research Department Report No. BBC RD 1993/10.

6. International Standard ISO/IEC 11172-3. Coding of moving pictures and associated audio for digital storage media at up to 1.5 Mbit/s. March 1993. Audio part, Layer II.

7. STOLL, G. 1992. Source coding for DAB and the evaluation of its performance: a major application of the new ISO coding standard. Proc. of First International Symposium on DAB, June 1992. pp. 83 - 98.

8. CHAMBERS, J. P. 1992. DAB system multiplex organisation. Proc. First International Symposium on DAB, June 1992. pp. 111 - 120.

9. Le FLOCH, B. 1992. Channel Coding and Modulation for DAB. Proc. EBU First International Symposium on DAB, June 1992. pp. 99 - 110.

10. PRICE, H. M. 1992. CD by radio. IEE Review, 38, 4, April 1992. pp. 131 - 135.

11. SHELSWELL, P., BELL, C. P., *et al*. 1991. Digital Audio Broadcasting: The first UK field trial. BBC Research Department Report No. BBC RD 1991/2.

12. GILCHRIST, N. H. C. 1990. Digital Sound: Subjective tests on low bit-rate codecs. BBC Research Department Report No. BBC RD 1990/16.

13. BELL, C. P. and STOTT, J. H. 1990. UK developments in digital audio broadcasting. Proc. of International Broadcasting Convention, 1990.

14. POMMIER, D., RATLIFF, P. A. and MEIER-ENGELEN, E. 1990.
The convergence of satellite and terrestrial system approaches to digital audio broadcasting with mobile and portable receivers.
EBU Review Tech., 241/242, June/August 1990. pp. 82 - 94.

15. Le FLOCH, B., HABART-LASALLE, R. and CASTELAIN, D. 1989.
Digital sound broadcasting to mobile receivers.
IEEE Trans. on consumer electronics, 35, 3, August 1989. pp 493 - 503.

16. POMMIER, D. and RATLIFF, P. A., New prospects for high-quality digital sound broadcasting to mobile, portable and fixed receivers. Proc. of International Broadcasting Convention, 1988. IEE conference publication No. 293, pp. 349 - 352.

17. EBU. 1988. Advanced digital techniques for UHF satellite sound broadcasting; collected papers on concepts for sound broadcasting into the 21st century.
EBU Technical Centre, August 1988.

Summary of contents:

- Introduction - Purpose of the demonstrations of experimental UHF digital sound broadcasting.
- EBU guiding principles: Satellite sound broadcasting in the frequency range 0.5 to 2 GHz.
- EBU technical studies on an advanced digital system for satellite sound broadcasting in the frequency range 0.5 to 2 GHz.
- New prospects for high-quality digital satellite sound broadcasting to mobile, portable and fixed receivers.
- Interleaving or spectrum-spreading in digital radio intended for vehicles.
- Principles of modulation and channel coding for digital broadcasting for mobile receivers.
- Low bit-rate coding of high-quality audio signals - An introduction to the MASCAM system.
- Real time software processing approach for digital sound broadcasting.

18. STOTT, J. H. 1985. Satellite sound broadcasting to fixed, portable and mobile receivers. BBC Research Department Report No. BBC RD 1985/19.

19. CCIR Recommendation 774: Digital Sound Broadcasting to vehicular, portable and fixed receivers using terrestrial transmitters in the VHF/UHF bands.

20. CCIR Recommendation 789: Digital Sound Broadcasting to vehicular, portable and fixed receivers for BSS (sound) in the frequency range 500-3000 MHz.

21. CCIR Report 1203: Digital Sound Broadcasting to mobile, portable and fixed receivers using terrestrial transmitters.

APPENDIX 1

OPERATION OF AN FFT

The implementation of the DAB system is made possible by virtue of the Fast Fourier Transform (FFT), and the following overview of the way in which the FFT works may assist in understanding the relevant stages of the signal path. An important point to keep in mind is that the processing in the transmitter and receiver needs to be carried out at great speed to support the kinds of bit rate involved.

A1.1 Introduction

In developing DAB to combat multipath propagation, attention has been paid to the effects of radio propagation in both the time domain and the frequency domain. As noted in the main text, these two domains provide different viewpoints for the same effects. Equally, when generating or receiving a signal, certain aspects of the processing can be carried out more easily in one or other of the domains. This is particularly true in the case of DAB, where the multiple-carrier RF signal is more easily synthesised and analysed in the frequency domain, but the symbol-by-symbol modulation is more easily treated in the time domain. Indeed, if a DAB signal is displayed on an oscilloscope, it appears similar to band-limited white noise punctuated by the null symbols (every 96 ms in Mode 1), which gives little clue to the existence of multiple discrete carriers.

Successive stages of a DAB transmitter, or receiver, operate on constituents of the DAB signal in both of these domains. In the channel encoder, the spectrum of the signal is constructed essentially as an array of numbers, each representing the instantaneous amplitude and phase of one of the QPSK carriers. From this frequency-domain spectrum, the equivalent time-domain signal[18] is produced, which can be up-converted to the final frequency and transmitted. The changes of modulation states from symbol to symbol are effected by changing the numbers input to the array. In the receiver, from the incoming time-domain signal, the spectrum is re-constructed as an array of numbers representing the individual modulated carriers, from which their modulation states can be determined.

The link between the time and frequency domains is process of transformation. Bearing in mind the number of carriers, 1536 in Mode 1, it would be out of the question to perform this process in a domestic receiver using analogue circuitry (e.g. banks of oscillators and filters).

The solution is to implement the transformation digitally, and there are several possible approaches, of which one of the most rapid is the FFT algorithm. The treatment in the main text showed, in practical terms, how the discrete Fourier transform (DFT) can be developed from a block diagram of the OFDM decomposition process. In this appendix, the DFT and then the FFT will be derived in stages starting from the fundamental Fourier transform.

---

[18] Any signal can be considered from the viewpoint of the time domain; the term ‘time-domain signal’ is used only to signify that, in this case, consideration is being given specifically to the variation of the signal voltage with the passage of time, and not from some other viewpoint.

A1.2 The Fourier transform

The basic principle behind transformation is that any arbitrary waveform can be synthesised by adding together a collection of continuous sinusoidal waves of different frequencies, having appropriate amplitudes and phases; or that the waveform can be broken down into its constituent sinusoidal components. A well known example is the continuous square wave, which can be decomposed into a fundamental sine-wave and a comb of odd harmonics with progressively decreasing amplitudes with increasing frequency. If the amplitude/frequency distribution of these sine-waves is plotted, this gives the spectrum as might be displayed (approximately) on a spectrum analyser. A general result of this process is that waveforms which change slowly have spectra which contain significant power only at low frequencies, and more-rapidly changing waveforms have spectra with greater bandwidths.

The Fourier transform is a mathematical process which identifies the frequencies, amplitudes and phases of the spectral components by the solution of an integral. The reverse process, constructing a time-domain waveform from the description of its spectrum, is known as the inverse Fourier transform, which uses a remarkably similar integral where frequency and time are interchanged.

In its fundamental form, the Fourier transform is continuous: it operates on a waveform that can be described continuously for all time; its solution describes the spectrum continuously over all frequencies; and both of these descriptions are in terms of analogue complex numbers[19], having infinitesimal resolution. In most practical applications, the subject waveform is treated as if it were continuous for all time, even though it cannot be. When transform techniques are applied to practical digital systems, naturally, some compromises have to be made.

A1.3 Digital implementation

In order to implement the Fourier transform digitally, the first step is to represent the quasi-continuous input signal as series of discrete samples represented by digital numbers with finite resolution. The integral now becomes a much simpler summation, of these numbers multiplied by fixed coefficients.

The next step is to impose time limits in order to limit the extent of the summation. In practice, the number of input samples which are available to be transformed may already be defined (e.g. by the symbol duration, in the case of DAB), and this automatically imposes time limits. It is convenient to think of this as the application of a ‘time window’; processing is only carried out on those samples which appear when the window is ‘open’. This action also imposes a limit on the range of frequency values which need to be considered in the summation, which will be explained later.

[19] When a sinusoidal wave is represented by a complex number, its amplitude is represented by the square-root of the sum of the squares of the imaginary and real parts, and the tangent of its phase is represented by the ratio of the imaginary and real parts.

The resolution with which time is treated is now no longer infinitesimal, and errors can be introduced if all significant components of the input signal are not faithfully represented by those samples taken. As is often the case in sampled systems, a compromise has to be made between accuracy and an acceptable amount of processing. Generally, the sampling frequency must be greater than twice that of the highest frequency component in the input signal, and this, so-called, Nyquist criterion is applicable in most cases of time-domain sampling. Frequency components above half the sampling frequency are not represented accurately and may need to be removed from the input signal by filtering. A frequency component at exactly half the sampling frequency can be considered to be at the Nyquist limit.

The result of the summation contains the required Fourier transform along with other 'distortion' products. Much of this distortion can be removed by (re-)sampling the result at an appropriate rate in the frequency domain.

A1.4 The discrete Fourier transform

The end product of these modifications is the DFT. When the extent of the summation is pre-determined, the values of the coefficients are known. They can either be calculated when required using an algorithm, such as a series expansion, or pre-calculated and stored as a look-up table if memory size permits. Then, the whole process can be carried out by a computer as a sequence of relatively simple multiply-and-add operations. Within some limitations, this can provide a good approximation to the fundamental Fourier transform.

The result of the DFT provides a series of samples of the spectrum, for negative and positive frequencies. The time-domain sampling of the input signal gives rise to repetitions of this two-sided spectrum at higher frequencies, centred on multiples of the sampling frequency (i.e. the spectrum is periodic in terms of frequency). If the sampling frequency is sufficiently great, these can be removed by filtering. Insufficient sampling frequency gives rise to overlapping spectra, so-called 'aliasing', which cannot easily be removed. A spectral component at the Nyquist limit must, by definition, contain an unwanted alias.

If the Nyquist criterion is just satisfied in the time-domain sampling, then the resulting sampled spectrum cannot contain useful information at frequencies above half the time-sampling frequency. If the time-sampling frequency is f_s and the time-window duration is T, then the number of samples processed $N = T \,.\, f_s$. The interval between the frequency-domain samples is $1/T$ and the useful range lies between $\pm f_s/2$, so the number of useful samples is $\pm (f_s/2)/(1/T) = \pm N/2$; that is, $N/2$ samples at positive frequencies and $N/2$ at negative frequencies. Thus, the total number of useful samples in the result is equal to the number of samples input. With N time-domain and N frequency-domain samples, a total of N^2 coefficients are needed in the summation.

The frequency-domain sampling of the result has a similar, although reciprocal, effect in the time domain; that is, the result of the DFT applies to the time-windowed input signal as if it were periodic with a period equal to the time-window duration. If the input signal really is periodic (i.e. it is composed of the same 'waveform' during consecutive and contiguous time windows), the results of consecutive DFT calculations will be the same. If it is not, subsequent DFT calculations will yield different results.

A1.5 Computation of a DFT

If a computer program was set up to implement a 16-sample DFT, it would take as its input 16 complex numbers representing consecutive samples of the time-domain signal to be transformed. For each sample in turn, the program would perform the complex multiplication of that sample and the appropriate coefficient for the first output frequency; the results would be added together and stored. This would then be repeated for the remaining 15 output frequencies, giving the result: 16 stored complex numbers representing samples of the spectrum at different frequencies. It is important to note that each output sample has contributions from every one of the input samples.

This would require 16^2 (i.e. 256) multiplication operations and 16 additions. Multiplications are more time-consuming operations for a computer, and generally the relationship between the number of multiplications and the number of input or output samples is a square law. This can lead to excessive computing time for large numbers of samples, which is a fundamental shortcoming of the DFT when implemented on a computer. However, there is a significant amount of redundancy in this 'long-hand' computation, and algorithms have been developed to exploit this. Notwithstanding this, in some cases, multiplication speed is less of a problem than other processes, such as memory access, and specialised integrated circuits are available which are designed to implement DFTs.

It was noted earlier that the inverse Fourier transform uses an integral which is very similar to that of the (forward) Fourier transform, so it follows that the inverse DFT uses a summation which is very similar to that of the (forward) DFT. Also, if the input frequency-domain array, remains static for consecutive inverse DFT calculations, the time-domain result will be periodic over consecutive time windows. If the window duration is equal to, or an integer multiple of, the period of this result, then the result will be a sampled waveform free from discontinuities (i.e. glitches). For example, with a 1 ms window, sinusoidal waves at 1 kHz and harmonics (within the Nyquist limit) can be portrayed without discontinuities.

A1.6 The fast Fourier transform

The FFT is a particularly efficient algorithm for implementing the DFT. It increases the speed of processing by cutting down the number of multiplications from n^2 to $(n/2)\log_2(n)$, where n is the number of input or output samples, in cases where n is a power of 2. Thus, representing a 1536-carrier DAB signal by means of 2048 samples, an FFT would require 11264 multiplications, whereas a DFT would require more than 4 million. The number $\log_2(n)$ has a value of 11 for DAB, and can be called the index of the FFT (i.e. $2^{11} = 2048$).

The FFT can be derived from the DFT by expressing the summation using matrix arithmetic. All of the computations relating the values of the output samples to the input samples can be expressed in a two-dimensional matrix. Individual elements of this matrix can be broken down into consecutive stages of simpler arithmetic; that is, they can be factorised, just as $x^2 + 3x + 2$ can be factorised into $(x + 1)(x + 2)$. The matrix, as a whole, can be factorised into a number of matrices containing simpler expressions, and when this process is taken as far as possible, the number of factored matrices is equal to the index.

This factorisation process, sometimes referred to as 'decimation', introduces several simplifications; some expressions always return zeros or ones, so they need not be calculated, and some others have counterparts in the same matrix which yield the same result but with the opposite sign. The overall benefit is the reduction in the number of multiplications required. There are different approaches to decimation which yield the same overall result but with greater internal complexity towards either the time-domain or frequency-domain end of the chain of matrices; these are known as 'decimation in time' and 'decimation in frequency', respectively. 'FFT' is really a generic name for this type of algorithm and there are many variants, the main differences being in the paths taken through the factored matrices.

The FFT algorithm is commonly based on a radix[20] of 2; that is, the numbers of input and output samples are equal to 2 raised to some power. A larger radix (e.g. 4, 8, etc.) is sometimes used for very large arrays of samples.

In the simplest form of FFT, the frequency-domain samples appearing in its output array cover the same frequency range as the 'parent' DFT, but their arrangement is rather different. It was noted earlier that the useful samples output by a DFT cover the range $\pm f_s/2$, and it was implied that they are symmetrically disposed about 0 Hz. However, it was also noted that the spectrum is repeated, centred about harmonics of f_s, so the negative-frequency samples re-appear between $f_s/2$ and f_s; the sample at exactly f_s is a replica of the sample at 0 Hz. By convention, it is the range 0 Hz to one sample below f_s which is represented in the output array of the FFT.

The inverse FFT can be derived from the inverse DFT in a similar way, and these comments apply equally to its input array of frequency-domain samples.

A1.7 Application to DAB

The DAB system uses 1536 carriers in Mode 1. This requires a 2048-sample inverse FFT in the transmitter and a 2048-sample FFT in the receiver.

The way that an FFT is implemented in hardware depends on the required balance of speed versus hardware complexity. Clearly, parallel processing should yield the greatest speed, whilst using several processes consecutively should reduce the amount of arithmetic hardware, although it may increase the requirement for temporary storage. For DAB, there are options which are more economical of hardware than using 2048 arithmetic devices in parallel, and more economical of processing speed than using one arithmetic device for all computations.

In the DAB channel encoder, the array of complex numbers representing the spectrum of the signal during each 1 ms symbol is applied to the inverse FFT which produces samples of the time-domain signal, for that symbol. These can be converted to analogue form, up-converted and transmitted. In this case, full parallel processing is not necessary because the time-domain samples need to be output consecutively, and not simultaneously, although all of the frequency-domain samples must be available for each computation.

[20] For example, 10 is the radix of the decimal numbering system.

In the DAB receiver, the frequency-domain spectrum is derived from the incoming time-domain signal, symbol by symbol, using the forward FFT. However, this signal appears via an ADC as a series of consecutive samples, so a mirror-image of this approach can be used. This will produce the spectrum for each symbol when all of the samples have been received, but computations can start when only a small number of time-domain samples are available; two, for example.

It might seem wasteful to have to use 2048 samples to represent the 1536 carriers, apparently wasting 512, but these can be put to good use by purposely setting their amplitudes to zero. This can be used in the encoder and the receiver to simulate a band-pass filter with an amplitude frequency response much steeper at the band edges than can be achieved using an analogue filter.

A1.8 Hardware examples

The second-generation experimental DAB receivers operate with a signal composed of 224 carriers, and use a 256-sample FFT which is performed using a single proprietary FFT chip; the TMC2310 made by TRW. The device has a resolution of 19 bits internally, and 16 bits at its input and output ports.

In the current third-generation experimental receiving and transmitting equipment, the FFTs are implemented using multiple general-purpose DSP (Digital Signal Processor) devices.

A1.9 Complex numbers

The Fourier transform operates with complex numbers in both domains, and this applies to the DFT and FFT derived from it, and their inverses. Whilst it is usually necessary to consider components of a spectrum as complex, having amplitudes and phases, the waveform of a radio signal is purely real; simply the variation of a voltage with the passage of time. This is not to say that such a waveform could not be specified using complex quantities, only that division of its specification into real and imaginary parts would require that they be combined in some way before the final waveform could be generated.

Essentially, the input and output arrays of the FFT can be divided into real and imaginary parts. When complex numbers are represented, each real sample has an imaginary sample associated with it. Conventionally in this field of engineering, the real and imaginary parts of the time-domain array (i.e. the input array of an FFT, or the output array of an inverse FFT) are referred to as the ‘I’ and ‘Q’ *ports*, for ‘In-phase’ and ‘Quadrature’, respectively.

A1.10 Negative frequencies

It was noted earlier that, up to the Nyquist limit, the DFT and the FFT produce as many output samples as are input, but in each case, half of the frequency-domain samples represent negative frequencies. Real radio signals are usually thought of as using only positive frequencies, but the mathematically rigorous definitions of their spectra should include components at negative, as well as positive, frequencies. All of the transforms being discussed require these full definitions.

For example, the spectrum of a cosine wave with frequency *f* contains two positive impulse functions (i.e. lines in the spectrum) at plus and minus *f*, each multiplied by half the amplitude of the wave. Of course, by trigonometry $\cos(-x) = \cos(x)$, so this is no different from the simplified view of a single positive impulse function at plus *f*, having both halves of the amplitude. In this simplified view, the negative frequencies are effectively 'folded' about 0 Hz, over into the positive frequency range. The spectrum is slightly more complicated for a sine wave of frequency *f*, because $\sin(-x) = -\sin(x)$; the two impulse functions at $\pm f$ are each multiplied by half the amplitude but with opposite signs, the positive-frequency one having negative sign.

When such spectra are calculated using the FFT or DFT, where the input waveform is expressed as a purely real function of time (i.e. it is presented to the I port, and zero is presented to the Q port), the transform of the cosine wave is purely real whilst that of the sine wave is purely imaginary. This might be expected in view of the orthogonal relationships of cosine and sine waves, or real and imaginary numbers. It can be shown that if a sine wave is expressed as a purely imaginary function (i.e. it is presented to the Q port, and zero is presented to the I port), the resulting transform is purely real.

It follows that for an inverse FFT to output a single sine or cosine wave, it must be presented with frequency-domain data for the negative-frequency component as well as for its positive-frequency counterpart, and the relationship between these data and their real/imaginary status must follow certain rules.

A1.11 Linearity of the transforms

At first sight, this would appear to imply that an *N*-sample FFT could only provide *N*/2 useful frequency-domain samples, so the 1536 carriers used by the DAB system would require the use of a 4096-sample FFT, and inverse. However, there is a simple way to halve this requirement by exploiting a useful property of the FFT and its forebears, that of linearity.

If two time-domain waveforms are added and the sum is transformed, the result is the sum of the two corresponding spectra. By reversing the sign of one of the input waveforms, the result is the difference between the two spectra. This property also applies, in reverse, to the inverse transforms.

Therefore, if a real cosine wave and an imaginary sine wave, of equal amplitudes and the same frequency *f*, are (complex) added and applied to an FFT, the result is one positive impulse function at minus *f*, at the real output port, multiplied by the amplitude of either wave; the positive-frequency component is cancelled. Since the two input waves are purely real and imaginary, respectively, the complex addition consists of no more than applying them simultaneously to the appropriate I and Q input ports. If the sign of the imaginary sine wave is reversed, the result is one positive-real impulse function at plus *f*, and the negative-frequency component is cancelled.

Combinations of two sine waves, or two cosine waves, do not cause cancellation of the second impulse function. An FFT would output a single impulse function as one sample, amongst many, having a non-zero value.

The various permutations which yield a single impulse function are listed below in Table A1.1.

I	**Q**	*-f*	**Re**	*+f*	*-f*	**Im**	*+f*
cos	sin	+		0	0		0
cos	-sin	0		+	0		0
-cos	sin	0		-	0		0
-cos	-sin	-		0	0		0
sin	cos	0		0	+		0
sin	-cos	0		0	0		-
-sin	cos	0		0	0		+
-sin	-cos	0		0	-		0

Table A1.1 - samples output by an FFT for simultaneous SIN and COS inputs

By the reverse argument, if a single positive-frequency sample is applied to the real input port of an inverse FFT, with a positive value, the time-domain result is a cosine wave at the I output port and a negative sine wave at the Q output port. The amplitudes of these sampled time waveforms are equal and are proportional to the input sample value, and their frequencies correspond to the position of the sample in the input array. With a negative-value input sample, the result is a negative-real cosine wave and a positive-imaginary sine wave. These, and other permutations can be derived from the above table by reading right to left.

Thus, it is possible to use the negative-frequency samples of an inverse FFT, independently of their positive-frequency counterparts, to produce combinations of sine and cosine waves. What is then needed is a method for combining the sampled waves appearing at the I and Q output ports to produce the required single output signal, and this can be achieved using a quadrature modulation system.

A1.12 Combination of I and Q

The I and Q data are applied separately to a pair of DACs (or one, with time multiplexing) to produce a pair of sampled baseband signals, which are then applied to low-pass filters to construct analogue signals (i.e. to remove the artefacts of sampling). Note that these filters are not used to implement matched filtering (e.g. cosine roll-off), as is used in some other digital modulation systems; that function is effectively carried out by the inverse FFT and the FFT in the receiver.

The filtered baseband signals are applied to a pair of mixers (i.e. multipliers), which are also fed with synchronous local-oscillator signals having 90^o (i.e. $\pi/2$) phase difference; that is, cosine and sine waves. The local-oscillator frequency is equal to the desired centre-frequency of the final signal. The outputs of the two mixers are then added to give the final signal which, after band-pass filtering to remove harmonics and other spurii, can be transmitted. This quadrature modulation system is illustrated in Fig. A1.

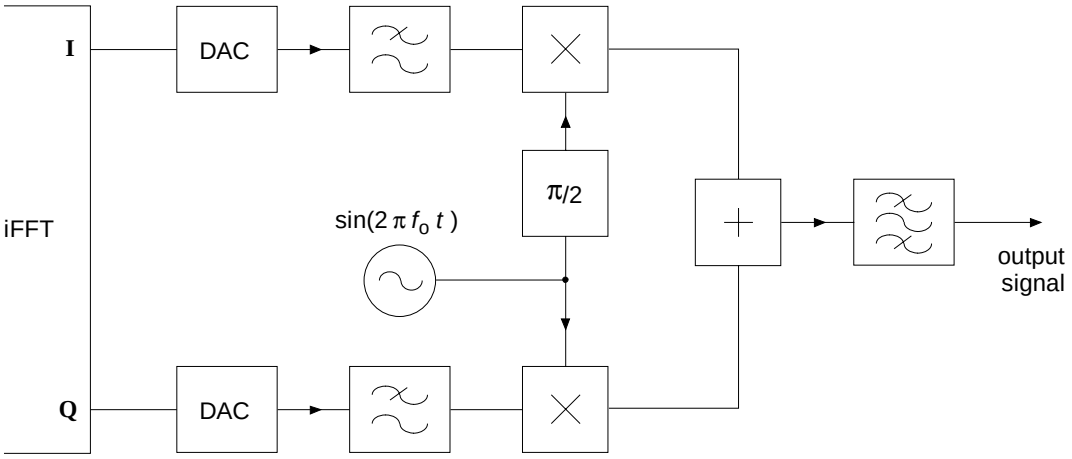


Fig. A1 - the quadrature modulation system

Taking, again, the example of a single positive-frequency sample applied to the real input port of the inverse FFT, with a positive value, the baseband cosine wave in the I channel is multiplied by the cosine local-oscillator signal and the baseband negative sine wave in the Q channel is multiplied by the sine local-oscillator signal. Each multiplication produces sum and difference-frequency cosine components (i.e. double-sideband suppressed-carrier AM), but in this case the difference-frequency components cancel when the mixer outputs are added; if in doubt, consult a table of trigonometric identities! Thus, the final signal contains a single cosine wave at the local-oscillator frequency plus the baseband frequency.

The various combinations are listed below in Table A1.2, where f_0 is the local-oscillator frequency; the mathematics have been simplified for the sake of clarity (i.e. $2\pi t$ has been omitted in several places).

-f	**Re**	*+f*	*-f*	**Im**	*+f*	**I** ($\times\sin f_0$)	**Q** ($\times\cos f_0$)	**Output**
0		+	0		0	cos	-sin	cos ($f_0 + f$)
0		-	0		0	-cos	sin	-cos ($f_0 + f$)
0		0	+		0	sin	cos	sin ($f_0 + f$)
0		0	-		0	-sin	-cos	-sin ($f_0 + f$)
+		0	0		0	cos	sin	cos ($f_0 - f$)
-		0	0		0	-cos	-sin	-cos ($f_0 - f$)
0		0	0		-	sin	-cos	sin ($f_0 - f$)
0		0	0		+	-sin	cos	-sin ($f_0 - f$)

Table A1.2 - waves output by an inverse FFT for a single input sample

The order of the rows in Table A1.2 has been chosen to show clearly that four phases are available, at either the sum or difference frequency, simply by selection of appropriate input data to the inverse FFT; hence a QPSK signal can be generated. Of course, the principle of linearity can be exploited further to produce simultaneously a sine and cosine wave, so any desired phase of output signal can be synthesised; the $\pi/4$ offset QPSK method used in the DAB system is readily achievable. Furthermore, this principle can be extended to cover the simultaneous generation of 1536 output waves; that is, the Mode 1 DAB signal.

In the third-generation channel encoder, a 2048-sample inverse FFT is, indeed, used to produce the 1536-carrier Mode 1 signal. The computations are processed with a time-window duration of 1 ms, the symbol duration. Therefore, the frequency-domain sampling has an interval of 1 kHz, so adjacent input samples correspond to sinusoidal output waves separated by 1 kHz. Assuming conventional ordering of the input array, the first 1024 samples represent positive frequencies, from 0 Hz to 1023 kHz in the baseband signals. The 1024th sample represents ±1024 kHz, which is at the Nyquist limit and is probably unusable. The remaining samples represent negative frequencies, from -1023 kHz to -1 kHz.

Of these, active data are applied to the 1536 surrounding 0 Hz; that is, those representing 1 kHz to 768 kHz and -1 kHz to -768 kHz, and static zeros are applied to the remainder. The inverse FFT then produces 2048 time-domain samples which are output to the DACs within 1 ms. Following up-conversion by the quadrature modulation system, carriers appear at frequencies from f_0 - 768 kHz to f_0 + 768 kHz.

The input data are changed from symbol to symbol giving the effect of QPSK modulated carriers. Of course, discontinuities occur when the signal is re-configured abruptly at the symbol boundaries, and this changes the fine detail of the spectrum. If the modulation data are random, then over many symbols the power spectrum of each modulated carrier takes on a $(\sin f/f)^2$ distribution, as described in the main text.

In this approach, the sample which represents 0 Hz in the baseband signals is not used. If data were input to this sample, the corresponding I and Q output signals would be static (DC) voltages, and if the mixers in the quadrature modulation system could handle such signals, the output signal would be a wave at the local-oscillator frequency. However, the accuracy with which the phase of this wave could be controlled could be compromised by drifts in the DACs, and the mixers themselves[21], so this 'zero carrier' is not used to carry modulation.

The mirror image of this approach is used in the experimental DAB receivers, where the incoming signal is converted to an IF and band-pass filtered, and then applied to a similar quadrature modulation system. In this case, the local-oscillator frequency is the centre-frequency of the IF band and, of course, ADCs are substituted for the DACs.

What has been described is the approach that has been taken, so far, in all successive generations of experimental DAB transmitter equipment. However, it is worth noting that the FFT can be applied to multi-carrier signal generation and reception in other ways, some of which do not require the quadrature modulation system. Also, it is possible to implement a quadrature modulation system in the digital domain, rather than the analogue domain.

A1.13 Addition of the guard interval

It was noted earlier that if the frequency-domain data input to an inverse DFT remain static for consecutive time windows, then the time-domain result will be periodic over the consecutive windows, and this applies equally to the inverse FFT. In the DAB channel encoder,

[21] The common form of RF mixer is composed of a ring of diodes, and this is not noted for its accuracy as an analogue multiplier. Its inaccuracy is of little consequence when AC signals are multiplied because the result is harmonic signals which can easily be removed by filtering.

the window duration is equal to the active symbol duration, 1 ms in Mode 1, and the signals output by the transform are all sinusoidal waves at harmonics of 1 kHz, so they are periodic over whole multiples of the active symbol duration. Therefore, consecutive inverse FFT operations with the same input data would yield continuous waves in the baseband signals. For example, if the data were held static for two consecutive symbols, the waves would be continuous over 2 ms.

In the DAB receiver, in order to accomplish a single FFT operation, it would not matter at what instant the time-window began as long as the input waves were continuous over 1 ms. A time displacement would only change the apparent phase of each of the waves, which would alter the absolute values of the data output by the FFT, but since the phase modulation is coded differentially, a slow change would not corrupt the decoded data. Thus, this example of an 'oversized' guard interval would permit time-agility in the receiver and the simultaneous reception of delayed signals as long as the limit of temporal coherence was not exceeded (i.e. the maximum vehicle speed would be limited further in this case).

Of course, there is no need to repeat the inverse FFT operation when the output samples can simply be stored, and read out to the DACs directly after the active symbol. Also, as noted in the main text, the DAB system actually uses the more-economical compromise of a 246 μs guard interval, so only about one quarter of the output samples need to be stored. According to available information, the samples making up the guard-interval appear to be applied to the DACs *before* each active symbol, so they are repetitions of the last quarter of those produced during the active symbol. This alternative arrangement does not imply additional storage, merely a re-ordering of the inverse FFT; either method is equally valid.

A1.15 Bibliography

The FFT algorithm is discussed in many books which deal with transform techniques, but these usually make extensive use of mathematics. The following book is very readable and includes intuitive developments of the Fourier transform, the DFT and the FFT, assisted by helpful diagrams:

1. BRIGHAM, E. O. 1974. The fast Fourier transform. Prentice-Hall Inc.

The use of dedicated FFT chips and DSP devices appears to be a relatively recent development so in this, and many other books from previous decades, the discussion of practical application of the FFT is limited to its implementation by means of a computer program. Integrated circuit data sheets can be more helpful in this respect, but, necessarily, they assume a great deal of background knowledge.

The application of FFTs specifically to an early incarnation of the DAB system is discussed in the following collection of EBU texts:

2. EBU. 1988. Advanced digital techniques for UHF satellite sound broadcasting; collected papers on concepts for sound broadcasting into the 21st century. EBU Technical Centre, August 1988. pp. 52 - 55.

APPENDIX 2

CONVOLUTIONAL ENCODING AND VITERBI DECODING

A2. Error correction coding

When the transmission channel is disturbed by noise or interference, the values of some of the bits in the sequence recovered from the received signal will be different from those that were transmitted; there will be bit-errors. The fraction of recovered bits that are in error is known as the Bit-Error Ratio (BER), and this is used to quantify the effect of a disturbance. Error correction coding provides an improvement in the BER of the decoded bit-stream in such conditions, relative to the un-coded case.

This is achieved by transmitting each possible sequence of bits as a unique 'code-word', using more than the minimum number of bits, so when a recovered code-word is altered by errors, this will show up as a sequence of bits which could not have been transmitted. The additional 'redundant' bits are chosen to increase the uniqueness of each code-word; to reduce the likelihood of an altered one appearing as one which could have been transmitted. The benefit of coding the data in this way increases with the amount of redundancy, but the cost is the additional capacity needed to transmit the redundant bits. The ratio of the number of bits input to the encoder to the number output is known as the 'code rate'; for example, at rate 1/3, three bits are output for each one input so the redundancy accounts for two thirds of the transmitted bits.

There are many different approaches to error correction coding, each with its own merits. In most cases, the received signal is decoded in a way that averages random effects, such as noise, over many bits. This means that the encoding must be spread over a number of consecutive input bits (i.e. a block). The benefit of averaging increases with the size of the block, but the cost is increased data storage (i.e. hardware complexity and processing delay). The optimum approach for a particular application depends on factors such as the required 'coding gain' (i.e. BER improvement), and the acceptable amounts of redundancy, delay and hardware complexity.

A2.1 Block coding

Block coding is relatively straightforward to explain. The serial bit-stream to be transmitted is divided into consecutive blocks of bits which are held static whilst the encoding operation is performed. Some arithmetic formula is applied to the contents of the block (e.g. a check-sum) and the resulting bits are interleaved with the original data bits for transmission; in that case, the code is referred to as 'systematic'. Alternatively, in the case of a 'non-systematic' code, only the bits resulting from the application of the formula are transmitted.

If two code-words are compared by counting the number of bit positions where the bit value is different, the resulting number is known as the 'Hamming distance'; for example, the Hamming distance between 000100 and 100001 is 3. A measure of the error correction capability provided by a block code can be gained by considering the Hamming distance between any two different code-words. The minimum value for all possible code-words is

known as the ‘minimum distance’, and from this can be calculated the maximum number of erroneous bits in a single code-word that can always be corrected; this is the integer part of ½ (minimum distance - 1). If the minimum distance is 3, then one error can always be corrected; the altered sequence differs from the correct sequence in one bit position but differs from all other sequences in at least two. Therefore, it would be possible to decide from all of the sequences that could have been transmitted, to which the erroneous sequence most closely corresponds.

A2.2 Convolutional coding

A different approach known as convolutional coding is used in the DAB system, and this is fairly common nowadays. It has been chosen as the best compromise for many data communication systems which use radio transmission channels.

In a convolutional encoder, the serial bit-stream can be considered as passing continuously through a shift register, with a fixed number of ‘taps’ known as the ‘constraint length’. The different taps provide versions of the bit-stream delayed by different amounts. Formulae, known as ‘generator polynomials’, are applied to the taps and these produce a set of resulting bits as each new bit is shifted in. The resulting bits alone are then interleaved to form the output sequence. The ‘output’ of the shift register is regarded as just one of several taps so a convolutional code is generally non-systematic, but if any of the polynomials uses only one tap then the output bit-stream will contain the original data (albeit interleaved) and the code will be systematic. The effect is analogous to the mathematical process of convolution, of the input bit-stream with the impulse response of the encoder, and hence the name. The impulse response corresponds to the output sequence for an input sequence of ....0001000...

A simple convolutional encoder is illustrated in Fig. A2.1, where the constraint length is 3 and two generator polynomials are used.

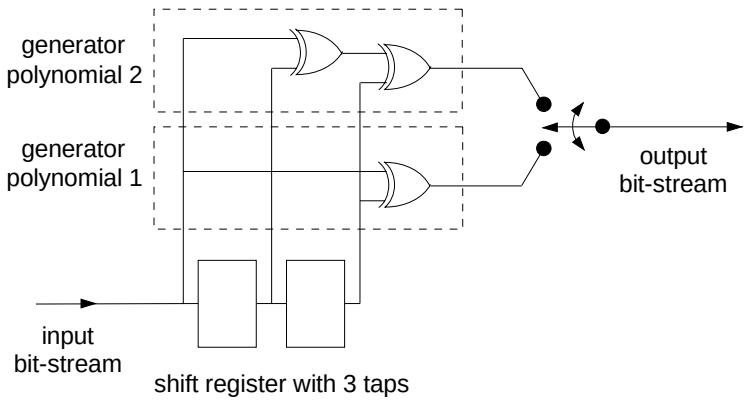


Fig. A2.1 - example of a simple convolutional encoder

An exclusive-OR gate has the effect of a modulo-2 adder; that is, the output represents the least significant bit of the sum of the two 1-bit input numbers. The combination of the bit-streams output by the two polynomials would be implemented digitally in practice;

a switch is shown only to simplify the illustration. For rate 1/2 coding, the shift register is clocked at half the rate at which the switch is toggled, so two bits are output for each one input.

The values of the bits stored in the first and second stages of the shift register (i.e. appearing at the second and third taps) can be taken to represent its 'state', and this can have four values: binary 00, 01, 10 and 11, or decimal 0, 1, 2 and 3. For a convolutional encoder, the block of input bits, and the resulting code-word, could be considered to have unbounded length. There is no unique relationship between the transmitted bits and the input bit at any point in the sequences; the relationship depends on the previous state of the encoder and the value of the current input bit, so it involves the 'history' of input bits over the span of the shift register.

A2.3 Tree and trellis diagrams

The possible output sequences for an arbitrary input sequence can be represented by a 'tree' diagram, as shown in Fig. A2.2, starting with the state 0.

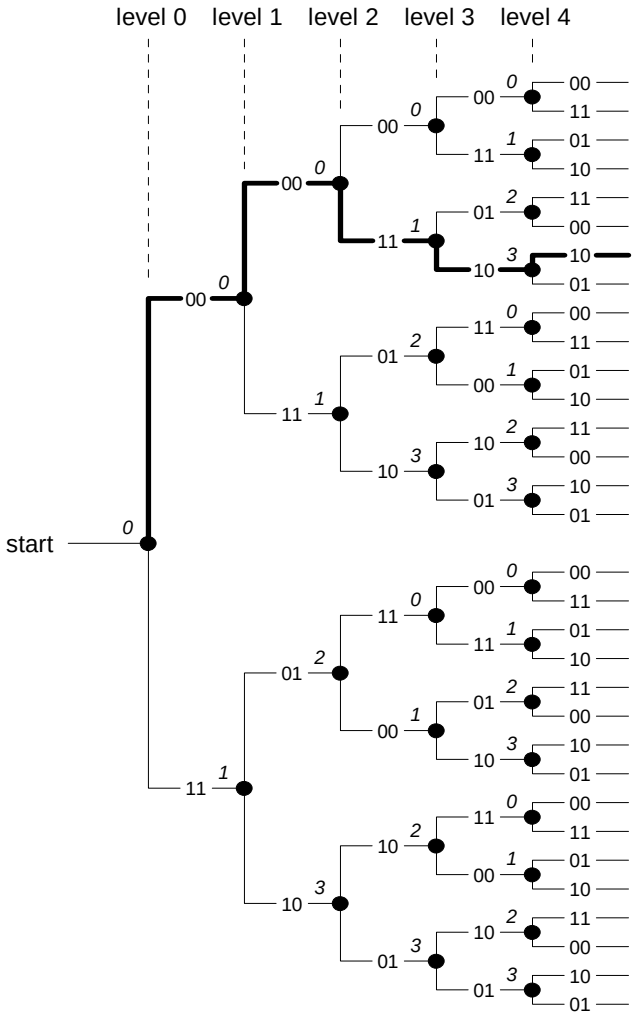


Fig. A2.2 - tree diagram for the encoder shown in Fig. A2.1

With each consecutive input bit, the output sequence can be built up by working from left to right, drawing a path between the nodes which are denoted by black spots. Each node is numbered (in *italics*) with the state of the encoder on the way in to that node, and each of the two onward branches shows the binary code which would be output by the encoder for an input bit of 0 (upper branch) or 1 (lower branch). For example, the path corresponding to an input sequence of 00110 is shown as a bold line, and this produces the code-word 00 00 11 10 10 (with the convention that the first occurring bit in a series is written at the left-hand end).

Only five 'levels' of nodes are shown here, but the tree could contain all possible code-words if it was sufficiently expanded. The number of nodes in each level would go on increasing if the tree was expanded, and the number of possible paths would rise exponentially. However, the nodes in the upper and lower halves of the tree at level 3 correspond to identical operations, so half of them can be omitted and the paths cross-linked to their counterparts leaving only four nodes; this is illustrated in Fig. A2.3.

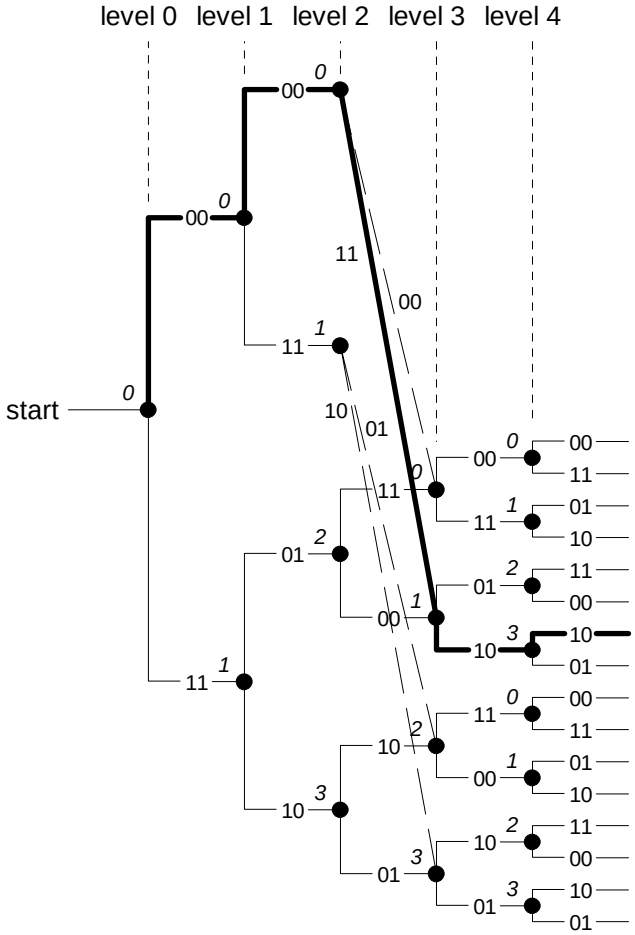


Fig. A2.3 - cross-linking at level 3

The cross-linking can be repeated at level 4, and thereafter, leaving only four nodes at each subsequent level.

The result can be re-drawn as a more-compact 'trellis' structure shown in Fig. 2.4; the same example path is shown here as a bold line. In this case, although the number of possible paths still rises exponentially with the number of levels, these paths are forced to pass through a limited number of nodes.

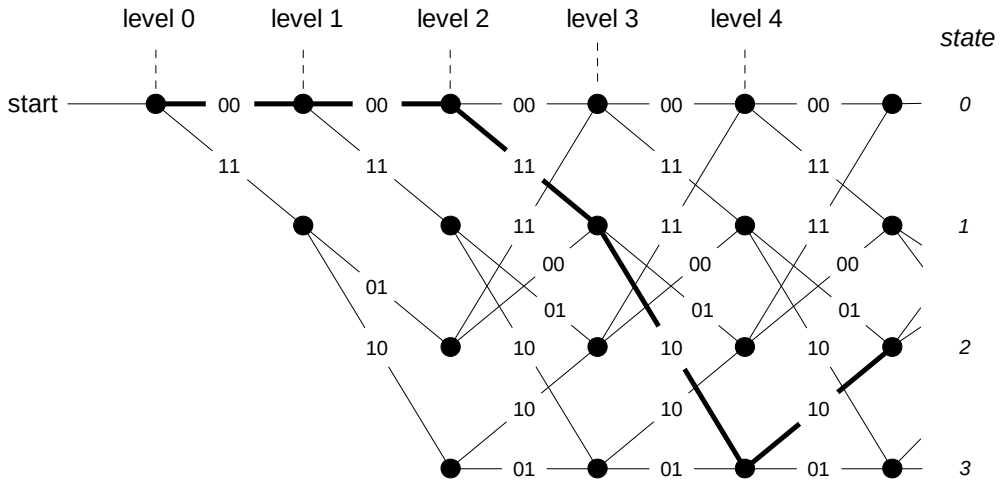


Fig. A2.4 - trellis diagram for the encoder shown in Fig. A2.1

A2.4 Decoding convolutional codes

The objective of the decoder is to estimate from the sequence of bits recovered from the received signal, the sequence of bits that were fed into the encoder. Inevitably, the accuracy of this estimate will be compromised at large values of BER. Generally, the best that can be done is to compare the recovered sequence with all of the possible transmitted sequences, to determine the relative likelihood of those which nearly match, and to choose the one most likely. This is known as 'maximum likelihood' decoding.

Maximum likelihood decoding could be achieved by comparing the recovered sequence with the sequence represented by every possible path in the tree shown in Fig. A2.2, and choosing the path which gives the smallest Hamming distance, accumulated from level to level. The accuracy of the estimation would increase as the sequence was lengthened, but the number of paths and the amount of computation required would increase exponentially. Fortunately, a better approach has been found and this is known as the Viterbi algorithm, after the person to whom its discovery (in the late 60s) is credited.

A2.5 The Viterbi algorithm

The key to this approach is representation of the possible transmitted sequences by the trellis shown in Fig. A2.4. As the recovered sequence develops, paths can be traced out between nodes from one level to the next, and the possible transmitted sequences that these represent can be compared with the recovered sequence. For each path, a 'metric' can be calculated which indicates the similarity between the sequence represented by that path and the

recovered sequence; a greater metric indicating greater similarity[22]. At some later stage the path with the largest metric can be judged to have the maximum likelihood, and the most likely sequence that was fed into the encoder can then be deduced.

Viterbi's breakthrough was to recognise a way to constrain the number of developing paths as the sequence is lengthened, and hence to constrain the amount of computation required. Up to level 3, the number of paths does increase exponentially but, thereafter, all paths must pass though the limited number of nodes in each level. At level 3 and beyond, each node has two paths leading into it and two paths leading out. If a decision is made at each node in these levels to pursue the path with the greater metric and to discard the other, this choice cannot prove to be incorrect later on because the two possible paths would have emerged from the same node; their metrics could only remain the same or be reduced thereafter.

Furthermore, by discarding one incoming path at each of these nodes the number of remaining paths to be considered (the 'survivors') will not increase further as the sequence is lengthened; one path is discarded for every new one generated. It is probable that any two survivors will meet at a node in some higher level and often the field can be narrowed down to only one survivor within a remarkably short sequence, two or three times the constraint length of the encoder. A decoder 'length' (i.e. the length of the equivalent trellis) of four or five constraint lengths is found to offer near-optimum performance.

Thus, a Viterbi decoder models the trellis of its counterpart encoder, and the structures of the two are intimately related. Generally, the number of nodes in each level, beyond level 2, is 2^{k-1}, where k is the constraint length of the encoder, and acceptable complexity sets the upper limit on the constraint length at perhaps 10. The algorithm can be implemented in hardware using conventional arithmetic and memory logic elements and, in practice, metric calculations and decisions are made sequentially for nodes in one level at a time. As soon as the input sequence exceeds the length of the decoder, bits can be output representing the maximum likelihood transmitted sequence; the rates at which bits are output and input are related by the code rate.

A2.6 Puncturing

Although the preceding example was configured for rate 1/2, the same coding scheme can be applied for other rates. However, there are major structural differences between a Viterbi decoder for rate 1/2, as described, and rate 2/3 for example, even if the constraint length is not changed. At rate 2/3 with a constraint length of 3, each branch of the trellis would correspond to a transmitted sequence of three bits rather than two, and each of the higher-level nodes would have four paths entering it and four paths emerging from it; the decisions would be of one path amongst four.

In the DAB system (and many other cases), different types of data have different sensitivities to errors and it would be wasteful of data capacity to use more than the necessary amount of redundancy. Consequently, there is a requirement for coding at a selectable rate but complete

[22] This could be based on the reciprocal of the accumulated Hamming distance, or on the cross-correlation of the two sequences (substituting +1 and -1 for the 1 and 0 bit values); correlation of sequences is explained in Appendix 3.

re-configuration of the encoder and decoder hardware would be complicated if a large number of different rates was required. If the hardware is configured for the most powerful code rate required then there is a simpler alternative known as 'puncturing'.

If the data are encoded at rate 2/4, which is essentially the same as rate 1/2, and then every fourth bit in the transmitted sequence is omitted (i.e. not transmitted, and the following bits are shuffled up to fill the 'hole'), the effective code rate becomes 2/3. The corresponding trellis diagram is illustrated in Fig. A2.5, where each possible omitted bit is represented by an '×'.

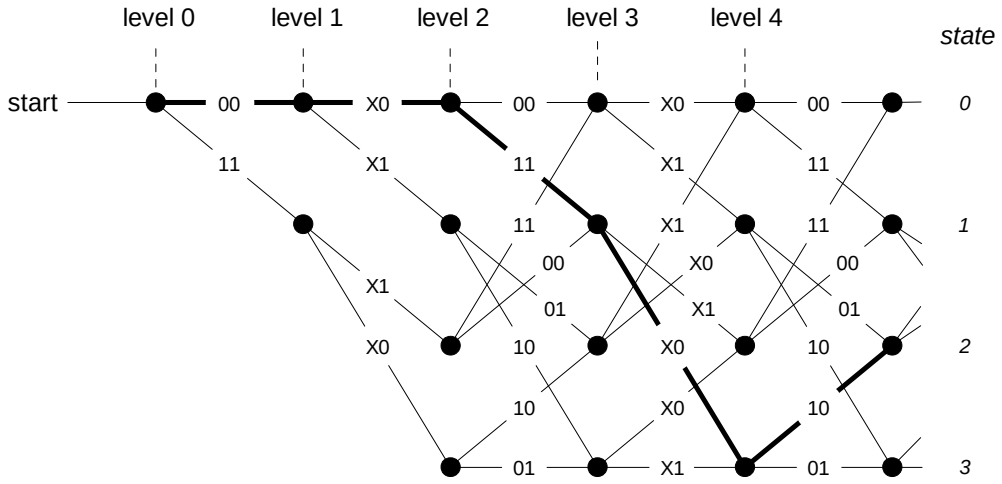


Fig. A2.5- trellis diagram for the encoder of Fig. A2.1 with every fourth output bit omitted

This trellis can be implemented by a Viterbi decoder provided that the occurrence of puncturing events (i.e. omitted bits) is known. Passage through the trellis is advanced by one bit at each of these events, as though a bit had been received, and this could be achieved simply by inserting bits of arbitrary value in the input bit-stream. It is clear from Fig. A2.5 that no decision will be made at an individual node on the basis of the value of one of these bits, and the operation of the decoder can be made independent of them by not including them in the computation of metrics. Thus, by means of puncturing, a Viterbi decoder can use the same trellis structure for several different code rates, with only minor modifications to the metric arithmetic.

Puncturing is simply a reduction in the amount of redundant bits that are sent; when less protection is required, more bits are punctured and more transmission bit-rate is saved. Generally, if the structures of the encoder and decoder are based on a rate m/n code, where m is less than n and both are integers, puncturing allows operation at rates m/k, where k is an integer in the range n to m (i.e. up to rate 1; no error correction).

If the code-word which results from puncturing is identical to that produced by a dedicated non-punctured encoder of the same rate, there is no loss of error correction performance. However, if a large range of different rates are required in practice, puncturing can cause a slight loss of performance relative to a non-punctured code of the same rate; this depends on the choice of generator polynomials and which bits are punctured.

A2.7 Application to DAB

The convolutional coding used in the DAB system employs these principles but is considerably more complicated than the examples given so far. The constraint length is 7 so the encoder has 64 states. Fundamentally, it uses 4 generator polynomials and the generated bits are transmitted in a fixed sequence cycling through the 4 bits, so the encoder operates at rate 1/4. This can provide very powerful error correction, albeit with extreme consumption of the available bit-rate. An example of an encoder with a constraint length of 7 and four generator polynomials is illustrated in Fig. A2.6. Again, the combination of the bit-streams from the four polynomials is implemented digitally in practice; a rotary switch is shown only to simplify the illustration. For rate 1/4 coding, the shift register is clocked at one quarter of the rate at which the switch is incremented.

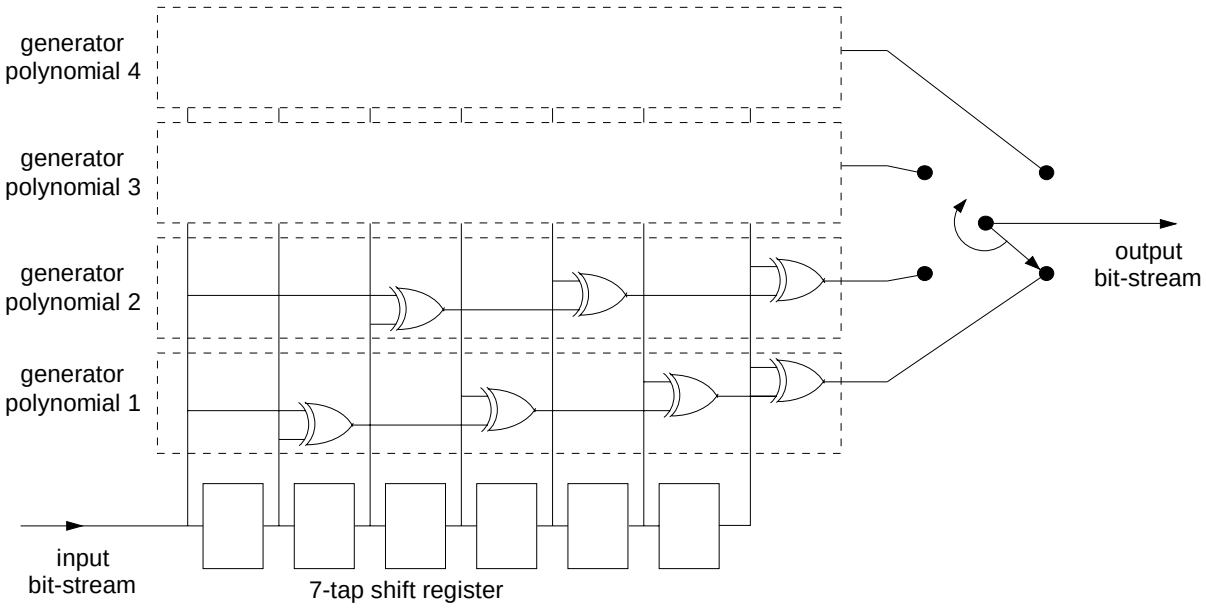


Fig. A2.6 - example of a convolutional encoder with four generator polynomials

Interpreting rate 1/4 as rate 8/32, puncturing allows operation at all code rates between 8/32 and 8/9 (8/8, no protection, is not used). The choice of which bits in each group of 32 are punctured is determined by a puncturing 'code', available to both the encoder, and the decoder in the receiver, as an entry in a look-up table. The index of this entry is transmitted regularly at predictable times by the MCI, to which a constant, known puncturing rule is always applied (rate 1/3). With reference to Fig. A2.6, when puncturing is applied, the rotary switch is incremented by more than one step at a time, skipping one or more of the polynomial outputs. In two particular cases, for rates 8/24 = 1/3 and 8/16 = 1/2, the operation is simplified because the outputs of one or two of the polynomials are not used at all.

In the corresponding trellis, and therefore the Viterbi decoder, four paths emerge from each node. At the higher levels, there are 64 nodes in each level and four paths enter each node; one path amongst two must be chosen at each of these nodes. Each branch represents a sequence of four transmitted bits. All three generations of experimental receiver use proprietary Viterbi decoder chips manufactured by SOREP.

A2.8 Performance

The coding gain of a convolutional code depends on the code rate, the constraint length and uniqueness properties endowed by the generator polynomials; their choice is not obvious, and may be the result of extensive computer searches (as may be the puncturing codes).

Assessment of the number of errors that can be corrected is more complicated than in the case of a block code (see Section A2.1) because of the unbounded length of the code-word; there is no single counterpart to the minimum distance. Nevertheless, one useful parameter in the context of Viterbi decoding is the 'free distance'. This is the minimum Hamming distance between any two code-words as the length, and the number of code-words considered, approaches infinity. On the basis that truncating the length of the Viterbi decoder (from infinity) to four or five times the constraint length incurs little penalty, the free distance is a first-order guide to the number of errors that can always be corrected in an input sequence the length of the decoder. The coding used in the DAB system is based on an un-punctured code for which the free distance is 10, so this indicates a capacity for always correcting up to 4 errors in the decoder trellis simultaneously. However, this is an incomplete assessment and some further explanation can be found in the books listed in the Bibliography (Section A2.12).

The combination of a convolutional encoder and a Viterbi decoder works well in the presence of a continuous stream of randomly placed errors, which may be caused by noise or inter-symbol interference. For the non-coded case, the BER of the recovered bit-stream would increase, with reducing S/N of the input signal, as illustrated in Fig. A2.7.

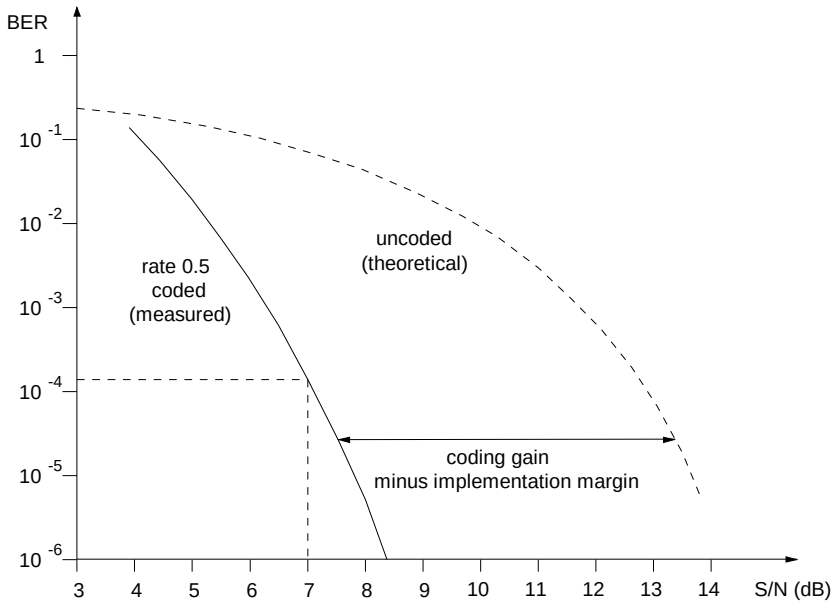


Fig. A2.7 - relationship between BER and S/N for the third-generation DAB receiver

Convolutional encoding, at rate 1/2 with a constraint length of 7, and Viterbi decoding offers a coding gain which approaches 7 dB at large values of S/N in the presence of Gaussian channel noise. The 'coded' curve in Fig. A2.7 corresponds to the measured BER of the bit-stream fed to the ISO source decoder in the third-generation experimental receiver, and this includes an element of 'implementation margin' of about 2 dB (i.e. an allowance for receiver imperfections, such as RF self-interference caused by fast logic circuitry). The effect of the remaining errors is clearly audible at a BER slightly greater than 10^{-4}, which corresponds to a S/N of about 7 dB.

At values of S/N smaller than shown in this figure, the coding gain would become fractional; the BER of the decoded bit-stream would be greater than in the un-coded case. The Viterbi decoder fails when an excessive number of the bits stored in the trellis are in error, and some proportion of these must be flushed out before normal error correction can resume. This causes extension of the duration of such events, so the output BER can be greatly increased. For this reason, the Viterbi decoder is not resilient to bursts of noise or interference, but the DAB system employs interleaving in order to overcome this limitation.

A2.9 Soft decision

One further stage of complication is introduced in the decoder to yield improved error correction performance in the presence of channel noise, and that is 'soft decision'.

If the computation of metrics was based on the Hamming distance, as described earlier, the decoder would sometimes be faced with a draw, cases for which the Hamming distance was the same, so an arbitrary choice would have to be made. This, and other 'rounding error' problems can be resolved to some extent by representing the input bit-stream, and the arithmetic in the decoder, with greater resolution than 'hard' binary ones and zeros. Indeed, the FFT used to decompose the OFDM signal carries out its arithmetic using 16 or more bits, so it is relatively straightforward to implement a so called 'soft-decision' Viterbi decoder in a DAB receiver.

For a transmission channel which introduces only Gaussian noise (such as a satellite down-link), it is found that a 3-bit representation is optimum, giving 2 dB improvement over hard-decision. In this case, an infinite number of bits would only increase the improvement to about 2.25 dB. When the channel also introduces fading which cannot be removed by conventional AGC, there can be value in increasing the resolution to 4 bits.

A multi-bit representation of the demodulator output signal contains an indication of the 'confidence' with which each bit was demodulated. For example with a three-bit word: given '000', it can be inferred with great confidence that a '0' was transmitted, but '010' suggests that it was probably a '0', etc. The two less-significant bits of each word can be considered as a weighting factor, and this can be applied in decisions made within the Viterbi decoder. A neutral value half-way between '000' and '111' would indicate no confidence at all, or zero weighting, so no decisions should be made on the basis of this word. If such words are inserted in the bit-stream input to the decoder at puncturing events, logically, this would allow the decoder to operate with different puncturing without the need for modifications to the metric arithmetic.

Confidence data have other potential uses in a DAB receiver because, independently of the cause of the transmission errors, they provide a guide as to how well the audio data are likely to be decoded. This information could be used to trigger a concealment strategy in the ISO decoder in cases of low confidence, when errors might otherwise lead to audible impairment.

A2.10 Channel equalisation

It is worth noting a matter of receiver implementation which is made possible by the use of soft decision and the availability of FFT processing.

Individual carriers of the OFDM signal may be subject to selective fading or narrow-band interference which can cause errors in the data recovered from them. In the simple case, the capability of the Viterbi decoder is then used to correct the errors so introduced. However, a means exists in the DAB receiver to anticipate some of these errors, so the incorrect bits can be substituted in the soft-decision bit-stream by words with zero (or, at least, smaller) weighting, thereby conserving some of the error correction capability.

All of the transmitted carriers are suppressed for the duration of the null symbol at the beginning of each transmission frame, and this provides a repetitive opportunity for the transmission channel to be inspected for noise or interference. The FFT in the receiver (which might not otherwise be used during the null symbol) can be used to analyse the spectrum of whatever is present in the transmission channel, and to measure the strength of any interfering signal at each of the carrier frequencies. This information can be used selectively to apply an artificial weighting to the data recovered from those carriers which are affected, for the duration of the following transmission frame.

This form of channel equalisation was implemented in the first-generation experimental receiver but it is not clear whether it is implemented in the current third-generation receiver. A measure of its potential benefit is given in Reference 2 to the main text of this document; about 1dB.

A2.11 Caveat

Notwithstanding all that has been said here about the use of the soft-decision Viterbi decoder, there is no reason why a receiver manufacturer wishing to produce a budget receiver, offering relatively meagre performance, could not use hard decision, and even some other form of decoder.

A2.12 Bibliography

The use of these techniques is now so widespread that almost any modern book on digital communications contains a section devoted to this topic. Viterbi's original work was first published in the late sixties, but it is highly mathematical. The following more-recent books contain detailed, but readable explanations of convolutional encoding and Viterbi decoding, including one co-written by Viterbi himself:

1. SKYLAR, B. 1988. Digital Communications - fundamentals and applications. Prentice-Hall International Inc. pp. 327 - 338.

2. BHARGAVA, V. K., HACCOUN, D., MATYAS, R. and NUSPL, P. P. 1981. Digital communications by satellite. John Wiley and sons. pp. 353 - 382.

3. CLARK, G. C. JR. and BIBB CAIN, J. 1981. Error-correction coding for digital communications. Plenum Press. pp. 227 - 238.

4. SPILKER, J. J. Jr. 1977. Digital Communications by Satellite. Prentice-Hall Inc. pp. 455 - 472.

5. HAYKIN. S. 1988. Digital Communications. John Wiley and sons Inc. pp. 393 - 414.

6. VITERBI, A. J. and OMURA, J. K. 1979. Principles of Digital Communication and Coding. McGraw-Hill Inc. pp. 227 - 286.

APPENDIX 3

TIME AND FREQUENCY INTERLEAVING

The DAB system uses interleaving with respect to time (i.e. between frames) and frequency (i.e. between carriers) in order to disperse clusters of errors in the received bit-stream. The interleaving/dis-interleaving processes for these two domains are independent but generally their beneficial effects are additive; certainly for the case of a moving receiver.

Both interleaving processes are based on scattering of the data bits to be transmitted, in order to disperse consecutive bits widely over time or frequency, and re-ordering of the bits received in order to restore the original sequence. In each case, the scattering is governed by a fixed series or sequence, and the way that it is applied will be described in this Appendix.

A3.1 Time interleaving

In the interleaver, 16 different time delays are applied in a fixed repetitive sequence to groups of 16 consecutive bits in a logical frame, so the first bit and the seventeenth bit, for example, are delayed by the same amounts[23]. Of course, all frames may not contain whole multiples of 16 bits, so the last 5 bits, for example, would be subjected to the first 5 delays in the sequence. The sequence starts afresh with the first bit of each new logical frame. The fixed sequence in which the different delays are applied is known as a 'bit-reversal' sequence, illustrated below in Table A3.1.

input sequence		bit-reversal sequence	
numeric	binary	binary	numeric
0	0000	0000	0
1	0001	1000	8
2	0010	0100	4
3	0011	1100	12
4	0100	0010	2
5	0101	1010	10
6	0110	0110	6
7	0111	1110	14
8	1000	0001	1
9	1001	1001	9
10	1010	0101	5
11	1011	1101	13
12	1100	0011	3
13	1101	1011	11
14	1110	0111	7
15	1111	1111	15

Table A3.1 - construction of a bit-reversal sequence

[23] In other words, the incoming bits are counted and given an index equal to the count, then the index is revised 'modulo 16'; this means that when the index is equal to, or greater than 16, the revised index is found by subtracting 16 repeatedly until the result is less than 16. The revised indices of all bits are then between 0 and 15.

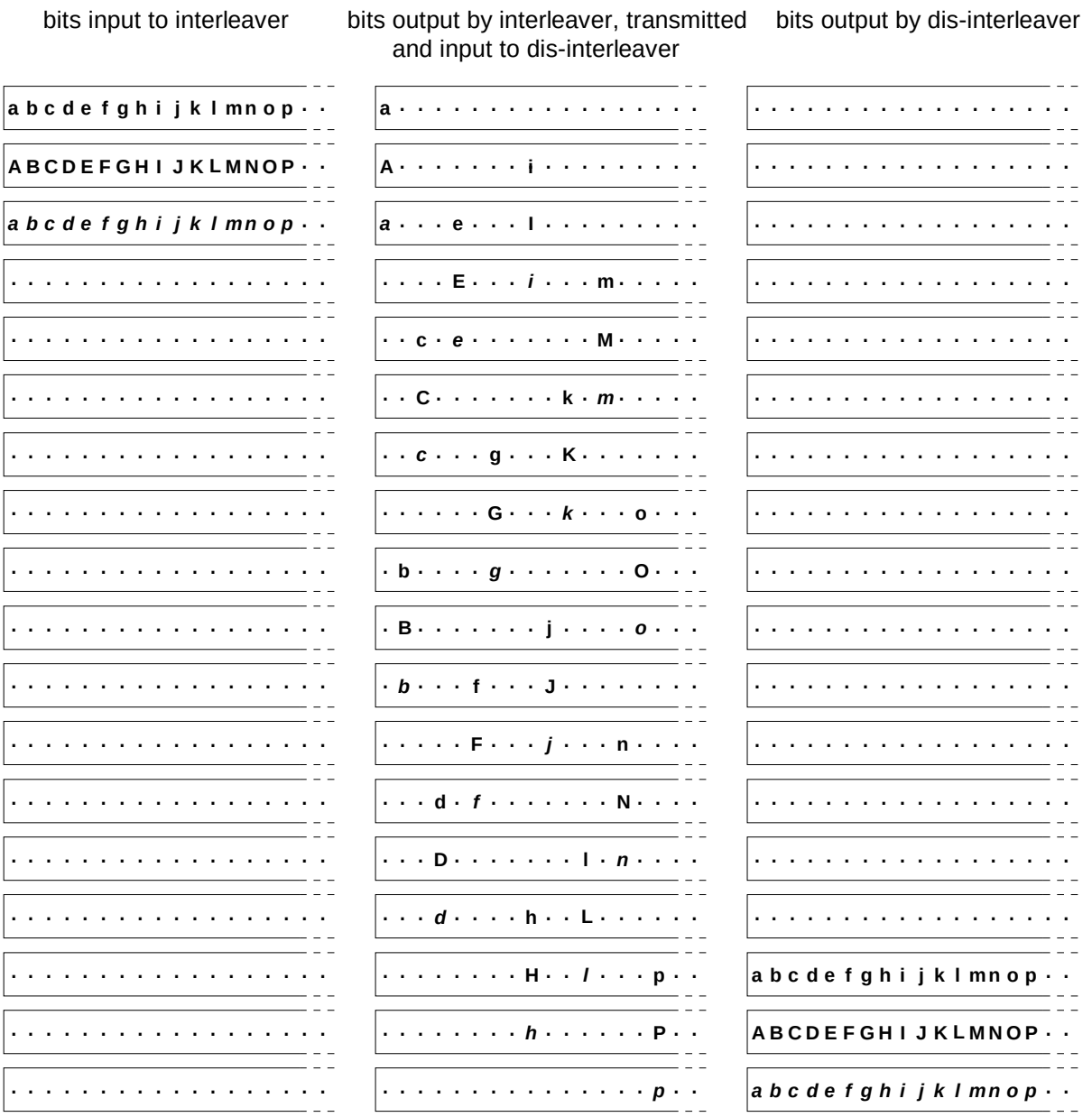
This is constructed by taking the numbers 0 to 15, expressed in binary form, and reversing the order of the four bits. This is achieved easily in hardware or software, simply by re-ordering the bit-pattern in a fixed manner, and in this application it provides sufficient dispersal.

If the numbers in the input sequence are considered as the indices of consecutive bits input to the interleaver, modulo 16, during the logical frame with an index of zero, the bit-reversal sequence gives the index of the later frame to which they are transferred. The magnitudes of the different time delays required for this are the values in the bit-reversal sequence multiplied by 24 ms. The maximum delay affects the last input bit (index = 15) in each group of 16, and its magnitude is 15×24 ms = 360 ms.

The time interleaving and dis-interleaving processes are illustrated in Fig. A3.1, where each box represents the beginning of a logical frame. The passage of time is from left to right within a frame, and from top to bottom from one frame to the next. The first 16 bits in three consecutive un-interleaved frames are indicated at the top left-hand corner, and these are represented by the letters A to P; **lower-case** letters are used in the frame which is transmitted first; **UPPER-CASE** and ***italic*** letters are used in the two subsequent frames. The dots represent later bits, beyond the scope of this discussion.

Take, for example, the fourth bit in the first frame, '**d**'. Its index is 3, for which the bit-reversal sequence gives a value of 12, so its transmission is delayed by 12 frames. At the output of the interleaver, it appears in the frame which is transmitted 12 frames later than its original first frame. The dis-interleaver makes up the delay of all bits to the same value, 15 frames or 360 ms, so bit '**d**' is delayed by a further 3 frames to appear in the restored sequence at the output of the dis-interleaver.

The minimum difference between adjacent values in the bit-reversal sequence is 4, so when consecutive bits are subject to transmission errors, after dis-interleaving they are separated by at least 96 ms. Within this 'depth' of time interleaving, consecutive bits are affected by fading or interference events which are uncorrelated in the time domain even at very low vehicle speeds. The choice of interleaving over a maximum of 16 frames appears to have been made pragmatically and this number may be less than optimum but, practically, a greater time-delay (e.g. twice 360 ms) would probably be intolerable.



(source: J. P. Chambers)

Fig. A3.1 - time interleaving and dis-interleaving

A3.2 Frequency interleaving

In this case, the series[24] is rather more complicated than the bit-reversal sequence used for time interleaving but, nevertheless, it is amenable to simple arithmetic. Each of the 1536 bit-pairs to be transmitted is given an index, 0 to 1535, and each of the 1536 carriers is given an index, -768 to 768, omitting 0 which corresponds to the un-modulated centre carrier. The series then relates the bit-pair index to the carrier index, and is constructed as follows.

First, an intermediate series is constructed, starting with 0 as the value of the first term. The value of the next term is calculated by multiplying the value of the previous term by 13 and adding 511, with the proviso that if the result is greater than 2047 then 2048 is repeatedly subtracted from it until the result is less than 2048 but positive (i.e. the result is taken *modulo* 2048). This is repeated to yield 2048 different values; 0, 511, 1010, 1353, ..., 1221.

Then, from the intermediate series, only those values are selected which lie in the range 256 to 1792, omitting all others and 1024 (the centre carrier again). This yields 1536 different values, and 1024 is subtracted from each value giving -513, -14, 329, ..., 197. The position of each value in this series gives the bit-pair index, 0, 1, 2, ..., 1535, and the value gives the carrier index, so the first bit-pair is directed to the carrier indexed -513, and so on.

Consecutive bit-pairs are separated by widely varying amounts, ranging from about 40 carriers (i.e. their information is transmitted using frequencies separated by at least 40 kHz), up to more than 1400 (i.e. 1.4 MHz separation). It is notable that the same series is applied, in the same way, for every frequency-interleaved symbol so some particular patterns of static selective fading must exist which could 'wreak havoc'; one supposes that these will not be encountered frequently in practice!

By this process, the resulting dispersal-in-time of errors in the received, dis-interleaved bit-stream is limited to within one symbol-block, but errors are further dispersed by the time interleaving which is applied 'outside' the frequency interleaving.

[24] The term 'sequence' has been avoided in this case, because all the elements of the 'series' are notionally used at the same time, not consecutively.

APPENDIX 4

RECEIVER SYNCHRONISATION

Many of the processes in the DAB receiver require accurate synchronisation with aspects of the frequency and timing of the incoming signal, and in mobile reception conditions, all of these aspects can be modified by the Doppler shift.

A4.1 Receiver architecture

An example of the architecture of a DAB receiver is outlined in Fig. A4.1. The frequencies given here are examples of what can be used.

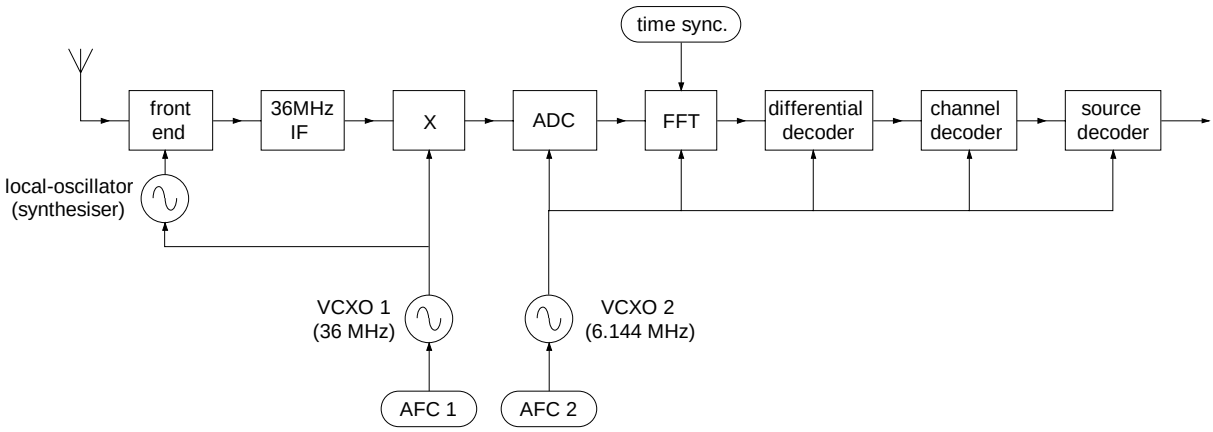


Fig. A4.1 - possible architecture of a DAB receiver

The two reference oscillators are both high-stability voltage-controlled crystal oscillators (VCXOs). VCXO 1 provides the reference frequency for the synthesised local-oscillator and the local-oscillator signal for down-conversion from IF to baseband frequencies. VCXO 2 provides clock signals for the digital processing parts of the receiver. The important point to note is that whilst both reference oscillators need to track Doppler shifts, VCXO 1 may also need to track drifts in oscillators used in the transmitter; the centre-frequency of the transmitted signal is not necessarily related to the frequencies of elements of the DAB signal. Therefore, the means to control the frequencies and phases of these two oscillators need to be independent.

This architecture is representative of the second-generation receiver following an experimental modification to apply Automatic Frequency Control (AFC) to the RF local-oscillator. It may not be fully representative of the current third-generation receiver; insufficient information is openly available to confirm this. The use of VCXO 1 for both frequency conversions is economical but technically not optimum because it implies some change of the intermediate frequency when AFC is applied (the potential change is greatest for low radio frequencies), and this conflicts with the use of a SAW IF filter with steep attenuation slopes. Nevertheless, this example serves to illustrate the important points.

The following aspects of synchronisation need to be considered:

(a) The rate at which FFT computations are carried out must be equal to the reciprocal of the *total* symbol duration of the incoming signal; to avoid inter-carrier crosstalk.

(b) The channel decoding and de-multiplexing process must be clocked at the same rate as data appear in the incoming signal; so the correct data are processed and selected.

(c) The frequencies of components of the baseband signal input to the FFT must be such that the carrier frequencies lie precisely on the correct 'teeth' of the comb with frequency separation equal to the reciprocal of the *active* symbol duration; to avoid corruption of the differential coding and to avoid inter-carrier crosstalk.

(d) The FFT processing window must be timed to start at the point in each total symbol which makes the greatest constructive use of receivable multipath signals.

The first two aspects are related to the frequency of the 'master' clock in the transmitter. The frequency of VCXO 2 must be agile and controlled to satisfy these conditions; AFC 2 can be derived from the symbol rate of the incoming signal. However, the source decoder must be clocked at a relatively constant rate to avoid noticeable pitch changes in the reproduced audio signals so, in practice, the source decoder may be provided with a separate clock signal which follows the average frequency of VCXO 2. In that case, the selected de-multiplexed data must be buffered before source decoding.

Aspect (c) is related to the radio frequency of the incoming signal, and the frequency of the synthesised local-oscillator must be agile and controlled to satisfy this condition. AFC 1 must be derived from the incoming signal in a manner that is essentially independent of AFC 2 and the symbol rate.

Aspect (d) requires analysis of the impulse response of the transmission channel. In practice, the time synchronisation (shown as 'time sync.' in Fig. A4.1) can be controlled by adjusting the phase of VCXO 2. The phase of VCXO 1 is not important because of the use of differential QPSK, as long as it does not change rapidly.

Facilities for AFC and time synchronisation are provided by the 'synchronisation channel' within the DAB signal. At the beginning of each transmission frame, this carries a null symbol followed by a Phase Reference symbol. These are used to control the reference oscillators in the receiver.

A4.2 Initial frequency and time synchronisation using the null symbol

All of the carriers are attenuated by at least 20 dB for the 1.297 ms duration of the null symbol (in Mode 1). This discontinuity can be detected from the envelope of the incoming RF signal, and can be used for coarse synchronisation in much the same way that sync. pulses are used in a television receiver.

The period of these discontinuities is the duration of the transmission frame, 96 ms, which corresponds to 589824 cycles of a 6.144 MHz clock. Initial frequency synchronisation of VCXO 2 can be achieved by counting the number of clock cycles in a frame and comparing the result with this number. Initial phase synchronisation of VCXO 2 can be achieved by timing the first FFT processing window per frame to start 246 μs (the duration of the guard interval) after the end of the null symbol, for example. If the frequency of VCXO 2 is correct, subsequent processing windows will then start 246 μs after the beginning of each symbol; that is, at the beginning of each 'active' symbol period.

The advantage of this simple approach is that it allows the first stage of synchronisation to be achieved rapidly and reliably, using simple circuitry. The disadvantage of using a null symbol is that it introduces a discontinuity into the differential coding of the QPSK modulation. The phase reference for decoding is normally taken from the preceding symbol, and the maximum symbol duration has been chosen for adequate correlation of the channel phase response from one symbol to the next in mobile reception conditions. However, it is likely that adequate correlation would not always be maintained from one symbol to the next-but-one. Therefore, a new phase reference needs to be established before the differential decoding can resume operation, and this implies the sacrifice of one symbol. Wastage is avoided by placing the multi-purpose 'Phase Reference' symbol immediately after the null symbol; despite its name, this is also used for fine time-synchronisation and AFC in the receiver.

A4.3 The Phase Reference symbol

During the Phase Reference symbol, all carriers are transmitted at normal power with the normal symbol duration and guard interval; these are necessary requirements for its use as the phase reference for all carriers.

The carriers are modulated in a fixed pattern and this pattern is reproduced when the incoming signal is analysed by the FFT in the receiver. However, the position of the pattern within the array of numbers output by the FFT depends on the frequencies of the received signal and the RF local-oscillator in the receiver. The objective is to locate the pattern precisely at a particular position so the data conveyed by the carriers during the following frame can be identified uniquely. The same pattern is stored in the receiver, so the frequency error can be determined by measuring the misalignment between the stored and received patterns. The measured error can then be converted into a control voltage to drive the local-oscillator frequency towards the point of minimum error.

The Phase Reference symbol also facilitates measurement of the impulse response of the transmission channel, and fine time synchronisation can be achieved using this; this will be discussed later.

The Phase Reference symbol-block has the normal capacity of 1536 bit-pairs, in Mode 1, and the whole of this capacity is used for synchronisation functions. The data are self-contained within the symbol-block and can be decoded without reference to an earlier symbol. The chosen fixed pattern has properties which assist its recognition in the presence of noise and interference; it is based on a CAZAC (Constant-Amplitude Zero-Autocorrelation) sequence.

A4.3.1 The CAZAC sequence

This is a sequence of complex numbers, which may be called elements, which has the following characteristics:

(a) Each element has the same magnitude; for example, its value can be +1, +j, -1 or -j. This accounts for 'Constant Amplitude'.

(b) If the sequence is multiplied, element by element, by its complex conjugate[25], the sum of the products is a large number; this would be true for many sequences. However, if the sequence is shifted by one element (i.e. the first number takes the place of the second one, etc., and the last number takes the place of the first one), and the shifted sequence is multiplied by the conjugate of the original sequence, the sum of the products is zero. Furthermore, the same zero result occurs for any amount of shift, less than the length of the sequence, in either direction. This means that the sequence has 'Zero Autocorrelation', and this property is relatively uncommon.

'Correlation' is the process of multiplying the element values and accumulating the result, and the 'auto' prefix indicates that the process is carried out on a sequence and the conjugate of the same sequence. An example of a 16 element CAZAC sequence, its conjugate and the autocorrelation with no shift are shown in Fig. A4.2; note that the conjugate has the signs of the 'j's reversed.

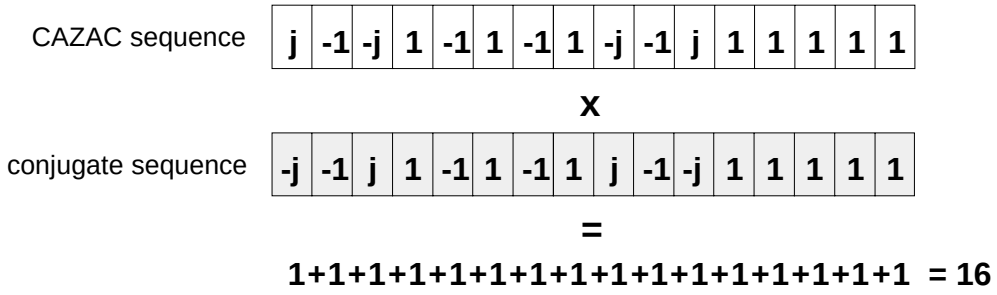


Fig. A4.2 - a 16-element CAZAC sequence, its conjugate, and the autocorrelation with no shift

Performing the autocorrelation with a shift of 1 element (or any other discrete amount up to 15 elements) gives the zero result, as shown in Fig. A4.3. Note that the zero result arises because of cancellation, so it is essential that all elements of the sequence be represented in the autocorrelation calculation; if the sequence was truncated for some reason, the 'ZAC' property would be lost.

[25] The conjugate of a complex number has the sign of the imaginary part reversed, so the conjugate of +j is -j, and vice versa, but the conjugates of +1 and -1 are themselves, because they have no imaginary parts. When complex numbers are multiplied, $j \times j = -1$.

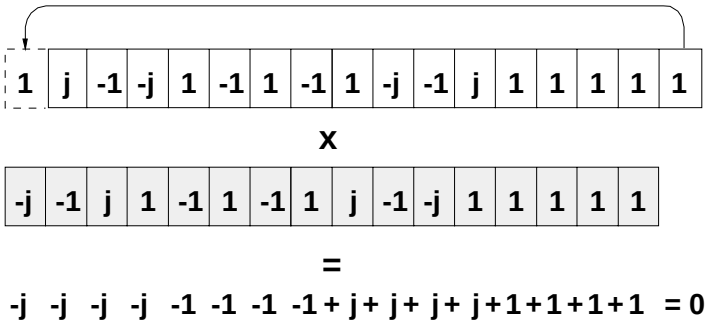


Fig. A4.3 - the autocorrelation with a shift of one element

The result for any discrete amount of shift can be presented as a histogram, as shown in Fig. A4.4. A shift of 16 elements, or any multiple, restores the original sequence so the autocorrelation peak is said to be 'periodic'.

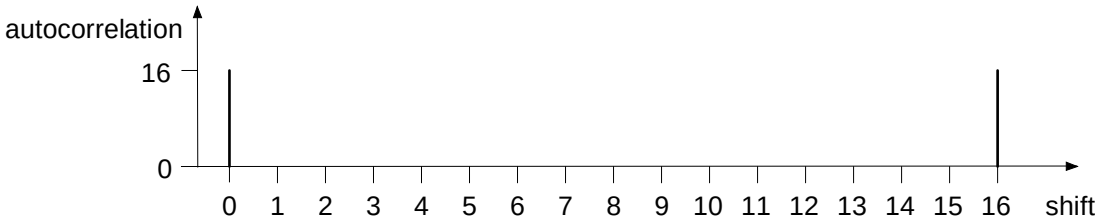


Fig. A4.4 - autocorrelation *versus* shift

Some useful properties of a CAZAC sequence are as follows:

(a) If all elements of the original sequence are multiplied by some constant factor, which may be complex, the magnitude of the result becomes multiplied by the same factor and the zero autocorrelation property is preserved.

(b) If the original sequence is corrupted by the addition of one or more shifted versions of the same sequence, individual correlation peaks are found for each of the sequences at appropriate shifts. Property (a) applies, so if the different sequences are multiplied by constants, the complex values of correlation are multiplied in the same way.

(c) If the original sequence is corrupted by noise or interference, some or all elements will have modified values and, generally, the result will be a complex number. In that case, up to a certain degree of corruption, the correlation[26] for zero shift will still have the greatest magnitude. The ruggedness of such a sequence can be demonstrated by adding a 16-element sequence of random numbers (from the set +1, +j, -1, -j), as shown in Fig. A4.5, overleaf. This corresponds to 0 dB signal-to-noise ratio, but the correlation peak for zero shift can be clearly identified.

[26] The 'auto' prefix will be dropped hereafter because a modified sequence is being considered. Strictly speaking, this should be the *cross*-correlation of the two sequences.

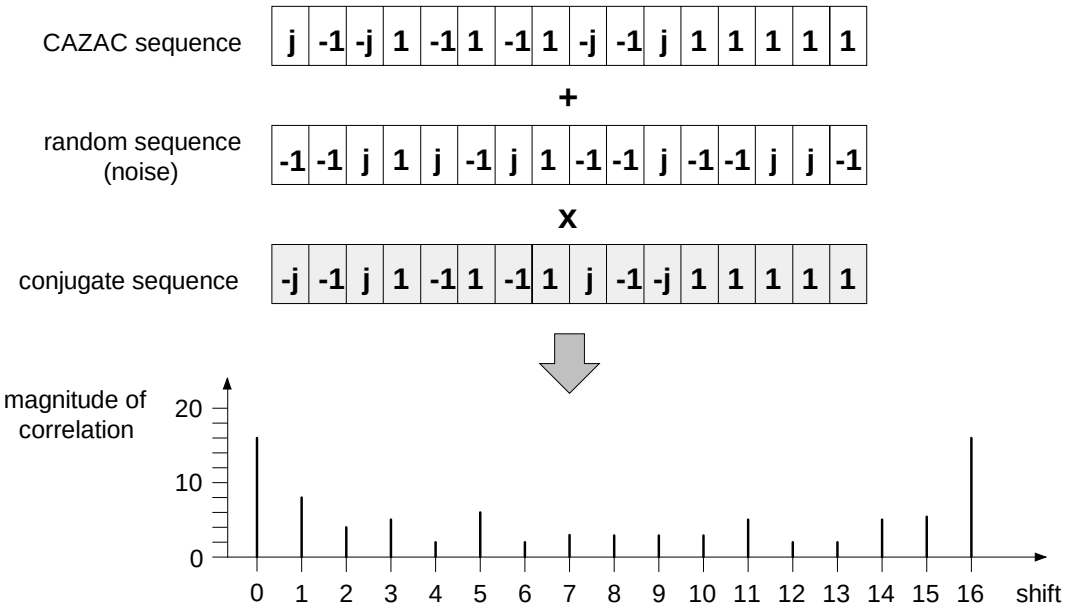


Fig. A4.5 - correlation with added noise

Essentially, the added sequence would need to mimic the original CAZAC sequence (or some multiplied and/or shifted version) for a large false correlation peak to be generated, and the likelihood of this occurring is reduced by lengthening the sequence.

(d) If several of the original CAZAC sequences are concatenated end-to-end and the correlation is performed between any group of 16 consecutive elements and the 16-element conjugate sequence, the result will be periodic correlation peaks. If the amount of shift is limited to less than ±16 elements, then a single central correlation peak can be produced, with reduced correlation either side. Fig. A4.6 shows the sequence extended by 8 elements at each end; in this case, the conjugate sequence can be shifted by up to 8 elements in either direction.

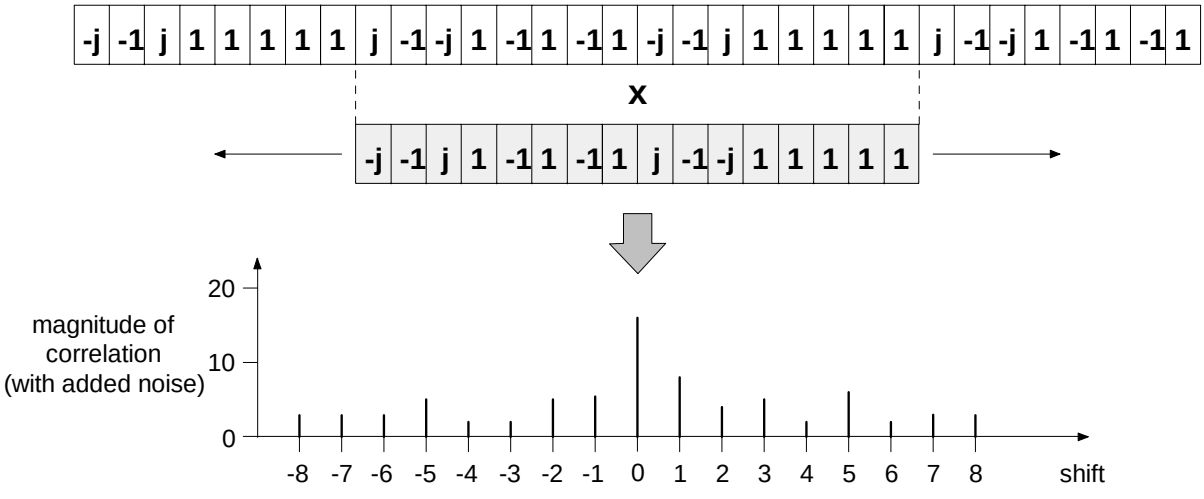


Fig. A4.6 - correlation with an extended sequence

A4.3.2 Coarse AFC

These properties are used to provide AFC in the DAB receiver. A large number of such 32-element extended CAZAC sequences are transmitted simultaneously in the Phase Reference symbol-block. Each element value is conveyed by the phase modulation one carrier, and each sequence is contained in the modulation of 32 consecutive carriers (working from the low-frequency end of the ensemble, for example).

In the receiver, the sequences are reproduced in the array of complex numbers output by the FFT, and the conjugate sequence is held in ROM, so the correlation can be performed as previously described. If the time-domain signal input to the FFT has no frequency error, then the reproduced sequence will be aligned in some particular way with the FFT outputs, but if a large frequency error exists (i.e. one or more times the carrier frequency separation), the reproduced sequence will be shifted with respect to those outputs. This can be detected by shifting the stored conjugate sequence or changing the local-oscillator frequency, and searching for the correlation peak. Coarse AFC can be provided in this way, with a capture range of at least ±8 carrier separations (i.e. ±8 kHz in Mode 1).

The Phase Reference symbol-block is not subjected to frequency-interleaving because this would require additional processing in the receiver and it would confer no advantage in conditions of selective fading; a CAZAC sequence is equally sensitive to corruption of adjacent or non-adjacent elements. Also, time interleaving cannot be used because the AFC needs to be updated from transmission frame to frame in order to optimise the receiver performance in conditions of changing radio frequency (i.e. oscillator drift or changing Doppler shift), and the attendant time-delay could not be tolerated.

Consequently, the ruggedness of this system is dictated by the length of the CAZAC sequences alone. However, very long sequences (i.e. across many carriers) could be corrupted by variations in the phase/frequency response of the transmission channel. The use of a large number of shorter sequences is relatively beneficial because the complex results of individual correlations can be added together as vectors, and the magnitude of the resultant used for AFC; this has the effect of averaging noise, interference and channel variations.

A4.3.3 Fine AFC

Having achieved coarse AFC, the residual frequency error will be less than the carrier frequency separation. One effect of a small frequency error is to cause crosstalk between the FFT outputs so the correlation peak will be dispersed to some degree; the magnitude of the correlation will rise and fall as the conjugate sequence is shifted, and the true peak will lie between two discrete amounts of shift.

Take, for example, four adjacent carriers (labelled ***D***, ***E***, ***F*** and ***G*** to facilitate this explanation) which carry part of one of the CAZAC sequences (elements labelled **P**, **Q**, **R** and **S**). The FFT in the receiver operates like a bank of band-pass filters followed by

demodulators; each filter has a sin *f*/*f* frequency response, *f* being the relative frequency, which is broad but exhibits distinct nulls. With no frequency error, the nulls coincide with the frequencies of the adjacent carriers as illustrated in Fig. A4.7, so there is no crosstalk. Only one of the frequency responses is shown fully in order to preserve clarity.

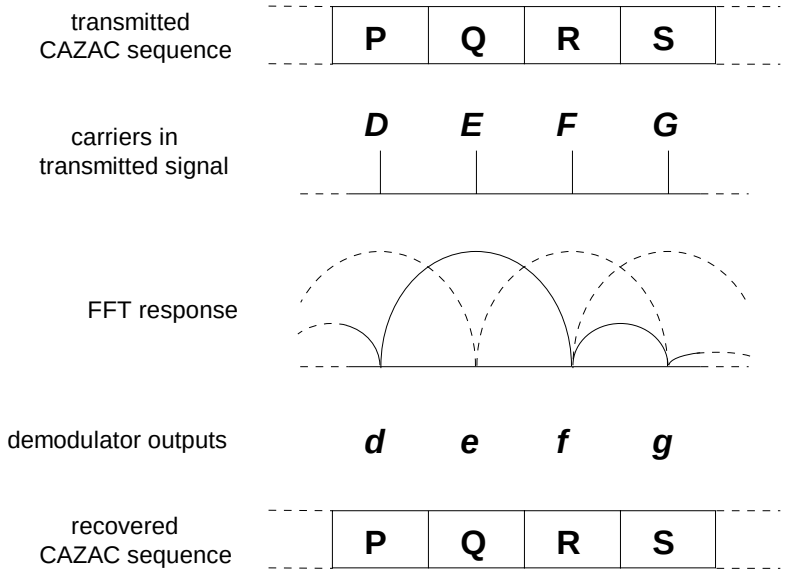


Fig. A4.7 - transmission and reception of four carriers with no frequency error

Each filter/demodulator outputs a complex number (labelled ***d***, ***e***, ***f*** and ***g***) which carries information about the phase of *one* carrier, and the received CAZAC sequence can be recovered from these.

In the presence of a frequency error, the number output by *each* filter contains contributions from *all* of the carriers; that is, crosstalk. The strongest component represents the closest carrier, and the relative amplitudes of the other components depend on the relative frequencies of the carriers they represent (with a sin *f*/*f* variation).

Considering the filter/demodulator which outputs the number ***e*** in the absence of an error; in the presence of an error, its output number will contain a component ***e′*** (the prime ′ indicates some difference from ***e***; principally a smaller amplitude), ***f″*** (with an amplitude even smaller than ***f***), ***d‴***, and many other components with much smaller amplitudes. The next higher-frequency filter/demodulator will output a number containing ***f′***, ***g″***, ***e‴***, etc., and so on for all filter/demodulators. This is illustrated in Fig. A4.8.

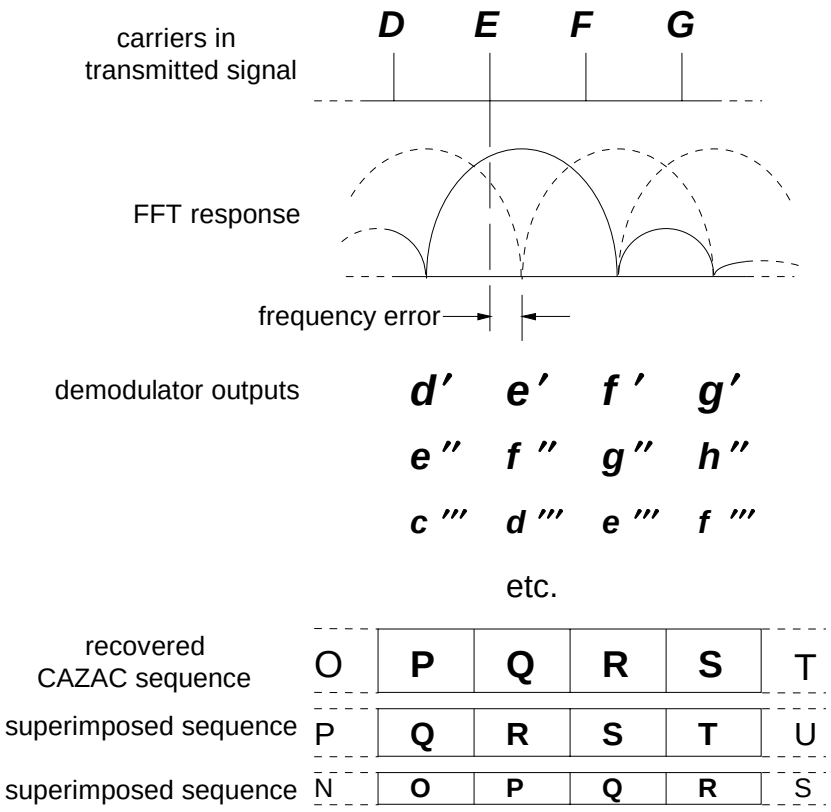


Fig. A4.8 - reception of four carriers with a frequency error

The numbers output by the FFT represent the phasor sums of these components, so the array of numbers contains many shifted versions of the CAZAC sequence superimposed with different amplitudes. This is indicated in Fig. A4.8 by the use of smaller typefaces.

When the correlation is performed, these different amplitudes are preserved in the numerous values of correlation which are found at different discrete shifts, as illustrated in Fig. A4.9.

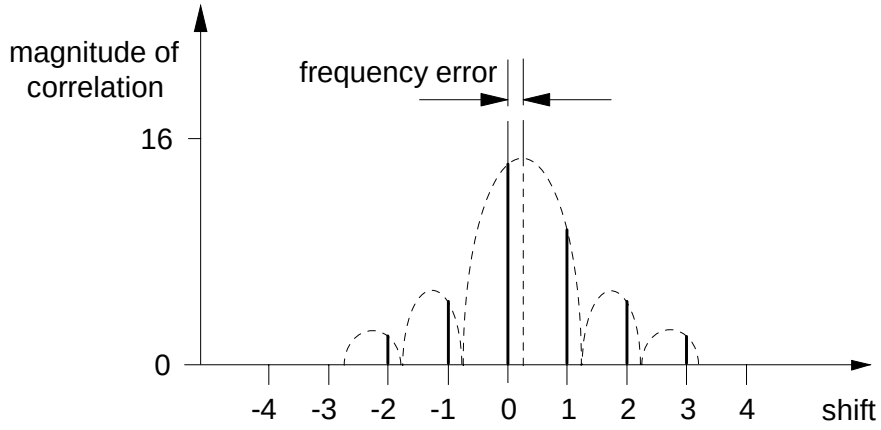


Fig. A4.9 - magnitude of correlation *vs*. shift with a small frequency error

The variation of the *magnitude* of the correlation, at each discrete shift, with increasing frequency error, follows a $(\sin x/x)^2$ law as illustrated by the dashed line in Fig. A4.9; the square is introduced by taking the magnitude. In this case, the variable x is the sum of the magnitude of the error and the amount of shift, multiplied by π.

The amount of shift is an integer, so for a given error, $\sin x$ has the same value for all shifts. Therefore, the magnitude at each shift is proportional to $(1/x)^2$ and individual values are related to one another by a quadratic equation (i.e. of the form $a + bx + cx^2 = 0$). The magnitude of the error can then be calculated from the magnitude of the correlation at different shifts, and the sign of the error can be established with an additional measurement. The middle three values offer the greatest signal-to-noise ratio and, therefore, the greatest accuracy. Some straightforward arithmetic involving these three values provides a number representing the magnitude and sign of the frequency error, and this can be translated by a DAC into a control voltage for the reference oscillator (VCXO 1) which provides the reference frequency for the RF local-oscillator.

A4.3.4 Fine time synchronisation

Having achieved fine frequency control, the Phase Reference symbol-block can then be used to achieve fine time synchronisation of the FFT processing window in the receiver. In order to make the greatest constructive use of multipath signals it is necessary to measure the impulse response of the transmission channel and to time the beginning of the processing window on the basis of a calculation which weights the relative importance of different multipath signals. The impulse response can be measured as follows.

The array of complex numbers output by the FFT (partly illustrated in Fig. A4.7) represents the amplitudes and phases of the carriers transmitted in the Phase Reference symbol-block multiplied by the complex (amplitude and phase) frequency response of the channel, albeit sampled at the carrier frequencies. The (time) impulse response is related to the channel frequency response by the inverse Fourier transform, so a sampled version of the impulse response is related to this sampled frequency response by the inverse DFT, and this can be performed using an inverse FFT; quite *separate* from the 'main' FFT used for OFDM decomposition.

The channel frequency response is revealed by holding the definition of the Phase Reference symbol-block in the receiver as an array of complex numbers in ROM, and by dividing the main FFT outputs by their counterparts in this stored array. The resulting array is then applied to an inverse FFT which outputs a further array representing the impulse response. The magnitudes of elements in this array are then calculated and can be used to derive the timing reference for the following transmission frame. This involves searching for the peaks and applying an algorithm for the chosen timing strategy in order to derive a number representing the amount by which the timing of the main FFT processing window should be advanced or retarded.

For accuracy in this method, the transmitted signal must have a constant Power Spectral Density (PSD) at different frequencies within the bandwidth of the DAB ensemble (i.e. its spectrum must be 'white') for the duration of the Phase Reference symbol. With this provision, the same 'gain' can be applied at all frequencies to avoid distorting the influence

of channel noise. To a first order, and certainly over the period of only one symbol, the PSD is effectively constant because the carriers are all transmitted with the same amplitudes. This might not be the case if the spectrum of the DAB signal were considered over a period of many symbols because it would also depend on the carrier phases.

The Phase Reference symbol-block has been described here as it is defined in the frequency domain, that is, at the input to the inverse FFT in the transmitter or the output of the main FFT in the receiver, and not in terms of the time-domain signal that is transmitted. The time-domain signal is related to this by the Fourier transform. Now, the PSD of a signal in one domain is related by the Fourier transform to the autocorrelation function of the transformed signal in the other domain. In this case, the autocorrelation function of the DAB signal described in the frequency domain is that of the CAZAC sequences, which could be described as an 'impulse'. The Fourier transform of an impulse is a constant value (*viz.* for a true impulse in the time domain, the frequency spectrum is 'white'), so it follows that the PSD of the transmitted DAB signal, described in the time domain, is constant. If the Phase Reference symbol was repeated, the PSD would be constant over the duration of any number of repeated symbols.

Another benefit of the constant amplitude nature of the Phase Reference symbol-block is that the process of division by the stored array in the receiver only has the effect of subtracting phases. The same effect can be achieved by multiplying the array output by the main FFT by a stored array representing the *conjugate* of the Phase Reference symbol-block, and multiplication is an easier process to carry out digitally. Generally, this would introduce a scaling factor, equal for all elements of the array, but in this case the stored values all have unit amplitude so there is no scaling factor.

Described as an array, the Phase Reference symbol-block has 1536 elements in Mode 1, so a 2048-sample inverse FFT must be used. The samples output by the inverse FFT represent the duration of the active symbol, so the resolution of the impulse response is then about 0.5 μs. In the current third-generation receiver, an analogue version of this response is available which can be displayed on an oscilloscope; this is most useful for investigating multipath effects. It is worth noting that absolute magnitudes are not important in this case; indeed they depend on the action of AGC which is entirely separate from the processing of the Phase Reference symbol. An example of an impulse response is illustrated in Fig. A4.10.

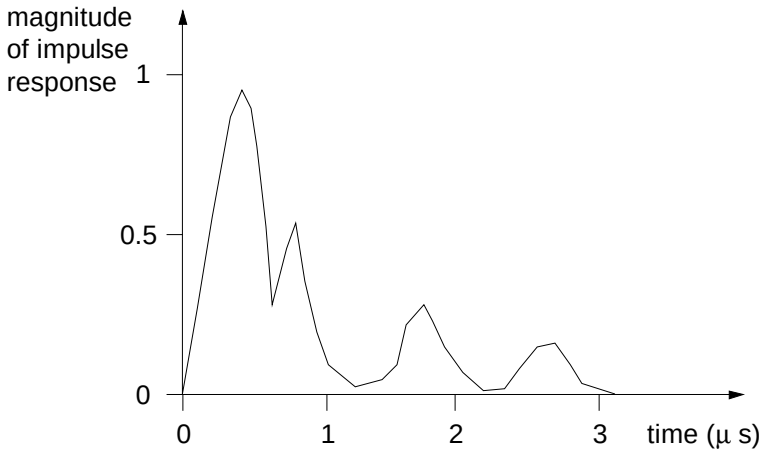


Fig. A4.10 - example of an impulse response for a multipath channel

The required inverse FFT need not be implemented by the same hardware that is used for the forward FFT used for OFDM decomposition. When the time and frequency synchronisation system was being developed (by A. Müller of Daimler Benz), it was found possible to modify a second-generation receiver to accomplish the necessary processing using an existing DSP device which would otherwise have been idle at the beginning of each transmission frame.

A4.3.5 Fine clock-frequency synchronisation

The timing of the FFT processing window depends on the phase of the clock signal provided for the FFT process. It can be varied either by adjusting the phase of the reference oscillator, VCXO 2 (which implies some frequency shift for major adjustments), or by 'slipping' clock cycles, effectively passing the clock signal through a shift register of variable length. The latter approach may be preferable to keep the frequency of VCXO 2 relatively stable.

In either case, if the frequency of VCXO 2 is correct, then the fine time synchronisation function will advance or retard the timing of the FFT window by a small amount each time a Phase Reference symbol is processed in order to compensate for changing multipath conditions. If the frequency is incorrect, the sense of compensation will be consistent from one transmission frame to the next. The average frequency error is proportional to the average phase error accumulated over each frame, so it can be calculated and an appropriate AFC signal derived for VCXO 2. Averaging over several frames corresponds to the function of a low-pass filter.

If the timing of the FFT window is simply related to the phase of VCXO 2, there may be no need for additional processing because the oscillator control circuitry could be configured as a conventional phase-locked loop, able to compensate for frequency as well as phase errors.

The control of VCXO 2 is independent of the AFC applied to the radio-frequency reference oscillator, VCXO 1.

A4.3.6 Enhancements

What has been described so far is the basis of how the time and frequency synchronisation system works, but in practice a straightforward modification is applied which provides a considerable improvement in its performance and ease of implementation.

The elements of the CAZAC sequences are not conveyed by the modulation of individual carriers but are coded differentially in the modulation of adjacent pairs of carriers; each element value is represented by the product of the complex values of one carrier and the conjugate of its (higher frequency) neighbour. The amplitudes of the carriers and the element values are all equal and nominally unity, so each element value is represented by the difference between the phases of two adjacent carriers. This is essentially the same as the method used for the bulk of the DAB data except that, in this case, the differential encoding is applied from carrier to carrier during one symbol, rather than from symbol to symbol for each carrier. The element values can be reproduced in the receiver by differentially decoding the complex numbers output by the FFT; the first element is derived from the numbers which represent the phases of the first and second carriers across the ensemble, and so on.

This modification has two important benefits. Firstly, it makes the operation largely independent of the phase response of the transmission channel; this need only be correlated from one carrier to the next, rather than over groups of 16. Secondly, it offers a simplification in the computation of the fine frequency error by avoiding the square root which is required to solve a quadratic equation. This can be explained as follows.

Take, for example, four adjacent carriers (***D***, ***E***, ***F*** and ***G***, as before) which carry part of one of the CAZAC sequences (elements **Q**, **R** and **S**). The way that the phases of the carriers are related to the sequence elements by the differential coding is illustrated in Fig. A4.11; for example, element **Q** is conveyed by the modulation of carriers ***D*** and ***E***, etc.

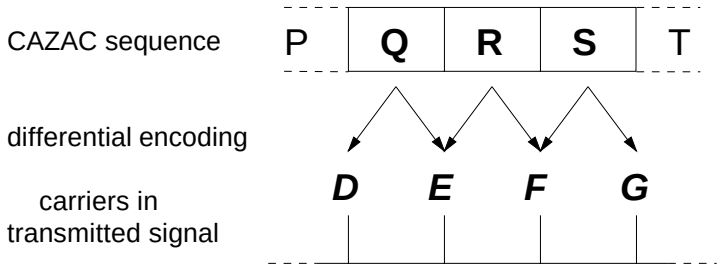


Fig. A4.11 - encoding of four carriers

With no frequency error, each filter/demodulator (of the FFT in the receiver) outputs a complex number which carries information about one carrier, and the received CAZAC sequence can be recovered from these by differential decoding.

In the presence of a small frequency error, the number output by each filter contains contributions from all of the carriers with a sin *f*/*f* distribution of relative amplitudes, as was illustrated in Fig. A4.9. Considering the filter/demodulator which outputs the number ***e*** in the absence of an error; in the presence of an error, its output number will contain components ***e′***, ***f″***, ***d‴***, etc., with progressively decreasing amplitudes. The output number for the next higher-frequency will contain ***f′***, ***g″***, ***e‴***, etc., and so on.

When differential decoding is applied, one number is multiplied by the conjugate of its neighbour so the result contains the difference between their phases and the product of their magnitudes. Subtraction is a linear process (i.e. it does not introduce distortion) so the results will contain many shifted versions of the CAZAC sequence superimposed with different amplitudes, as before; this is illustrated in Fig. A4.12 on the next page. When the correlation is performed, these different amplitudes are preserved in the different values of correlation which are found at different shifts, and the same method could be used to calculate the frequency error as was described in Section A4.3.3.

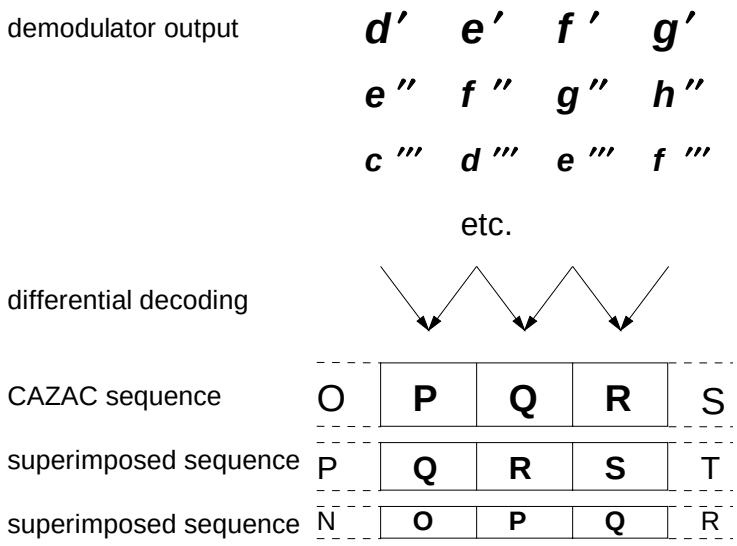


Fig. A4.12 - decoding of four carriers with a frequency error

However, because the number output by each filter/demodulator contains contributions from all of the carriers, it is also possible to reproduce all elements of the CAZAC sequence by differentially decoding within each individual output number. For example, if the number containing components ***e′***, ***f″*** and ***d‴*** is multiplied by its own conjugate, the result contains the squared magnitude of each component and products of each component with the conjugate of another. The product of ***d‴*** and the conjugate of ***e′*** gives element **P** of the sequence with one magnitude, ***e′***and ***f″*** give element **Q** with a different magnitude, and so on. The same principle applies to the next FFT output number, where element **Q** is reproduced with the same magnitude as **P** was in the previous case, and so on as illustrated in Fig. A4.13.

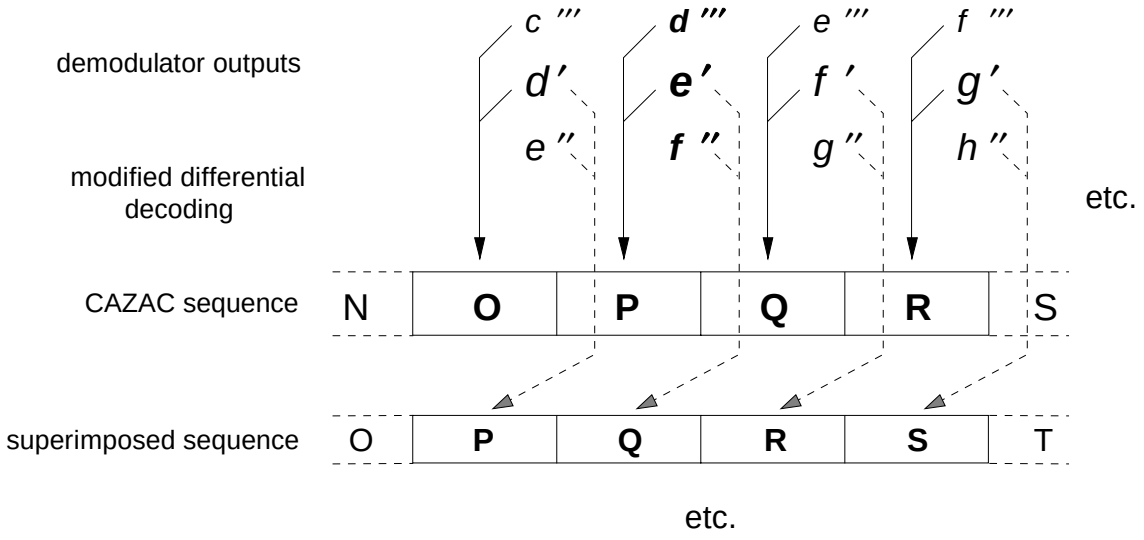


Fig. A4.13 - modified differential decoding

For the sake of clarity, the components of each demodulator output number have been re-ordered and only one superimposed sequence is shown in Fig. A4.13. When an array of such results is built up for all carriers, this 'modified' differential decoding process reveals another set of superimposed, shifted versions of the CAZAC sequence with different magnitudes in this array.

When the correlation is performed between this array and the conjugate sequence, again these different magnitudes are preserved, but in this case the variation of the magnitude of the correlation, at each discrete shift, with increasing frequency error, follows a law which involves products but not a perfect square. For a given error, the relationship between one of these values and one derived by conventional differential decoding is a linear equation, rather than quadratic, and this is easier to solve digitally. The magnitude and sign of the error can be calculated using two pairs of results from conventional and modified decoding.

Differential encoding does not alter the PSD of the Phase Reference symbol-block, so this has no effect on measurement of the channel impulse response.

A4.3.7 Additional considerations

A potential problem is introduced by the use of differential coding. The product of one component and the conjugate of the next one in the alphabet, with this labelling (e.g. $\boldsymbol{d'''}$ and the conjugate of $\boldsymbol{e'}$), should produce one element of the CAZAC sequence (i.e. **P**) and, ideally, all other combinations (e.g. $\boldsymbol{d'''}$ and $\boldsymbol{f''}$) should produce zero; but they do not. In fact, no 16-element CAZAC sequence has this property, and all the possible sequences produce a predictable 'residue' (i.e. false correlation) when differentially encoded and decoded, and then correlated. This effect would reduce the signal-to-noise ratio of the AFC control voltage, but there is a way to counteract it.

If all elements of a CAZAC sequence are multiplied by a constant (+1, +j, -1 or -j) and the correlation is performed with the conjugate sequence which has been multiplied by the same constant, the result is independent of the value of the constant. This is not affected by differential encoding and decoding of the sequence from one element to the next (i.e. with an 'offset' of 1). However, if the correlation is performed between the same multiplied conjugate sequence and the sequence produced by differential decoding with any other offset, the results have values which are dependent on the constant (i.e. the unwanted 'cross terms' like $\boldsymbol{d'''}$ and $\boldsymbol{f''}$ have values which depend on the constant). The dependency is different for different offsets, but if four identical CAZAC sequences are each multiplied by a different value of the constant (i.e. one each by +1, +j, -1, and -j) and the complex sum of the four values of correlation is taken for each shift, all significant components of the residue are cancelled. Additionally, if these four sequences are subject to similar channel disturbances (e.g. noise or interference), then processing in this way can also reduce the impact of the disturbances; this is assisted by transmitting the four sequences on adjacent frequencies.

The 1536 carriers in Mode 1 allow the transmission of 48 differentially encoded, 32-element extended CAZAC sequences in the one symbol-block, and in practice they are all derived from the same 16-element 'kernel' CAZAC sequence. After extension to 32 elements, the sequences are arranged in groups of four across the ensemble where each of the four is

multiplied by a different constant, +1, +j, -1 and -j, respectively; there are 12 such groups. The multiplied sequences are then differentially encoded from element to element and presented to the inverse FFT for transmission by OFDM.

There is an obvious hitch in this description; the use of differential coding implies that 33 carriers are needed to represent 32 sequence elements. If the absolute phase of one carrier in a group is chosen, the phases of the remaining 31 are determined as well as that of one other, to one side of the group. However, the absolute phases of the carriers in any group of 32 are not important to the processing of the CAZAC sequences, only the relative phases of adjacent carriers, and the relative phases from one group to another are unimportant. Therefore, the phase of the lowest-frequency carrier in each group (for example) can be chosen at will, and this could be used to provide continuity in the differential coding from one group to the next; the 32nd element being defined by the phases of the 32nd carrier in the corresponding group and the first carrier in the subsequent group. In that case, only the highest-frequency group would be left with an un-defined 32nd element.

In practice, the lengths of the extended sequences are reduced to 31 elements, making the groups completely independent. The drawback is a small reduction in the AFC capture range, to at least ±7 times the carrier separation. This facility could be used for additional signalling; in Mode 1, it provides 48 2-bit values every 96 ms, equivalent to 1 kbit/s. However, in practice, the phases are chosen to minimise the peak-to-mean ratio of the transmitted DAB signal during the Phase Reference symbol.

In the receiver, conventional and modified differential decoding can be applied to the 48 sequences effectively in parallel, and the 48 recovered sequences for each case can be correlated with appropriately multiplied versions of the stored conjugate sequence. The complex results of these correlations can then be summed, and the magnitudes of these sums used to derive the fine frequency error. The effects of distributed disturbances, such as noise or selective fading, are averaged by summing the results in this way.

A4.3.8 Other Modes

In Transmission Modes 2 and 3, the number of carriers is reduced to 384 and 192, respectively. Smaller numbers of the same 31-element extended sequences are used, 12 in Mode 2 and 6 in Mode 3. This reduces the scope for averaging, but the demands for fine frequency control are lesser in these modes because the carrier frequency separations are correspondingly greater. Note that in Mode 3, cancellation of the differential coding residue is incomplete.

The greater separation of the carrier frequencies in the higher modes means that the AFC capture range is correspondingly greater; at least ±28 kHz in Mode 2, and at least ±56 kHz in Mode 3. In terms of fractional bandwidth, it is kept relatively constant.

Throughout this discussion, values of AFC capture range have been preceded by the term ‘at least’. This is because one of the possible methods for initial acquisition is to change the local-oscillator frequency, for example by impressing a ramp waveform on the control voltage fed to VCXO 1. In that case, the capture range could extend well beyond the fundamental range provided by the extended CAZAC sequence.

Inverse FFTs with smaller numbers of samples must be used in the measurement of the impulse response so the results will contain fewer samples, but the active symbol durations are correspondingly shorter (250 μs in Mode 2, and 125 μs in Mode 3) so the absolute resolution is constant; about 0.5 μs.

A4.3.9 Caveat

Of course, a receiver manufacturer may choose to simplify some parts of the time and frequency synchronisation processing in order to economise on processing power. A smaller number of the transmitted sequences could be used for the AFC function, and an inverse FFT with less samples could be used to derive the impulse response, yielding less resolution. Ideally, the processing should be carried out as each Phase Reference symbol arrives so the control signals for the reference oscillators can be updated with the least delay, but a manufacturer could choose to distribute the processing over the following transmission frame. Such changes would be likely to impair the receiver performance in some changing multipath conditions.