|  |  |
| --- | --- |
| 3GPP TR 26.933 V0.4.0 (2024-01) | |
| Technical Report | |
| 3rd Generation Partnership Project;  Technical Specification Group SA;  Study on Diverse Audio Capturing system (Release 19) | |
|  | |
|  |  |
|  | |
| The present document has been developed within the 3rd Generation Partnership Project (3GPP TM) and may be further elaborated for the purposes of 3GPP. The present document has not been subject to any approval process by the 3GPPOrganizational Partners and shall not be implemented. This Specification is provided for future development work within 3GPPonly. The Organizational Partners accept no liability for any use of this Specification. Specifications and Reports for implementation of the 3GPP TM system should be obtained via the 3GPP Organizational Partners' Publications Offices. | |

|  |
| --- |
|  |
| ***3GPP***  Postal address  3GPP support office address  650 Route des Lucioles - Sophia Antipolis  Valbonne - FRANCE  Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16  Internet  http://www.3gpp.org |
| ***Copyright Notification***  No part may be reproduced except as authorized by written permission. The copyright and the foregoing restriction extend to reproduction in all media.  © 2022, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).  All rights reserved.  UMTS™ is a Trade Mark of ETSI registered for the benefit of its members  3GPP™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners LTE™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners  GSM® and the GSM logo are registered and owned by the GSM Association |

Contents

Foreword 7

Introduction 8

1 Scope 9

2 References 9

3 Definitions of terms, symbols and abbreviations 10

3.1 Terms 10

3.2 Symbols 10

3.3 Abbreviations 10

4 Factors of different UE categories related to audio capture 10

4.1 Mobile phones 10

4.1.1 Structure Size 10

4.1.1.1 Length 11

4.1.1.2 Width 11

4.1.1.3 Depth 12

4.1.1.4 Summary 12

4.2 Earbuds 13

4.3 Tablets 13

4.4 Laptops 14

4.5 Watches 17

4.6 AR glasses 18

4.7 Cars 18

5 Components used in audio capture 19

5.1 Component 19

5.1.1 Microphone 19

5.1.1.1 Dynamic microphone 20

5.1.1.2 Condenser microphone 20

5.1.1.3 Micro-Electro-Mechanical Systems microphone 20

5.1.1.4 Contact microphone 20

5.1.1.5 Other microphones 20

5.1.1.6 Summary 20

5.2 Preamps 21

5.3 ADC 21

5.4 Clock 21

5.5 Directivity 21

5.5.1 Traditional approaches used in immersive audio 21

5.5.1.1 Directional microphone capsule 21

5.5.1.2 Interference tube 21

5.5.1.3 Binaural acoustic stimulation 21

5.5.2 Beamforming microphone array 22

5.5.2.1 Delay-sum microphone array 22

5.5.2.2 Differential microphone array 23

6 Acoustic design 24

6.1 Stereo capture 24

6.1.1. Principle of stereo signal representation 24

6.1.2. Characteristic of stereo capture 24

6.1.3. Factors that affect stereo capture 24

6.1.4. Stereo microphone configurations 24

6.1.4.1. Near-Coincident 24

6.1.4.1.1. ORTF 25

6.1.4.2. Baffled 25

6.1.4.2.1. OSS (Optimal Stereo System) 25

6.1.4.3. Coincident 26

6.1.4.3.1. X/Y 26

6.1.4.3.2. Blumlein 27

6.1.4.3.3. M/S 27

6.1.4.4. Spaced 28

6.2. Spatial audio capture 29

6.2.1. Binaural capture 29

6.2.1.1. Principle of binaural signal representation 29

6.2.1.2. Possible issues in binaural capture 29

6.2.1.3. Factors that affect binaural capture 29

6.2.1.4. Differences between binaural and stereo audio 29

6.2.2. Parametric spatial audio capture 30

6.2.2.1. Principle of parametric spatial audio representation 30

6.2.2.2. Characteristics of parametric spatial audio capture and representation 31

6.2.2.3. Factors that affect parametric spatial audio capture 31

6.2.2.4. Multi-microphone configurations in parametric spatial audio 31

6.2.3. Non-parametric spatial audio capture 31

6.2.3.4.1. Immersive audio ORTF configuration[6] 32

6.2.3.4.1.1 ORTF-surround 32

6.2.3.4.1.2 ORTF-3D 33

6.2.3.4.2. Immersive audio M/S configuration 34

6.2.3.4.2.1 Double-M/S[7] 34

6.2.3.4.2.2 M/S-3D[8][9] 34

6.2.3.4.3. IRT-cross 35

7 Signal processing 36

7.1 AEC 36

7.1.1. Principle of mono audio AEC 36

7.1.2 Challenges for immersive audio AEC 37

7.1.3 The current status of the research 37

7.1.4 AEC for XXX UE 37

8 Example audio capture processing solutions 37

8.1 Capture Scenario: Telephony Communications 38

9 Conclusions and Recommendations 39

A.1 Mobile phone size 39

A.2 Earbud size 40

A.3Tablet size 41

A.4 Laptop size 42

A.5 Watch size 42

A.6 AR glass size 43

A.7 Car exterior size 43

A.8 Car exterior and interior size 44

] 45

Annex <X> (informative): Change history 46

# Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

x the first digit:

1 presented to TSG for information;

2 presented to TSG for approval;

3 or greater indicates TSG approved document under change control.

y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.

z the third digit is incremented when editorial only changes have been incorporated in the document.

In the present document, modal verbs have the following meanings:

**shall** indicates a mandatory requirement to do something

**shall not** indicates an interdiction (prohibition) to do something

The constructions "shall" and "shall not" are confined to the context of normative provisions, and do not appear in Technical Reports.

The constructions "must" and "must not" are not used as substitutes for "shall" and "shall not". Their use is avoided insofar as possible, and they are not used in a normative context except in a direct citation from an external, referenced, non-3GPP document, or so as to maintain continuity of style when extending or modifying the provisions of such a referenced document.

**should** indicates a recommendation to do something

**should not** indicates a recommendation not to do something

**may** indicates permission to do something

**need not** indicates permission not to do something

The construction "may not" is ambiguous and is not used in normative elements. The unambiguous constructions "might not" or "shall not" are used instead, depending upon the meaning intended.

**can** indicates that something is possible

**cannot** indicates that something is impossible

The constructions "can" and "cannot" are not substitutes for "may" and "need not".

**will** indicates that something is certain or expected to happen as a result of action taken by an agency the behaviour of which is outside the scope of the present document

**will not** indicates that something is certain or expected not to happen as a result of action taken by an agency the behaviour of which is outside the scope of the present document

**might** indicates a likelihood that something will happen as a result of action taken by some agency the behaviour of which is outside the scope of the present document

**might not** indicates a likelihood that something will not happen as a result of action taken by some agency the behaviour of which is outside the scope of the present document

In addition:

**is** (or any other verb in the indicative mood) indicates a statement of fact

**is not** (or any other negative verb in the indicative mood) indicates a statement of fact

The constructions "is" and "is not" do not indicate requirements.

# Introduction

Providing immersive voice and audio services by end-user devices is becoming more and more practicable with the development of 4G/5G technologies. Related requirements have been investigated in 3GPP TR 22.891. Several use cases for VR are envisioned in TR 26.918, and for these cases the corresponding audio capturing system are generally considered. As such, capturing capability is crucial for making truly immersive voice and audio experiences.

Due to physical constraints on their outline shapes and sizes, the end-user devices are usually configured with different numbers of microphones and also different microphone setup configurations, hence different audio capturing capabilities are expected. Based on this, the present document gives diverse audio capturing system.

# 1 Scope

This document addresses audio capturing configurations for end-user devices, which is to make the devices to have audio capturing capability in order to provide truly immersive voice and audio service.

This document aims to study the following aspects:

1) Factors of different UE categories related to audio capture.

2) Components used in audio capture.

3) Acoustic design for audio capture.

4) Signal processing, e.g., microphone array beamforming processing, AEC processing etc.

5) Example of audio capture processing solutions.

Editor’s Note: the scope is for further detailed based on the objectives and input contribution.

# 2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non‑specific.

- For a specific reference, subsequent revisions do not apply.

- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[2] 3GPP TR 26.891: "5G enhanced mobile broadband; Media distribution".

[3] 3GPP TR 26.918: "Virtual Reality (VR) media services over 3GPP".

[4] 3GPP TS 26.119: "Media Capabilities for Augmented Reality".

[5] 3GPP TS 26.258: "Codec for Immersive Voice and Audio Services (IVAS); C code (floating point)"

[6] H. Wittek and G. Theile, “Development and application of a stereophonic multichannel recording technique for 3D Audio and VR,” in AES Convention 143, New York, 2017.

[7] Wittek, Haut, Keinath: “Double M/S – a Surround recording technique put to test”, 24. Tonmeistertagung 2006

[8] P. Geluso. “Capturing Height: The Addition of Z Microphones to Stereo and Surround Microphone Arrays,” presented at the 132nd Convention of the Audio Engineering Society (2012 Apr.), convention paper 8595.

[9] Fischer, C., Zingler, D., Medina Victoria, J.: “MS-3D: Extending the Double-MS Array for 3D-Audio applications” (29th Tonmeistertagung – VDT International Convention, November 2016)

[10] J. Benesty, D. R. Morgan and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," in IEEE Transactions on Speech and Audio Processing, vol. 6, no. 2, pp. 156-165, March 1998.

[11] A.W. H. Khong, J. Benesty and P. A. Naylor, "Stereophonic acoustic echo cancellation: analysis of the misalignment in the frequency domain," in IEEE Signal Processing Letters, vol. 13, no. 1, pp. 33-36, Jan. 2006.

[12] 3GPP TS 26.250: "Codec for Immersive Voice and Audio Services (IVAS); General overview

[...] ……

# 3 Definitions of terms, symbols and abbreviations

This clause and its three subclauses are mandatory. The contents shall be shown as "void" if the TS/TR does not define any terms, symbols, or abbreviations.

## 3.1 Terms

For the purposes of the present document, the terms given in 3GPP TR 21.905 [1] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in 3GPP TR 21.905 [1].

Definition format (Normal)

**<defined term>:** <definition>.

**example:** text used to clarify abstract rules by applying them literally.

## 3.2 Symbols

For the purposes of the present document, the following symbols apply:

Symbol format (EW)

<symbol> <Explanation>

## 3.3 Abbreviations

For the purposes of the present document, the abbreviations given in 3GPP TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in 3GPP TR 21.905 [1].

VR Virtual Reality

[

# 4 Factors of different UE categories related to audio capture

Editor’s Note:

* *Collect relevant information on potential UEs like smartphone, headphone, XR glasses, etc.*
* *The shape, structure of UE.*
* *Available computer power according to current device and tendency.*

## 4.1 Mobile phones

### 4.1.1 Structure Size

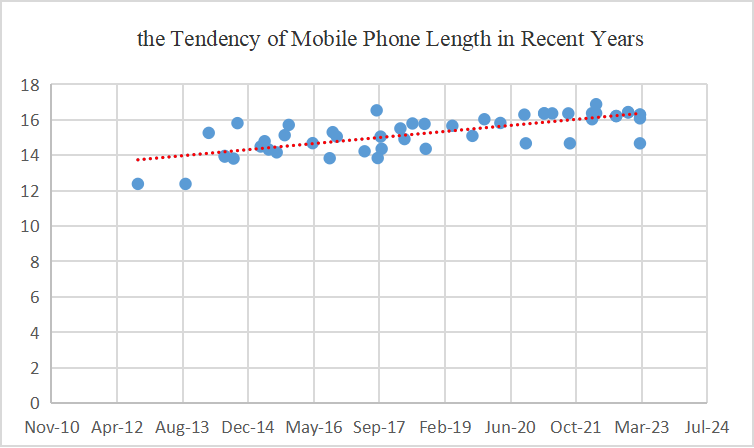
Since 2012, the structure size of mobile phones has been rising in length, width, and thickness, indicating a continuous increase in mobile phone size. This may be in response to the strong demand from consumers for multimedia and gaming functions on their phones, as well as the increasing requirements for microphone and camera quantity and battery consumption. The evolution trend of mobile phones is towards full-screen, which provides a market foundation for the increase in screen size.

The detailed data of structure size are listed in Annex A.1

#### 4.1.1.1 Length

The length of mobile phones has gradually increased from 12.38cm in 2012 to 16.88cm in 2022, with an average length of 15.26cm, according to the investigation, showing an upward trend. With the development of mobile phone models, some phones are no longer limited to the 16:9 aspect ratio, e.g., there are now styles with 18.5:9 and 19.5:9. Although high aspect ratio screens can display more information, most video contents are still in the traditional 16:9 format, so too high an aspect ratio is not conducive to the video display.

The tendency of mobile phone length in recent years is shown in Figure 4.1.1.1-1

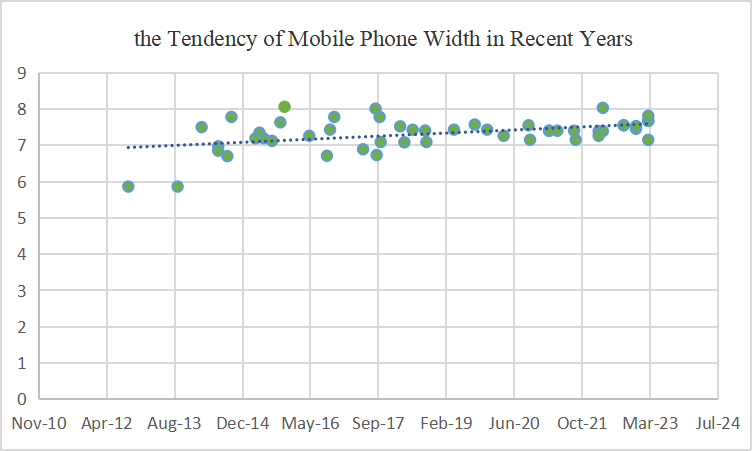


**Figure 4.1.1.1-1 The tendency of mobile phone length**

#### 4.1.1.2 Width

According to the mobile phone data surveyed, the width of mobile phones was around 5.86cm in 2012, while in 2022, the width of mobile phones was changed to around 7.55cm, with a maximum value of 8.06cm. In recent years, the average width of mobile phones has been 7.32cm, and an increment in the length of mobile phones generally follows an increment in width. This is a reasonable evolution tendency for the purpose of function requirements and appearance

The tendency of mobile phone width in recent years is shown in Figure 4.1.1.2-1

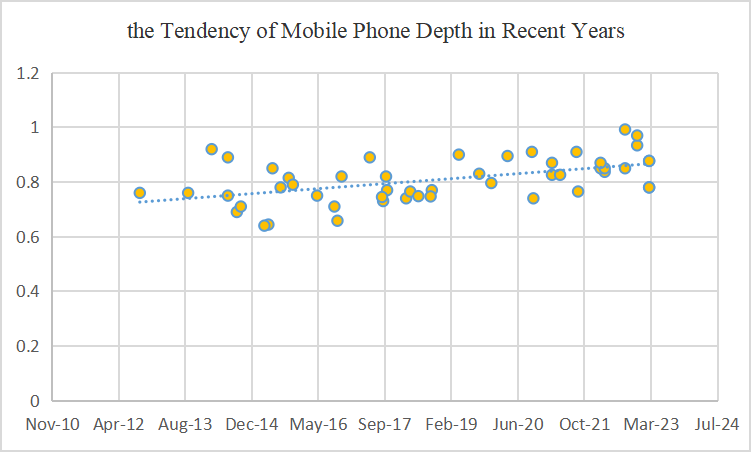


**Figure 4.1.1.2-1 The tendency of mobile phone width**

#### 4.1.1.3 Depth

Among the phones investigated, the thinnest one measures 0.64cm, the thickest one measures 0.992cm, and the average thickness is 0.81cm.

The tendency of mobile phone width in recent years is shown in Figure 4.1.1.3-1



**Figure 4.1.1.3-1 The tendency of mobile phone depth**

#### 4.1.1.4 Summary

According to the investigations, the summary is as follows:

* The maximum values of length, width, and height are 168.78mm,80.6mm and 9.92mm.
* The minimum values are 123.8mm,58.6mm, and 6.4mm.
* The average values are 152.65mm,73.17mm and 8.08mm.
* The 95% Confidence Interval (CI) are (149.60 mm,155.69 mm), (71.92 mm,74.42 mm) and (7.85 mm,8.31 mm).

### 4.1.2 Multi-Microphones Capture on mobile phone

Modern smartphone devices come with inbuilt MEMS microphones supporting more than one microphone, while the top end devices have inbuilt microphones between 3 to 4; mid-range devices support dual microphones. With multi-microphones as a feature, they offer several advantages over devices with mono including improved audio quality, better spatial awareness, more accurate noise cancellation, better sound localization.

The spacing between the multi-array microphones, its polar pattern and the number of microphones varies from device to device. However, generally the spacing ranges from 3cm – 17cm. The following table covers the specifications of these microphones in smartphones.

Table 1: MEMS Microphone example characteristics for Multi-array microphones in Smartphone.

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of Microphones** | **Placement of Microphones** | **Spacing** | **Polar Pattern** |
| Dual Microphone  array | Top and Bottom of Bezel | 12 to 17 cm | Omni-directional or Cardioid |
| Bottom of Bezel and rear facing | 10 to 15 cm |
| Triple Microphone array | Top, Bottom of Bezel and Rear Facing | 5 to 17 cm | Omni-direction, Cardioid, directional |
| Quad Microphone array | Top, Dual Bottom of Bezel and Rear Facing | 5 to 17 cm | Omni-direction, Cardioid, directional |

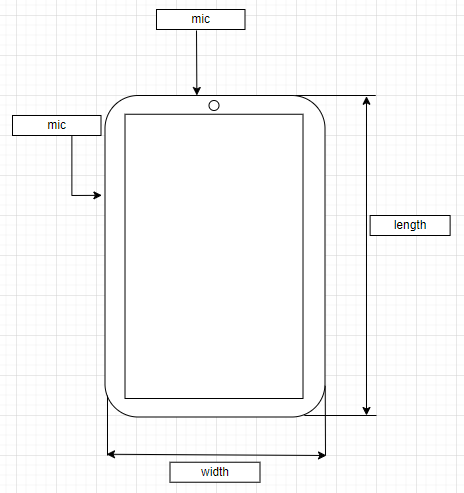
Microphones with 3cm stereo spacing are usually located at the bottom of the bezel (left and right side of USB C-Port), while the larger spacing is achieved with a combination of top, bottom bezel microphone and/or rear facing microphone. Because the user can operate and hold the device in a variety of ways, a combination of these microphones can be activated for stereo voice communication and efficient capturing configuration is possible with advancements in audio processing by means of fine-tune of audio signals through pre-processing which offers great potential for enhancing stereo performance for different user end consumption scenarios such as loudspeaker and/or headphones.

## 4.2 Earbuds

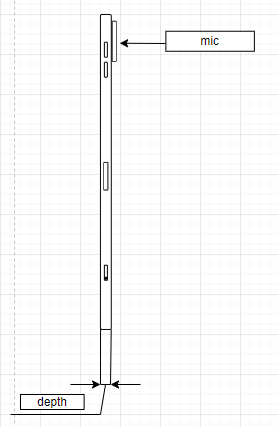
Nowadays, the mainstream earbuds are TWS headphones. Therefore, we investigated some TWS Earbuds. Specifically, the length of all these TWS earbuds measures less than 4cm. Additionally, the depth and width of these earbuds are around 2cm.

## 4.3 Tablets

Tablets, as popular UE, equipped with speakers and microphones. We list 21 devices from 7 brands, it is evident that the size of tablets is significantly larger than that of mobile phones, yet they share a very similar shape. Most of them exceed a height of 20 cm. The minimum length also reaches 19.54 cm. The width is close to 20 cm. The models in the statistics range from 13.48 to 21.49 cm.



**Figure 4.3-1 Tablet front view**



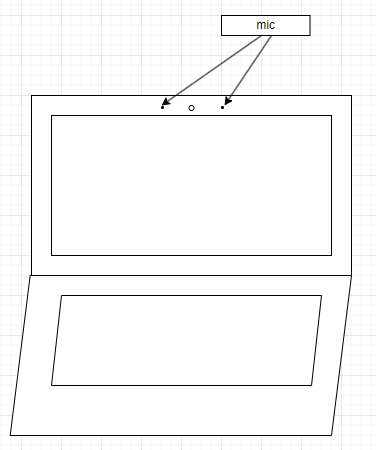
**Figure 4.3-2 Tablet side view**

In addition, the microphone design of tablets now is also very similar to that of mobile phones.

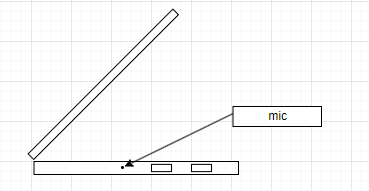
## 4.4 Laptops

Compared to tablets, the primary distinction in laptops is their clamshell structure, featuring a hinged screen and an attached keyboard. Consequently, in current devices, there are more options for microphone placement. Some are positioned on the screen part, while others are located on the keyboard section.

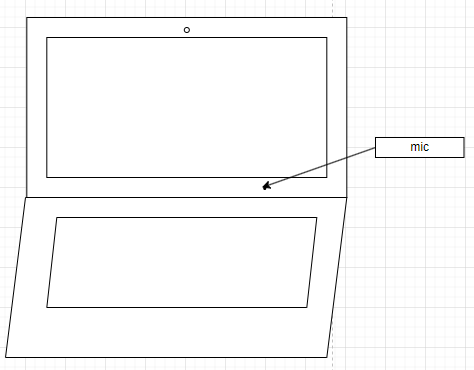
Some popular microphone placement on laptop is explained in the following figure:



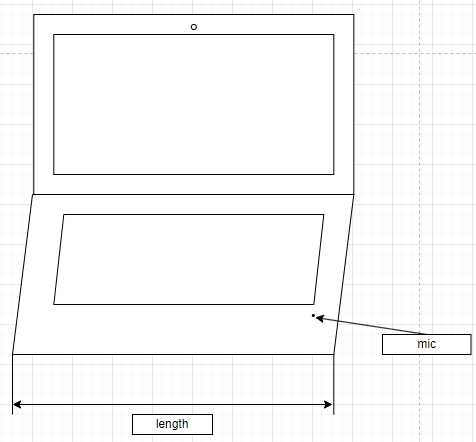
**Figure 4.4-1 Laptop microphone placement 1**



**Figure 4.4-2 Laptop microphone placement 2**



**Figure 4.4-3** **Laptop microphone placement 3**



**Figure 4.4-4 Laptop microphone placement 4**

Also, most laptops have larger structure size that of mobile phones. The range in length goes from approximately 28 cm to around 41 cm, resulting in a range of about 13 cm. The widths span from about 18.5 cm to approximately 32 cm. The height range extends from a few tenths of a centimetre (around 0.7) to over 2.6 cm.

图示

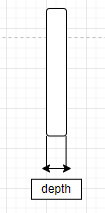
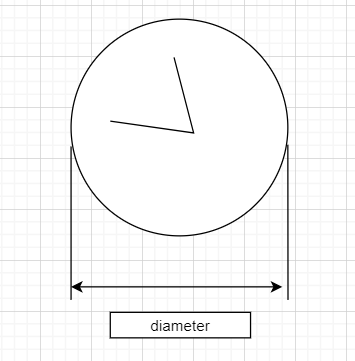
描述已自动生成

**Figure 4.4-5 Laptop side view**

## 4.5 Watches

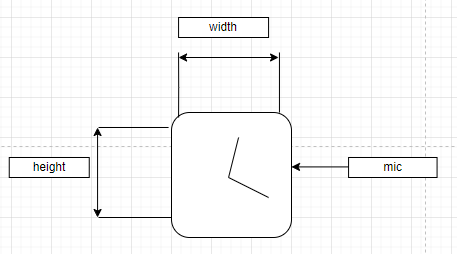
Many watches nowadays also have calling capabilities, but most of them are equipped with only one microphone. Nowadays, the smartwatches capable of making calls mainly come in two shapes: circular and rectangle.

For circular type the diameter is around 4.7 cm. And the depth is range from 10.9 to 13 cm



**Figure 4.5-1 circular watch**

For rectangle type, the lengths vary from 4 cm to 5.7 cm. The widths range from 3.4 cm to 4.57 cm. Heights range from 1.07 cm to 1.49 cm.



**Figure 4.5-2 rectangle watch** **front view**

门上的瓷砖

描述已自动生成

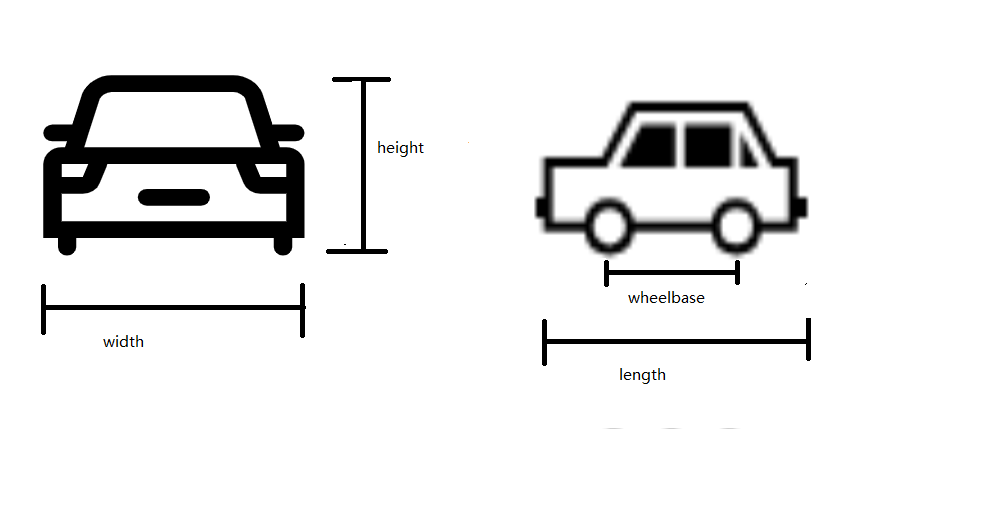
**Figure 4.5-3 rectangle watch** **side view**

## 4.6 AR glasses

Presently, AR glasses come in a wide range of sizes. Some are substantial and weighty, akin to headband-style headsets, while others are lightweight, resembling ordinary glasses. The larger variants often adopt a headband-style design, whereas the smaller ones are designed to the appearance of regular eyewear, Therefore, the average size of the investigated model is approximately 157.48 cm\*130.92cm\*41.22cm on width, length, and height respectively.

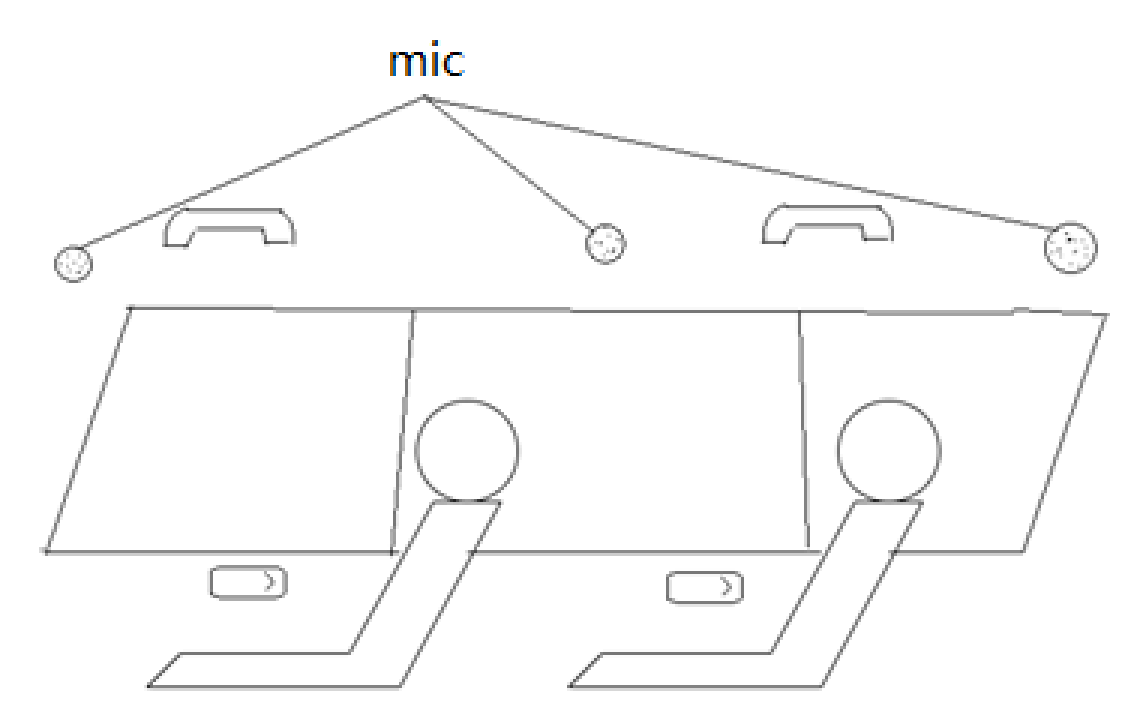
## 4.7 Cars

We have also investigated the dimensions of some mainstream civilian cars. The lengths vary from 445.8cm to 532cm. Widths range from 180.6cm to 208.9cm. Heights fluctuate between 144.2cm and 180.0cm.



**Figure 4.7-1 car front and side view**

Due to the demand for voice services, modern cars are increasingly equipped with more microphones, especially electric vehicles. One of the popular microphone placements is to place microphone arrays on one side of the car roof. There are also other configurations, including centralized placement and distributed placement.

****

**Figure 4.7-2 microphone placement in car**

Some models with exterior and interior size is listed inAnnex A.7and A.8.

Editor’s Note: this is basis for further work

# 5 Components used in audio capture

Editor’s Note:

* Documentation of components may be used in diverse audio capture.
* Relevant components like microphone, AD converter, etc.

## 5.1 Component

### 5.1.1 Microphone

The function of microphone is to convert sound pressure signal to analog electrical signal in circuit.

4 types of microphones popular in the market are described in this proposal. These microphones have unique advantages in UE's immersive audio. They are classified to dynamic microphone, condenser microphone, Micro-Electro-Mechanical Systems (MEMS), contact microphone.

#### 5.1.1.1 Dynamic microphone

Dynamic microphone is one of the popular microphones on market. The most advantage of dynamic microphone for UE is it doesn’t need for external power; the entire recording system will be easier. Another advantage is durability, make it more suitable for loud and high-pressure situation. But it usually has a disadvantage that it is less sensitive to high frequencies.

Dynamic microphone uses a small movable induction coil, which positioned in the magnetic field and is attached to the diaphragm. The current signal generates when the movement of the diaphragm causes the coil to also move within a magnetic field.

#### 5.1.1.2 Condenser microphone

Condenser microphone is another popular microphone on market, especially for immersive audio. Most immersive system is using condenser microphones, like ambisonic microphone and external stereo microphone for mobile phone. It’s popular for its high sensitivity, wide frequency response, low noise. However, the condenser microphone requires a power source, and in the case of most professional condenser microphones, it specifically requires 48V phantom power. Meeting this requirement can be challenging for UE device consider the channel number of immersive audio.

Condenser microphone uses capacitor to convert sound waves to electrical signal. The capacitor consists of two plates, one of them is a diaphragm that vibrates in response to sound waves. The diaphragm vibrates and changes the distance between the two plates. Then the capacitance changes which influences the electrical signal.

#### 5.1.1.3 Micro-Electro-Mechanical Systems microphone

In the past decades, microphone for UE has change from carbon microphones to electret condenser microphones. Recently the MEMS microphone is spread rapidly, benefited from its advantages of high stability and small volumes.

According to the techniques of microfabrication, the MEMS microphone is much smaller and allow integrate other components including preamps, ADC with transducer in one package under the control of integrated microelectronics.

Which means for manufacturers, it much easier to build the capture system, MEMS microphone can output the digital signal directly. In other hand, it allows need to select the component more carefully. Since the microphone is much smaller and very uniform in their mechanical properties, it's suitable for UE and make immersive audio become possible for economic portable UE like mobile phone.

#### 5.1.1.4 Contact microphone

Contact microphone is a type of microphone that senses solid vibrations through direct contact with a surface.

Compared to the acoustic microphones, the contact microphones have the benefit of not to capture sound waves in the air, but to capture mechanical vibrations of the target object. Hence, it’s resistant to noise in air.

Nowadays, bone conduction microphone, which is a special kind of contact microphone, is very popular on TWS headphones. It is used to capture high SNR speech signal even in complex scenarios.

#### 5.1.1.5 Other microphones

TBD

#### 5.1.1.6 Summary

From a size perspective, the MEMS microphones are the best choice for most portable UE (like mobile phone, headphone). The study will mainly focus on this miniature microphone consider the immersive audio system is much more complex.

Other microphones will also be considered, like the dynamic microphone and condenser microphone still dominate the professional audio industry.

## 5.2 Preamps

TBD

## 5.3 ADC

TBD

## 5.4 Clock

TBD

## 5.5 Directivity

Directivity is a very important part in immersive audio, every immersive audio format has requirement on directivity. Even for objective audio, we also need take care of the directivity to avoid the influence of environment noise.

### 5.5.1 Traditional approaches used in immersive audio

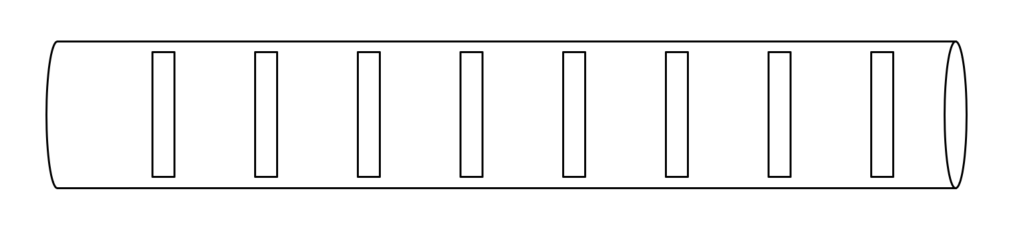
#### 5.5.1.1 Directional microphone capsule

Most directional microphone is using two closely diaphragms that electrically subtracted from each other to provide a range of polar patterns.

#### 5.5.1.2 Interference tube

Interference tube is usually used on shotgun microphones. Make it the more directional than a typical cardioid or supercardioid microphone.

Interference tube is a long, narrow extended tube that is placed in front of the microphone capsule and has multiple small holes along its length. It creates phase shfit for sounds arriving from off-axis directions, the off-axis sound will arrive at the diaphragm with varying phase relationships and so partially cancel one another out.



**Figure 5.5.1.2-1 The schematic diagram of interference tube**

#### 5.5.1.3 Binaural acoustic stimulation

TBD

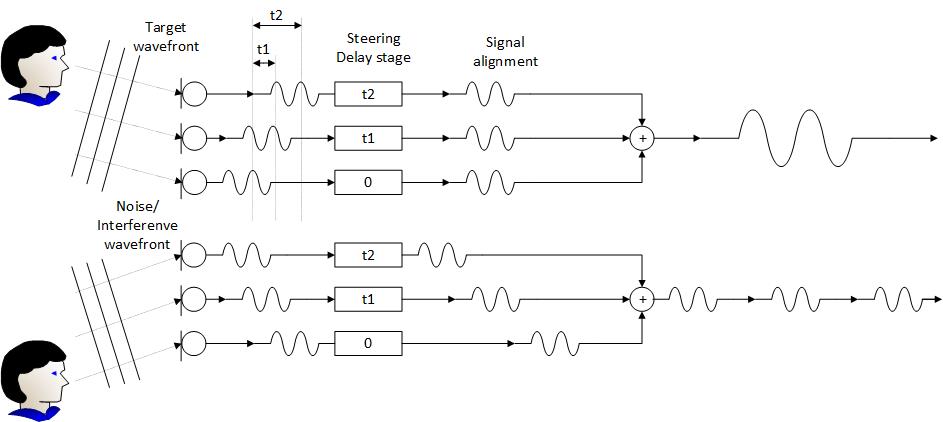
### 5.5.2 Beamforming microphone array

Research on microphone array beamforming began in the late 1960s, although some basic principles can be traced back to the 1930s when directional microphones were invented. Early work in this field was strongly influenced by sensor array theory developed in the radar and sonar fields.

Beamforming is a very popular technology to achieve target directivity, though it’s mostly used for mono speech now, it is great potential in immersive audio. There are also many studies in this area.

This proposal starts with two fundamental technologies: Delay-sum and differential. And aim for the suitable solution for immersive audio on UE.

#### 5.5.2.1 Delay-sum microphone array

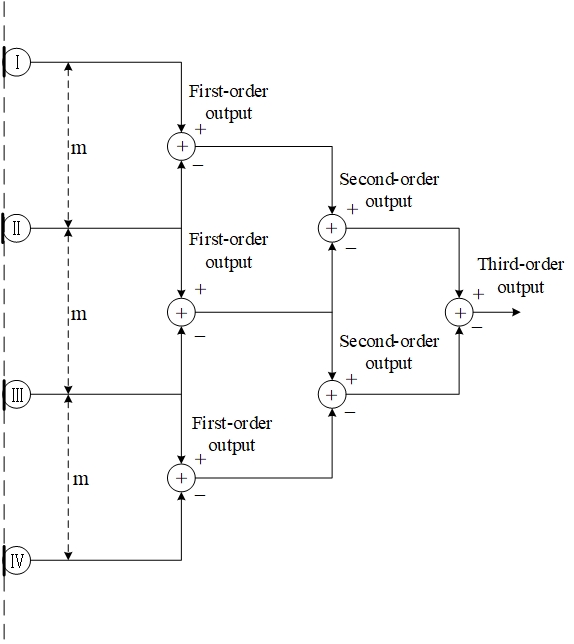


**Figure 5.5.2.1-1 The diagram of Delay-sum microphone array**

The basic idea of this technology is to delay the output of each microphone by an appropriate amount of time. The phase relationship between the microphones is carefully controlled to ensure that the signal form desired direction have the same phase so that they can be reinforced.

Though the delay-sum microphone array can obtain a very sharp directionality. However, the biggest problem with this beamformer is that its beam pattern changes significantly with frequency.

#### 5.5.2.2 Differential microphone array



**Figure 5.5.2.2-1 The diagram of Differential microphone array**

In Differential Microphone Array (DMA), the signals from two or more microphones are subtracted from each other to create a special directivity. The traditional directional microphone can also be seen as a special kind of differential beamforming.

By adjusting the weight and phase of the differential signal, we can all get different directivity like: cardioid, bidirectional (Figure-8), supercardioid, hypercardioid, subcardioid (wide cardioid).

Due to the smaller spacing between microphones, the size of array is usually smaller, making it easy to integrate into UE such as earphones, mobile phones, etc. Another characteristic of DMA is that its directivity is frequency-invariant; therefore, they are suitable for processing broadband speech and audio signals.

Editor’s Note: this is basis for further work

# 6 Acoustic design

Editor’s Note

* Relevant acoustic design content is envisioned.
* Including acoustic structure, microphone array design, etc.

## 6.1 Stereo capture

### Principle of stereo signal representation

The basic idea behind the stereo recording technique is to capture two signals with a proper relationship. By controlling the relationship between the two signals, it creates sound image with spaciousness, direction and depth feeling for listeners. And it can be reproduced through headphones or loudspeakers.

### Characteristic of stereo capture

Compared to other formats, stereo capture does not aim to accurately reproduce the original sound field. Instead, its focus is on creating convincing illusory sound images for listeners, which is achieved by generating enough perceptual cues. It can provide a natural and realistic experience to the listeners in a limited range of listening zone. And it is more technically mature.

### Factors that affect stereo capture

The key cues that may influence the quality of stereo capture are interchannel time differences, interchannel level differences and frequency range, which have been discussed since the emergence of stereo audio.

In the past, the discussion of factors that affect stereo capture always revolves around microphone properties (such as directionality and frequency range) and the placement of microphones.

With advancements in audio processing, we now have more methods to control audio signals, which is highly promising for stereo applications. This is especially relevant since UE imposes strict restrictions on hardware due to space constraints. The ability to fine-tune audio signals through processing offers great potential for enhancing stereo performance despite various limitations, but it may also import more influence on audio experience, which needs to be carefully analyzed. Therefore acoustic design also needs to consider the characteristics of relevant processing.

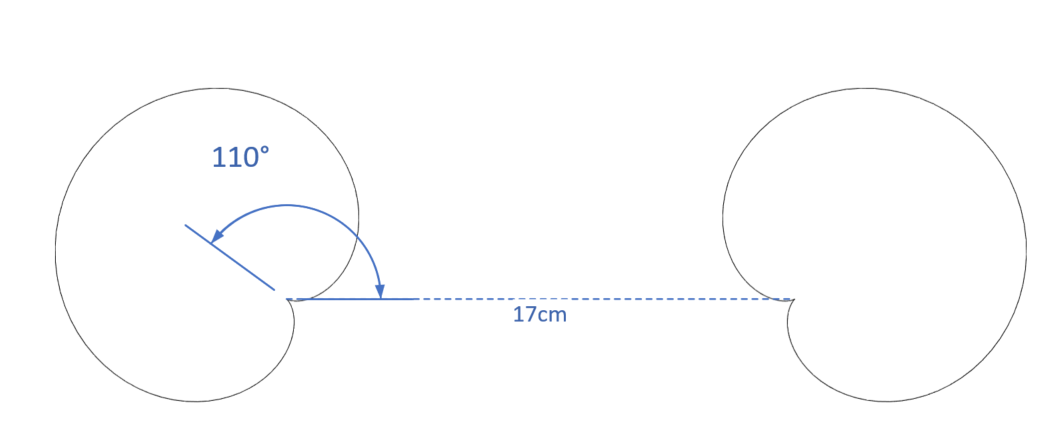
### Stereo microphone configurations

Stereo microphones can generally be classified as spaced, near coincident, baffled, coincident, and it is based on the directional characteristics of the microphones, their angle and distance between them. The following sub-clause presents some of these configurations.

#### Near-Coincident

Near-Coincident using two directional microphones placed close together at an angle is to capture stereo audio. This configuration utilizes the angle and distance between the microphones to create a suitable time and level inter-channel difference. There are number of near coincident configurations which are arranged at specific angles and distances, one of the well know near coincident stereo microphone is

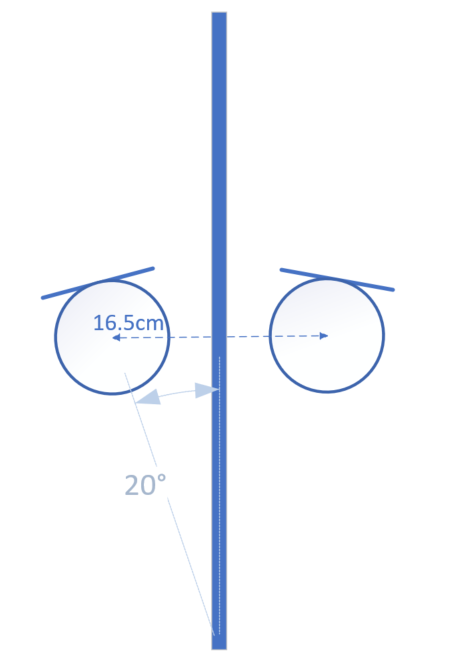
ORTF (Office de Radiodiffusion-Television Français). In the ORTF configuration, microphone pair is arranged to mimic the placement of human ears, it uses two cardioid microphones with 17cm apart and at a 110° angle from each other.



**Figure 6.1.4-1 The configuration of ORTF stereo microphone**

#### Baffled

A baffled configuration is a stereo recording technique that utilizes an acoustic baffle to increase the separation between the left and right audio channels. The baffle is typically a physical barrier that is placed between the two microphones. One of the examples of baffled microphone setup is OSS (Optimal Stereo System).OSS method utilizes a specially designed 30-cm disk covered with foam, with two omni-directional microphones mounted on opposite sides of the disk and angled slightly outward at 20°. The capsules of the two microphones are positioned 16.5 cm apart.



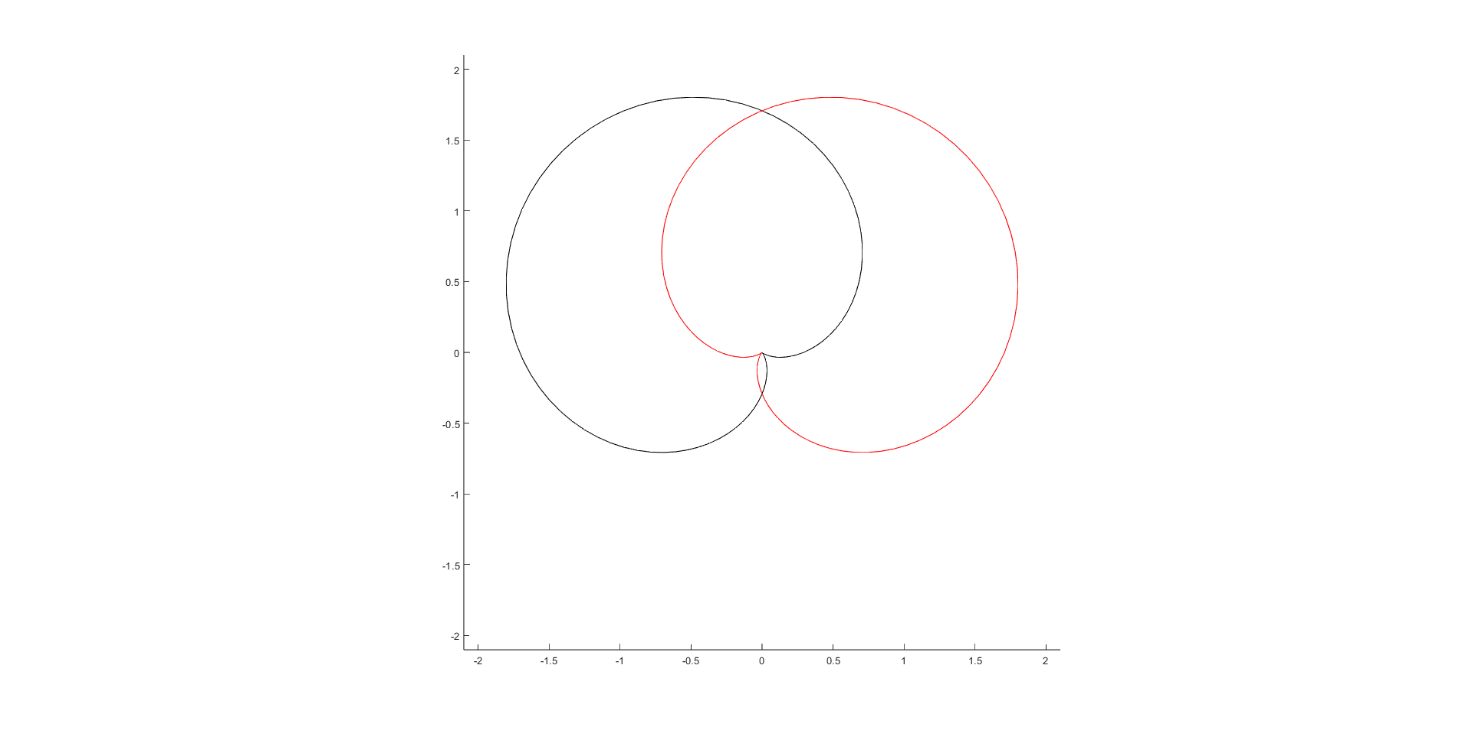
**Figure 6.1.4-2 The configuration of OSS stereo microphone**

#### Coincident

A coincident stereo microphone consists of two directional microphones placed at an appropriate angle at the smallest-possible spacings. Therefore, sound arrives with equal delay and different level and phase at microphones.

##### X/Y

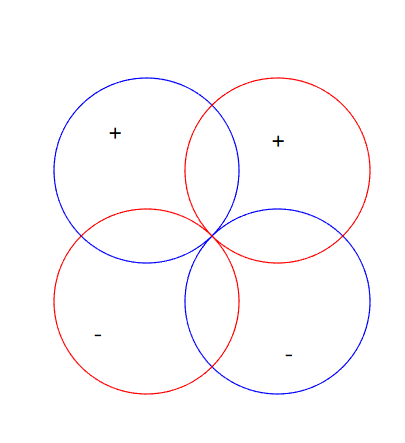
X/Y stereo microphone is commonly using two cardioid microphones ranging from 90-135°.



**Figure 6.1.4-3 The configuration of X/Y stereo microphone**

##### Blumlein

Blumlein stereo microphone consists of two bidirectional (figure-eight) microphones with 90° angle at the same place.

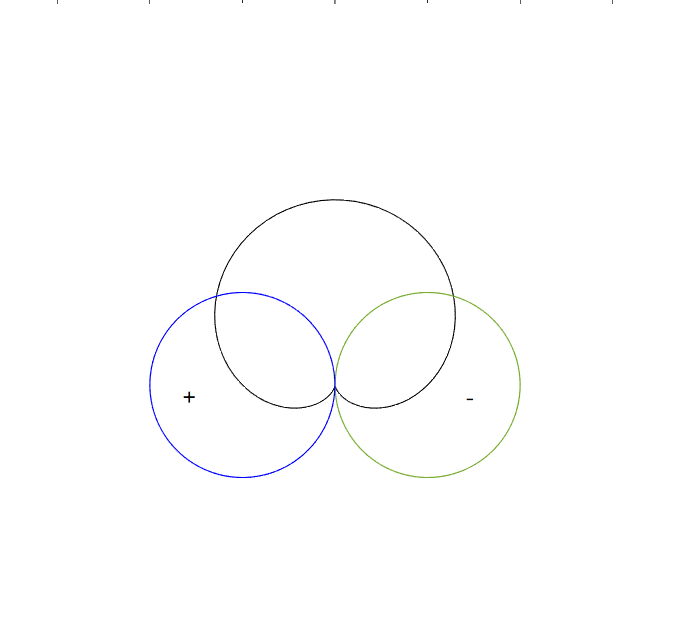
****

**Figure 6.1.4-4 The configuration of Blumlein stereo microphone**

##### M/S

M/S (mid-side) stereo microphone using a microphone (usually a cardioid) pointed forward, and a bidirectional (figure-eight) microphone oriented perpendicular to the directional microphone. The figure-eight microphone captures side signal, and the cardioid microphone capture mid signal. Therefore, we can obtain the left and right channel signal through the simple addition and subtraction.

In addition, controlling the ratio of the two signals, different angles can be obtained.



**Figure 6.1.4-5 The configuration of M/S stereo microphone**

#### Spaced

The spaced stereo microphone, also known as A/B stereo, is a stereo microphone technique that involves placing two omnidirectional microphones some distance apart from each other. This technique is commonly used with microphone spacings ranging from 0.3-1 meter.

The spaced stereo microphone technique utilizes the distance between two microphones to create a time difference and level difference between the left and right channels. This is caused by the difference in arrival time of sound waves at each microphone, as well as the absorption of sound by the air between the microphones.

NOTE: As most classic spaced configurations involve microphone distances greater than 30cm, which exceeds the size of current mobile phones, this aspect can only be listed for further study.

## Spatial audio capture

Several device form factors require spatial audio capture processing that is carefully designed and tuned for the specific multi-microphone array. In practice, a suitable parametric spatial audio capture analysis and processing is often implied for such devices.

### Binaural capture

#### Principle of binaural signal representation

The basic idea behind the binaural recording technique is to capture the two signals that form the input to our hearing. By capture these signals in the ears of a listener, it can retain the both timbre and spatial aspects, even keep the personal feature in binaural. And it can be reproduced accurately though headphones.

Editor’s note: number of microphones to be clarified, some processing could apply to get binaural signals from more than two microphones

#### Possible issues in binaural capture

Binaural audio can be defined as follows:

“Binaural audio is defined as a two-channel spatial representation of a soundfield as typically captured at the entrance of the ear canals and intended for direct presentation to the left and right ears over headphones”

However, the situation is not always so ideal. In most case, it’s hard to place the microphone just at the entrance of ear canals. So, it may be helpful to figure out what will influence binaural capture, therefore we can get better signal under limited conditions.

#### Factors that affect binaural capture

There are many cues that may influence the quality of binaural capture, e.g., interaural time differences, interaural level differences, interaural phase differences and spectral characteristics. The cues are influenced by the listener’s pinnae, head and body.

Earbuds usually have transducers blocked at the entrance of ear canals for playback, which occupy the most important location for binaural record and the microphone need to be set a few millimetres outside the entrance of the ear. The surface of earbud may also cause the reflection. It can be seen that the reflection from pinnae capture in microphone will be influenced.

#### Differences between binaural and stereo audio

Both binaural and stereo formats consist of two left and right channels. Several differences are outlined in Table 1.

**Table 1: Differences between binaural and stereo audio**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Format** | **Distance between left and right channels** | **Spatial cues** | **Suggest playback** | **Relationship between interaural differences and interchannel differences** | **Sound image** | **Binaural renderring** |
| **Stereo** | 0 to few meters | Interaural time differences and interaural level differences, | Headphone | Interaural differences equal to interchannel differences. | -90° to 90 ° (see NOTE) | Allowed |
| Loudspeaker | Interaural differences equal to interchannel differences plus differences caused by propagation from speakers to ears | Between left and right loudspeakers | Not allowed |
| **Binaural** | Equal to distance between ears | Interaural time differences, interaural level differences, interaural phase differences and spectral characteristics | Headphone | Interaural differences equal to interchannel differences. | All directions. | Not allowed |
| NOTE: When stereo audio playback on headphones is processed with binaural rendering, the sound image transforms to be positioned between left and right virtual loudspeakers. | | | | | | |

### Parametric spatial audio capture

#### Principle of parametric spatial audio representation

Spatial, or immersive, audio representations generally enable reproduction of audio scenes where the listener experience can optimally correspond with the real recorded situations and environments. This means, e.g., that a listener can hear audio sources around them in their original directions relative to the listening position and orientation. Faithful reproduction is generally possible via headphones, including head-tracked binaural rendering, or a loudspeaker setup that provides sufficient spatial capability.

Parametric spatial audio describes a spatial sound field according to a parametric representation. In a typical solution, a multi-microphone capture is compressed into a lower number of audio channels and associated spatial parameters. The parameters define the perceptually relevant properties of the sound field (e.g., directional information, how directional or diffuse the sound field is, etc.), while the audio channels define the actual energetic representation of the captured sound field.

The analyzed parameters and audio channels can be further utilized for the synthesis of an accurate spatial audio representation in a desired output configuration. For efficient and realistic synthesis, the directional parameters are mapped into perceptual spatial cues, which are further applied for the transport audio channels. To enhance the quality of the reproduction, the parameters can be analysed based on the desired auditory frequency bands.

Typical processing flow of the parametric spatial audio capture is illustrated in the Figure 6.2.2.1.1.1.

A diagram of a cell phone

Description automatically generated

Figure 6.2.2.1.1.1 Overview of parametric spatial audio capture

A prominent example of a parametric spatial audio representation is the Metadata-assisted spatial audio (MASA) format defined in 3GPP TS 26.258 [5]. Specifically, MASA comprises one (mono) or two (stereo) transport audio signals and metadata.

#### Characteristics of parametric spatial audio capture and representation

Parametric spatial audio capture is typically purpose-fit for the device form factor that utilizes it. There can also be different representations, or a specific capture algorithm can utilize only a subset of parameters that another capture algorithm uses. Therefore, two substantially similar devices can have different capture algorithms or at least different tunings.

Parametric spatial audio capture and representation typically enables relatively low computational complexity for capture processing and encoding, largely because the number of channels in the representation can be lower than the number of originally captured channels.

#### Factors that affect parametric spatial audio capture

Multiple factors may affect the quality and accuracy of the parametric spatial audio capture. A few factors are:

• Dimensions of the capture device (distance and the placement of the microphones)

• Number and the characteristics of the microphones (directivity and frequency response)

• Applied signal processing techniques for the captured multi-microphone signals, e.g., noise suppression and filtering

• Quality of the spatial parameter analysis algorithm and the device specific tuning

Furthermore, due to the above factors, the quality and the accuracy of the parametric spatial capture is not necessarily similar for all the capture directions. This is highly dependent, e.g., on the specific device form factor and its microphone placement/spacing.

#### Multi-microphone configurations in parametric spatial audio

The number and the configuration of the microphones can be arbitrary, but in principle the lowest number of microphones for accurate 2D planar representation is typically 2 for 180° (frontal) capture, and 3 for 360° capture. Furthermore, by increasing the number of the microphones, the whole 3D sound field can be captured accurately. Typically, minimum number for microphones in this case is 4.

Editor’s note: as opposed to binaural audio, this enables head tracking.

The associated direction parameters can be obtained, e.g., by assessing the inter-channel properties of the captured multi-microphone signals. Such properties could be inter-channel time-difference, coherence, and/or level difference. The analysis can be based on suitable frequency bands.

### Non-parametric spatial audio capture

* + - 1. Principle of Non-parametric spatial audio representation

Non-parametric spatial audio representation is used to provide spatial audio service at a reference point or area using certain number of audio channel data which have corresponding placements. the key point to the performance of the spatial audio service is to have appropriate audio data based on either standard or non-standard placements.

Due to the constraints of the UE device shape, it is very hard to generate the spatial audio representation directly from their embedded acquisition units or even from selected accessory devices. the ordinary solution is to use microphone array to catch raw signals and then do mathematical processing to output the expected results.

Example processing flow of the non-parametric spatial audio capture is referring to Figure 6.2.3.1-1

A black screen with blue lines

Description automatically generated

Figure 6.2.3.1-1 Overview of non-parametric spatial audio capture

* + - 1. Characteristics of non-parametric spatial audio capture and representation

The placement of microphones is subject to various restrictions of the end-user devices, the non-parametric spatial audio capture can be used to generate both standard format audio and non-standard format audio. Several standard audio formats listed in IVAS-4 P-doc are surround (5.1 and 7.1), surround + height (5.1+4 and 7.1+4), FOA, HOA2, HOA3, Object-based audio. The standard format audio is a necessary part of the interoperable solution between different kind of end-user devices.

Non-parametric spatial audio capture and representation is an important intermediate link joint between originally captured raw signals at sending end and rendered spatial signals at receiving end, it can allocate computational complexity of the end-to-end real-time spatial audio solution into two ends.

The accuracy of the non-parametric spatial audio representation can be significantly different because of the corresponding non-parametric spatial audio capture solution, it is necessary to carefully define the minimum performance requirements for the non-parametric spatial audio representation, based on this, higher performance is always pursued with better solutions.

* + - 1. Factors that affect non-parametric spatial audio capture

It is the same as parametric spatial audio capture, refer to section 6.2.2.3

* + - 1. Microphone configurations in non-parametric spatial audio

Immersive audio capture technology by microphones has been developed for decades, however, its corresponding microphone configuration is not fit for current mobile phones.

Numerous stereo microphone configurations have been developed to create immersive audio experiences. Several immersive configurations that are compatible with mobile phones are listed here.

##### Immersive audio ORTF configuration[6]

###### 6.2.3.4.1.1 ORTF-surround

The "ORTF-surround" configuration evolves from the ORTF stereo technique, consisting of two back-to-back ORTF stereo setups. It utilizes four super-cardioid microphones arranged in a rectangular formation, with each side measuring 10 cm by 20 cm and forming azimuth angles of 80º and 100º. The output from each microphone is individually routed to the corresponding Left (L), Right (R), Left Surround (LS), and Right Surround (RS) speakers to create an immersive audio experience. Refer to Figure 6.2.3.4.1.1-1.

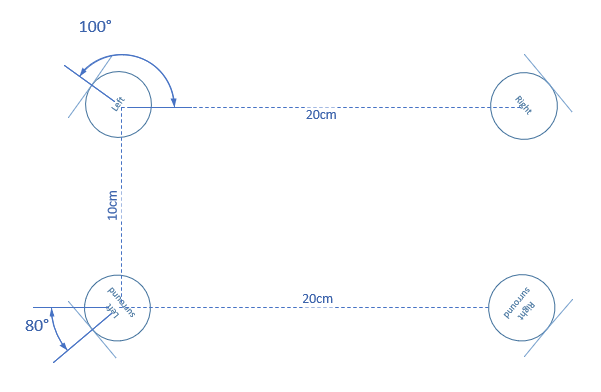


Figure 6.2.3.4.1.1-1 The configuration of ORTF-surround microphone

###### 6.2.3.4.1.2 ORTF-3D

The "ORTF-3D" consisting of two "ORTF-surround" configuration, one is placed directly on top of one another with 90º on elevation Angle, Refer to Figure 6.2.3.4.1.2-1.

A diagram of a circle with circles and lines

Description automatically generated

Figure 6.2.3.4.1.2-1 Left view of ORTF-3D microphone front channels

##### Immersive audio M/S configuration

###### 6.2.3.4.2.1 Double-M/S[7]

The "Double-M/S" configuration enhances the traditional M/S configuration by adding an additional rear-facing cardioid microphone. This rear-oriented mic integrates with the existing figure-8 microphone's signal, creating a pair of back-to-back M/S systems that capture a surrounding sound field. And corresponding channel signal can be obtained through the following equation, Refer to Figure 6.2.3.4.2.1-1.:

A diagram of a diagram

Description automatically generated

Figure 6.2.3.4.2.1-1 The configuration of Double-M/S microphone

###### 6.2.3.4.2.2 M/S-3D[8][9]

By incorporating a vertically oriented figure-8 microphone as “Z” signal into the "Double-M/S" configuration, the "M/S-3D" setup is capable of capturing the height channel, Refer to Figure 6.2.3.4.2.2-1.:

A circle with letters and numbers

Description automatically generated with medium confidence

Figure 6.2.3.4.2.2-1 Left view of M/S-3D microphone

##### IRT-cross

Another well-known configuration is the "IRT Cross," which is an equal segment microphone array. This array can be configured with either four cardioid microphones placed 20 cm apart or four supercardioid microphones spaced 14 cm apart. Refer to Figure 6.2.3.4.3-1.

A diagram of a diagram

Description automatically generated

Figure 6.2.3.4.3-1 The configuration of IRT-cross microphone

6.2.3.5 Comparisons between non-parametric spatial audio capture and parametric spatial audio capture

Parametric spatial audio employs a suitable parametric representation for the captured sound field, which is based on the analysis of the raw microphone signals to produce parameter metadata and potentially converting the raw microphone signal into specialized audio channel data.

In contrast, non-parametric spatial audio can only produce the audio channel data without parameter metadata. Once the microphones and their configuration are determined, the raw microphone signals can be converted into expected audio channel data by certain signal processing.

# 7 Signal processing

Editor’s Note

* Relevant signal processing content is envisioned
* Including relevant processing for audio format, enhancement solution for immersive, speech enhancement, etc.
* Relevant characterization of the audio capture performance.

## AEC

### 7.1.1. Principle of mono audio AEC

The aim of AEC is to minimize or eliminate the acoustic echo that occurs during a full-duplex communication from the other side. A case is described as in Figure 1, a speaker uses a sending device with voice communication capability to play an audio clip (red colour waveform signal), the audio clip is sent through the acoustic environment and received by the device's own microphone which generates the echo signal (blue colour waveform signal) for the sending audio clip. If an adaptive filter can be designed so that its output signal is just the opposite of the echo signal, like the green colour waveform signal, then the following step is to superimpose the blue colour waveform signal with the green colour waveform signal, thus the echo signal is eliminated depending on the performance of the adaptative filter.

A diagram of a speaker

Description automatically generated

**Figure 7.1.1.1: diagram of AEC**

AEC uses adaptive filters to counteract the impact of echoes and reverberations in the input signal, with the goal of a minimizing error. The general algorithm equations are as follows:







Where, is the input signal, is the desired signal, is the estimated echo signal, is the error signal, is the coefficient of the adaptive filter, is the step size parameter, and M is the length of the adaptive filter.

By minimizing the error and continuously updating the adaptive filter coefficients in an interactive way, the estimated echo signal becomes closer and closer to the desired signal, and finally making the error signal in an acceptable range.

### 7.1.2 Challenges for immersive audio AEC

For immersive audio services, the sound is reproduced through multiple speakers while simultaneously using several microphones. When the UE consists of N speakers and M microphones, the number of echo paths generated is the number of speakers multiplied by the number of microphones, i.e., N×M, which is several times the number of echo paths generated by a mono audio communication system. Taking stereo audio as an example, if a stereo audio device uses two speakers and two microphones to create a two-way audio service. In this case, Stereophonic Acoustic Echo Cancellation setup needs to estimate four echo paths. The significant increment in the number of echo paths poses a challenge to the computational complexity of adaptive algorithms. If the mono audio echo cancellation algorithm is directly extended to multi-channel, the algorithm is difficult to converge due to different people, different moments, and different positions of speech, and the adaptive filtering algorithm for mono audio case is not suitable for this situation. Obviously, extending mono audio echo cancellation algorithm directly to multi-channel will result in poor acoustic echo cancellation performance.

### 7.1.3 The current status of the research

The core of multi-channel acoustic echo cancellation algorithm is the same as that of single-channel echo cancellation algorithm - adaptive filtering algorithm. However, multi-channel acoustic echo cancellation faces more difficulties than single-channel AEC, the key point is the non-uniqueness of the solution. The strong correlation of multi-channel input signals will result in non-unique solutions when solving for the optimal filter coefficients, and the echo canceller cannot provide a unique echo path solution. The adaptive filter needs to fit a long impulse response to handle the same length of echo, and this larger filter order requires more historical data, leading to increased algorithm complexity and reduced convergence performance. Therefore, researchers currently address the problem of multi-channel acoustic echo cancellation mainly from two aspects: firstly, removing the correlation of input signals without affecting spatial sound perception, solving non-uniqueness; secondly, improving the adaptive algorithm, allowing the adaptive algorithm to have faster convergence speed and lower computational complexity to form multi-channel situations.

Example of stereo-AEC is referred to Annex B

Editor’s note: should add references

### 7.1.4 AEC for XXX UE

TBD

7.2 Microphone Array Signal Processing on device

According to previous investigations, Microphone Array Signal Processing (MASP) is an essential step for stereo capture, parametric spatial audio capture and non-parametric spatial audio capture, the basic processing is to transform the raw microphone signals into an expected audio representation, and enhancement could be done if necessary.

7.2.1 MASP for Channel-based

7.2.1.1 MASP for Stereo

In Clause 4.1, current mobile phones with dimensions of approximately 15cm in length and 7cm in width. Furthermore, a minimum of two microphones is required for stereo audio capture. Consequently, it is logical to use the most basic microphone array configuration as an example: two microphones positioned at a distance of less than 15cm or 7cm, depending on the orientation of mobile phones.So this subchapter will focus on how to complete microphone array signal processing for producing stereo signals based on such microphone array.

7.2.2 MASP for Binaural

According to clause 6.2.1, binaural capture on UE appears to be the sole format capable of directly obtaining audio from the raw microphone signals via the earbuds' microphones, enhancement processing could be done for better performance.

7.2.3 MASP for Scene-based

7.2.3.1 FOA

FOA signal model is very clear, it consists of four coincident signals: W, X, Y and Z, where W is an omnidirectional signal, while X, Y and Z are figure 8 directional signals aligned with the cartesian coordinate axes. So, the aim of the microphone array signal processing for FOA is to generate standard four coincident signals [2].

7.2.3.2 Matrix on current FOA microphones

The current FOA microphones utilize a tetrahedral configuration, which comprises four cardioid microphones oriented in the directions of Front-Left-Up (FLU), Front-Right-Down (FRD), Back-Left-Down (BLD), and Back-Right-Up (BRU). The W, X, Y, Z component are produced through matrix multiplying of the four cardioid signals, refer to equation(1). not follow the SN3D and ACN channel order.

7.2.3.3 HOA

HOA is for further study.

7.2.4 MASP for MASA

MASA format signals consist of audio signals and metadata. The metadata refer to 26.250[12] are derived from analysis of microphone raw signals, so microphone array signal processing is an essential module for producing MASA signal.

7.2.5 MASP for Object-based

According to 26.250 [12], Object-based audio consists of 1-4 individual mono object streams with associated metadata. Many existing technologies can be used to obtain object-based audio.

7.2.5.1 Mono object stream

The mono object stream may need audio with high quality and sufficient SNR, characteristics that closely match the existing mono audio solution. Consequently, the mono object stream may be derived from the current mono audio solution provided by UE.

7.2.5.2Associated Object Metadata

A minimal set of object metadata associated is the object position in the polar coordinate system described using azimuth [-180°,180°] and elevation [-90°, 90°] angle.

Direction Of Arrival (DOA) is commonly utilized in current audio services, it is to determine the direction of the audio that needs to be processed. The direction information can also be set as associated object metadata to describe the position of one audio object.

# 8 Example audio capture processing solutions

Editor’s Note

* Example solutions can be guidance on usage in conjunction with immersive voice and audio services codecs.
* Contributions are invited providing at least overview descriptions of the example audio capture solutions, illustrating the signal processing from the raw microphone feeds to the stereo/spatial audio formats.

To identify processing solutions at the capturing end, capturing scenarios are described as follows.

## Capture Scenario: Telephony Communications

**Capturing Type**: Multi-Microphone Capturing

**Description**

**Summary**

Call was established between Tom and Harry.

Tom device has multi-microphone capturing capability and Harry conversing via headphones connected to his communication device.

During the conversation with Harry, Tom wishes to share his experience and he changes the orientation of the portable communication device from portrait to landscape.

Tom device activates suitable microphone array configuration based on orientation of the device to maintain its intended position and to allow the listener (Harry) feel immersed in the experience sharing, providing a natural and enjoyable listening experience.

**User Story:**

Tom planned vacation with his friends, and they are in the Bhutan airport. One of Tom’s friend, Harry who is part of travel had to drop out of the vacation at the last minute. Harry felt devastated as he called Tom to break the news. To lighten up Harry’s mood and ensure she didn’t feel left out, Tom came up with an idea of sharing vacation experience with Harry daily. Tom knew it wasn’t the same as having Harry there in person, but he determined to make Harry feel like a part of the trip, even from afar. As Tom and his rest of the friends are at the observatory deck of the airport, decided to share his experience from day-0 (flights land off - takeoff, airport ambience, picturesque mountains in the backdrop etc.,) to Harry. Tom extends his hand holding his communication device in landscape mode towards the flight landing and takeoff with beautiful mountain view at the background. Harry can now view and listen to the airport observatory deck scene clearly which brought a smile on Harry’s face, as she feels she is part of the scene. As Tom and rest of their friends embarked on their vacation, he stayed true to his promise by giving virtual tour of day’s highlights to Harry.

**Device**

UE (Smartphone, Tablet), Headphones (Over the ear, on ear, In ear)

**Pre-condition:**

Tom’s UE implements multi-microphone capture with activation of relevant microphones.

Headphone connected to Harry’s UE

**Feasibility**

Availability of Multi-microphone capture is getting more common on smartphones.

**Potential Processing Solutions**

* Analog-Digital Conversion (ADC), Echo Cancellation, Noise Suppression, Automatic Gain Control, activation of suitable microphones based on device orientation.

[

Editor’sNote: this 8.2 section is used as a baseline for further refining

## Example design of spatial audio capture for multi-microphone UE devices

**Design**

**1 Overview**

A diagram of a wind turbine

Description automatically generated

A spatial capture on a modern mobile device may be implemented by utilizing the various microphones placed across the device, capture enhancement processing chain and a Ambisonic Static Upmixer. The capture stack incorporates simultaneous capture of 3 or more on-device microphones and an optional mono/stereo echo reference signal as inputs to produces a FOA output which can stored/transmitted using the IVAS (Immersive Voice and Audio Services) codec.

The various components of a spatial capture chain are as follows:

1. Raw Microphones Unprocessed Input and Echo Reference Signal
2. Compensation of the raw microphones self noise and frequency response based on measured and simulated data.
3. Acoustic Echo Cancelation
4. Wind Noise Estimation/Reduction
5. Content Based Processing
6. Static Ambisonic Upmix based on:
   1. Acoustic modelling of the phone and microphones via Finite Element Method analysis.
   2. Perceptually-based complex optimisation of upmixing coefficients accounting for the acoustic modelling.
7. Levelling and Limiting
   1. **Raw Microphone Input and Echo Reference**

Simultaneous capture of unprocessed raw inputs from all microphones is essential to derive the very best spatial capture utilizing a static ambisonic upmix. For generating first order ambisonic with good horizontal accuracy on the horizontal plane, at least 3 microphones are needed, which may be placed spatially on the mobile device to cover different planes. A typical placement of Top, Bottom and Rear microphones is seen on most modern mobile devices, however coplanar configuration is not optimal to resolve front/back confusion. An ideal arrangement of microphones should target to cover the three axes and spread evenly across the device to provide more spatial resolution, though practical constraints of phone design limit the spacing along the front-back axis to less than one centimetre.

A raw unprocessed capture is essential for spatial processing as any uncontrolled processing such as noise reduction, automatic gain control or echo cancellation on individual microphones will lead to loss of spatially sensitive background noise and in general cause front/back confusion.

It is however beneficial to apply controlled processing to each of the raw signals, to attenuate their electrical noise floor and to compensate for any anomalies in the frequency response due to the inherent response of the mic or to the effects of its placement inside ported holes on the edges of the device. These processings are based on measured data.

The echo reference is a mono or stereo signal representative of the audio played out from the loudspeaker and used by the AEC processing. It is necessary in full duplex communications if there exists an acoustic path between the loudspeaker and the device microphones. If none exists, such as when using headphones, echo reference is optional.

* 1. **AEC**

The baseline approach is to apply traditional Acoustic Echo Cancellation on the individual microphone channels.

* 1. **Wind Noise Reduction**

The baseline approach is to apply wind noise reduction on the individual microphone channel by non-linear signal processing. A DNN based noise reduction/speech isolation trained on wind noise is also found to be effective to reduce wind noise. Such a system performs well to maintain speech but incurs a higher processing costs and latency.

* 1. **Content Based Processing**

Content based processing is an optional set of processing that can greatly enhance spatial experience by identifying the content type and applying specific enhancements. Since the processing is content dependent, a general audio classifier or a speech isolation processing is necessary to identify regions of interest. Classification across Music, Speech and Background Noise has been found suitable for relevant scenarios. A real-time classifier might have delayed response to events, and false transitions might be triggered due to low or no lookahead. A classifier confidence smoothing or a state machine can be employed to minimize false transitions and identify class switching events.

A diagram of a speech process

Description automatically generated

Some of the processing that utilise events based on classification to apply specific enhancements are:

* Adaptive Background Noise Estimation and Reduction
* Remixing Speech/Music levels vs Ambient noise
  1. **Static Ambisonic Upmix**

This block forms the most important piece of the spatial capture and is responsible for mapping the multi-microphone capture to first order ambisonics or higher. Out of the various techniques to upmix a multi-microphone capture to FOA, a perceptually designed static upmix matrix yields reasonably accurate spatial performance with few or no drawbacks. The response per frequency band is generated across each microphone towards each ambisonic channel using: i) a model of how each microphone responds to the incoming sound field from a dense set of directions, based on Finite element method simulations which account for the 3D placement of microphones across device body structure and resultant acoustic energy transfer; ii) a perceptually-motivated optimisation where the complex upmix matrix coefficients are determined based on the data computed at the simulation step, and a target perceptual spatial accuracy. These steps of analysis, simulation and upmix design are performed offline once per device by assuming a simpler 3D geometry of the device and microphones as point sinks.

Another technique to generate the same matrix is to measure impulse responses from various direction of arrivals to each microphone and then inverting the matrix, but such technique would lead to obtaining the Ambisonics components independently from each other, with no guarantee of a perceptually-optimal result once the components are combined in the renderer..

The Upmixer utilizes the static matrix and applies per frequency band transformation to achieve ambisonic channel output.

A diagram of a matrix

Description automatically generated with medium confidence

* 1. **Levelling and Limiting**

As a final step to spatial audio capture, loudness correction on the ambisonic output may be desired to provide uniform loudness in all captures also preventing signal overload on rendering side, e.g. binaural stereo rendering. It is difficult to predict final loudness after render in an ambisonic domain and so a partial stereo downmix might be created to access the loudness and apply desired gain offset in the ambisonic domain. Such a gain application may be coupled with a limiter that can be used to prevent signal saturation.

]

# 9 Conclusions and Recommendations

Editor’s Note

* Provide recommendation on potential work for audio capturing based on the findings in this study.

]

Editor’s Note: the chapter structures are for further update.

Annex A: UE size

[

## A.1 Mobile phone size

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Date  (MMM-YY) | Length(mm) | Width(mm) | Depth(mm) |
| 1 | Mar-23 | 146.7 | 71.5 | 7.8 |
| 2 | Mar-23 | 160.8 | 78.1 | 7.8 |
| 3 | Mar-23 | 162.9 | 76.7 | 8.77 |
| 4 | Mar-23 | 162.9 | 76.7 | 8.77 |
| 5 | Dec-22 | 164.07 | 74.53 | 9.34 |
| 6 | Dec-22 | 164.35 | 75.29 | 9.7 |
| 7 | Sep-22 | 162.1 | 75.5 | 8.5 |
| 8 | Sep-22 | 162.1 | 75.5 | 9.92 |
| 9 | Apr-22 | 168.78 | 80.31 | 8.37 |
| 10 | Apr-22 | 163.7 | 73.9 | 8.5 |
| 11 | Mar-22 | 163.7 | 73.9 | 8.5 |
| 12 | Mar-22 | 160.3 | 72.6 | 8.7 |
| 13 | Sep-21 | 146.7 | 71.5 | 7.65 |
| 14 | Sep-21 | 163.6 | 74 | 9.1 |
| 15 | May-21 | 163.6 | 74 | 8.26 |
| 16 | Mar-21 | 163.6 | 74 | 8.7 |
| 17 | Mar-21 | 163.6 | 74 | 8.26 |
| 18 | Oct-20 | 146.7 | 71.5 | 7.4 |
| 19 | Oct-20 | 162.9 | 75.5 | 9.1 |
| 20 | Apr-20 | 158.2 | 72.6 | 8.95 |
| 21 | Dec-19 | 160.3 | 74.3 | 7.96 |
| 22 | Sep-19 | 150.9 | 75.7 | 8.3 |
| 23 | Apr-19 | 156.6 | 74.3 | 9 |
| 24 | Sep-18 | 143.6 | 70.9 | 7.7 |
| 25 | Sep-18 | 157.68 | 74.06 | 7.47 |
| 26 | Jun-18 | 157.91 | 74.27 | 7.48 |
| 27 | Apr-18 | 149.1 | 70.8 | 7.65 |
| 28 | Mar-18 | 155.1 | 75.2 | 7.4 |
| 29 | Oct-17 | 143.6 | 70.9 | 7.7 |
| 30 | Oct-17 | 150.5 | 77.8 | 8.2 |
| 31 | Sep-17 | 138.4 | 67.3 | 7.3 |
| 32 | Sep-17 | 165.32 | 80.09 | 7.45 |
| 33 | Jun-17 | 142.2 | 68.9 | 8.9 |
| 34 | Nov-16 | 150.5 | 77.8 | 8.2 |
| 35 | Oct-16 | 153 | 74.3 | 6.58 |
| 36 | Sep-16 | 138.3 | 67.1 | 7.1 |
| 37 | May-16 | 146.8 | 72.6 | 7.5 |
| 38 | Nov-15 | 157.1 | 80.6 | 7.9 |
| 39 | Oct-15 | 151.3 | 76.3 | 8.15 |
| 40 | Aug-15 | 141.6 | 71.2 | 7.8 |
| 41 | Jun-15 | 143.2 | 71.9 | 8.5 |
| 42 | May-15 | 147.9 | 73.45 | 6.44 |
| 43 | Apr-15 | 144.9 | 71.9 | 6.4 |
| 44 | Oct-14 | 158.1 | 77.8 | 7.1 |
| 45 | Sep-14 | 138.1 | 67 | 6.9 |
| 46 | Jul-14 | 139.6 | 69.7 | 7.5 |
| 47 | Jul-14 | 139.2 | 68.5 | 8.9 |
| 48 | Mar-14 | 152.6 | 75 | 9.2 |
| 49 | Sep-13 | 123.8 | 58.6 | 7.6 |
| 50 | Sep-12 | 123.8 | 58.6 | 7.6 |

## A.2 Earbud size

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Length(mm) | Width(mm) | Depth(mm) |
| 1 | 35.8 | 18.9 | 17.7 |
| 2 | 35.9 | 18.5 | 17 |
| 3 | 35.3 | 20.7 | 23.3 |
| 4 | 27.5 | 21.05 | 24.4 |
| 5 | 31.97 | 21.13 | 23.18 |
| 6 | 33 | 22.44 | 21.81 |
| 7 | ‎28.4 | 21.3 | 23.4 |
| 8 | 40 | 30 | 24 |
| 9 | 28.45 | 21.34 | 23.37 |
| 10 | 43.6 | 17.8 | 23.2 |
| 11 | 26 | 29.6 | 21.7 |
| 12 | 29.1 | 21.8 | 23.7 |
| 13 | 38.1 | 20.6 | 20 |
| 14 | 33.66 | 17.83 | 18.13 |
| 15 | 29.1 | 21.8 | 23.7 |
| 16 | 41.5 | 20.4 | 17.8 |
| 17 | 41.4 | 18.5 | 16.8 |
| 18 | 37.5 | 23.9 | 21 |
| 19 | 30.9 | 23.9 | 21.7 |
| 20 | 40.5 | 16.5 | 18 |
| 21 | 30.79 | 18.26 | 19.21 |
| 22 | 33 | 17.4 | 18.4 |
| 23 | 33.2 | 21.9 | 24.9 |
| 24 | 38.82 | 18.6 | 16.81 |
| 25 | 30.9 | 20.9 | 23.5 |
| 26 | 30.5 | 24.3 | 21.6 |
| 27 | 41.8 | 23.7 | 19.8 |
| 28 | 20.9 | 17 | 21.1 |
| 29 | 21.6 | 19.9 | 18.7 |
| 30 | 33.9 | 21.9 | 19.7 |
| 31 | 30.08 | 16.55 | 18.21 |
| 32 | 30.2 | 23.8 | 22.2 |
| 33 | 20.5 | 18.5 | 15 |
| 34 | 30 | 19 | 24 |
| 35 | 28.2 | 19.6 | 18.8 |
| 36 | 22.2 | 23.3 | 16.6 |
| 37 | 23 | 20 | 15 |

## A.3Tablet size

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Length(mm) | Width(mm) | Depth(mm) |
| 1 | 280.6 | 214.9 | 6.4 |
| 2 | 248.6 | 179.5 | 7 |
| 3 | 195.4 | 134.8 | 6.3 |
| 4 | 247.6 | 178.5 | 6.1 |
| 5 | 289.1 | 196.1 | 5.5 |
| 6 | 261.89 | 178.17 | 6.4 |
| 7 | 246.9 | 156.7 | 7.85 |
| 8 | 260.88 | 176.82 | 6.85 |
| 9 | 253.8 | 165.3 | 6.3 |
| 10 | 326.4 | 208.6 | 5.5 |
| 11 | 253.8 | 165.3 | 6.3 |
| 12 | 246.8 | 161.9 | 6.9 |
| 13 | 291.71 | 191.12 | 6.49 |
| 14 | 267.3 | 167.4 | 6.9 |
| 15 | 252.1 | 163.64 | 7.35 |
| 16 | 277 | 178.95 | 6.99 |
| 17 | 253.95 | 165.18 | 6.51 |
| 18 | 245.08 | 154.84 | 6.94 |
| 19 | 252.2 | 163.8 | 6.99 |
| 20 | 259.73 | 176 | 6.67 |
| 21 | 266.03 | 191.6 | 6.59 |

## A.4 Laptop size

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Length(mm) | Width(mm) | Depth(mm) |
| 1 | 363.4 | 260.25 | 22 |
| 2 | 312 | 221 | 15.99 |
| 3 | 369 | 259.4 | 23.5 |
| 4 | 321.9 | 213.9 | 19.9 |
| 5 | 296.68 | 213.5 | 15.65 |
| 6 | 356.98 | 288.73 | 25.65 |
| 7 | 286.5 | 184.7 | 7.99 |
| 8 | 313.8 | 229.8 | 16.7 |
| 9 | 354.9 | 251.9 | 22.45 |
| 10 | 296.2 | 216.5 | 10.9 |
| 11 | 340.4 | 237.6 | 11.5 |
| 12 | 312.6 | 221.2 | 15.5 |
| 13 | 356 | 247.7 | 20.2 |
| 14 | 315.6 | 222.5 | 14.9 |
| 15 | 364.81 | 289.98 | 18.5 |
| 16 | 410.3 | 319.9 | 26.7 |
| 17 | 296.5 | 205.5 | 12 |
| 18 | 294 | 197 | 8.95 |
| 19 | 305.7 | 199.8 | 12.9 |
| 20 | 359.5 | 238.3 | 15.9 |

## A.5 Watch size

Rectangle type:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Length(mm) | Width(mm) | Depth(mm) |
| 1 | 45 | 38 | 10.7 |
| 2 | 49 | 44 | 14.4 |
| 3 | 40 | 34 | 10.7 |
| 4 | 53 | 45.7 | 14.7 |
| 5 | 52 | 41 | 14.55 |
| 6 | 50.96 | 42.4 | 14.9 |
| 7 | 57 | 44.5 | 14 |

Circular type:

|  |  |  |
| --- | --- | --- |
| Model | Diameter(mm) | Depth(mm) |
| 1 | 46.5 | 10.9 |
| 2 | 48.5 | 13 |
| 3 | 46 | 13 |

## A.6 AR glass size

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Length(mm) | Width(mm) | Depth(mm) |
| 1 | 167 | 173 | 52 |
| 2 | 179 | 159 | 48 |
| 3 | 152.5 | 159 | 54.7 |
| 4 | 175 | 146 | 44 |
| 5 | 159 | 148 | 52 |
| 6 | 290 | 200 | 57 |

## A.7 Car exterior size

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Length(mm) | Width(mm) | Depth(mm) |
| 1 | 5020 | 1945 | 1760 |
| 2 | 4770 | 1930 | 1625 |
| 3 | 4480 | 1970 | 1601 |
| 4 | 5099 | 1989 | 1750 |
| 5 | 4880 | 1896 | 1450 |
| 6 | 4720 | 2089 | 1442 |
| 7 | 5021 | 1987 | 1478 |
| 8 | 5218 | 1998 | 1800 |
| 9 | 4458 | 1841 | 1632 |
| 10 | 4678 | 1806 | 1474 |
| 11 | 5320 | 1945 | 1488 |

## A.8 Car exterior and interior size

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Exterior | | | Interior | | | | | | | |
| Length (mm) | Width (mm) | Height  (mm) | Head room front(mm) | Head room rear(mm) | Leg room front(mm) | Leg room rear(mm) | Hip room front (mm) | Hip room rear(mm) | Shoulder room front(mm) | Shoulder room rear(mm) |
| 1 | 4648.2 | 1838.96 | 1689.1 | 1043.94 | 1099.82 | 1054.1 | 977.9 | 1374.14 | 1419.86 | 1450.34 | 1419.86 |
| 2 | 4495.8 | 1739.9 | 1460.5 | 1003.3 | 922.02 | 1130.3 | 787.4 | 1292.86 | 1272.54 | 1348.74 | 1361.44 |
| 3 | 4750 | 1921 | 1624 | 1041 | 1001 | 1063 | 1029 | 1367 | 1286 | 1432 | 1372 |
| 4 | 5021 | 1987 | 1431 | 1008 | 968 | 1077 | 901 | 1393 | 1278 | 1484 | 1399 |

# Annex B: Stereo AEC

In the context of multi-channel AEC, stereo AEC is considered a typical basis, therefore, it takes stereo as an example for AEC algorithm analysis and issue investigation. The stereo AEC is shown in Fig.B-1.

A screenshot of a computer

Description automatically generated

Fig. B-1. Stereo AEC system.

The two input signals  and  are derived from the same sound source convolving with two distinct impulse responses  and , so they typically have high correlations. After transmission of the distant stereo signal to the receiving room (near-end) and playback through the loudspeakers, the signals converge through the impulse responses  and of the near-end room to a single microphone, forming the desired signal  together with the noise signal . Subsequently, two adaptive FIR filters  and are employed to estimate the impulse responses  and  of the near-end room, yielding the estimated desired signal . Finally, the difference between the desired signal and the estimated desired signal which termed as the error signal is transmitted back to the distant end. Meanwhile, both the error signal and the input signals contribute to the updates of the adaptive filter coefficients. The following description focuses on a single microphone, and the other microphone follows a similar process.

B.1 Intuitive understanding of Stereo AEC

As illustrated in Fig. 7.1.3.1-1, assuming is the truncated vector of the actual echo path impulse response in the near-end room, and is the coefficient vector of the FIR filter. We define vector  as the filter misalignment vector. Ideally,

 (1)

The impulse responses of the distant room are denoted as  and , then substitute and into equation (1), and to do Fourier transformation as follows:

 (2)

 and  are the Fourier transforms of and , and  are the Fourier transforms of  and , represents the Fourier transform of the signal source 𝒔. For the single-channel AEC, is zero, so as long as  is not zero at the frequency point of interest, ensuring  to zero enables the filter estimating the path perfectly. However, for stereo AEC,  must not be zero, and  as the distant sound source is not zero either. Hence, we can now simplify (2), giving

 (3)

It can be seen from (3) that it is obviously impossible to deduce the conclusion of and .

B.1.1 De-correlation based method for stereo AEC

The de-correlation based method is a common solution for stereo AEC, the principles are described in the following.

B.1.1.1 Analysis of stereo audio covariance matrix[10]

In practical scenarios, the length of the actual room impulse response is infinite, but the trailing amplitude is generally small, and the effective length of the amplitude is limited. The impulse response length of the far-end room is set as M, the impulse response length of the near-end room is set as N, and the adaptive filter coefficient length is set as L. The error signal at time n between the output signal of the adaptive filter and the desired signal can be expressed as

 (4)

Use recursive least square error formula, giving

 (5)

where ( ) is an exponential forgetting factor. The minimization of (5) leads to the normal equation

 (6)

Where

 (7)

is an estimate of the input signal covariance matrix and

 (8)

is an estimation of the cross-correlation vector between the input and output signals. In this scenario, if  is not full rank, the normal equation does not have a unique solution, and the solution to the adaptive filter convergence may deviate from the actual room impulse response. Here examining the problem in terms of the sizes of the adaptive filter coefficient length (L) and the impulse response length of the distant room (M) so as to consider the uniqueness of the regularization equation solution.

Because the impulse response of the far-end length is infinite, L<M accords with the actual situation. Construct a new vector of length 2L, where  is the truncated vector of , giving

 (9)

where

 (10)

 (11)

Hence, matrix  is non-singular from the perspective of the Wiener solution, and the adaptive filter in stereo AEC has a unique solution, however, due to the relatively small values of  and , the covariance matrix  is very ill-conditioned, exhibiting significant divergence in eigenvalues, resulting in a slow convergence rate of the adaptive filter. This is commonly referred to as the "non-uniqueness" issue in the multi-channel acoustic echo problems. Under the premise of , the misalignment of the solution is considered.

The length L of the adaptive filter is actually smaller than the impulse response length N of the near-end room. The near-end room impulse response  is divided into two parts, one is the vector  matching the first L points of the adaptive filter length, and the other is the trailing vector  with the length of N-L. At the same time, we make repeated supplementary definitions for the input signal to match the room impulse response length. After a series of derivation, the Wiener solution can be got as follows:

 (12)

It can be concluded that the estimated adaptive solution is closely related to  from the point of view of Wiener solution.

B.1.1.2 Relationship between channel correlation and condition number of the covariance matrix[11]

From the above description, we can know that the coefficients of the adaptive filter are closely related to the norm of covariance matrix R, and further deduce that the condition number is used to measure the ill-conditioned degree of covariance matrix. The concept of matrix condition number is the product of the norm of the matrix and the norm of its inverse matrix, which is used to express the sensitivity of matrix calculation to error signals. Therefore, we use the condition number to establish the relationship between the correlation of stereo signals and the covariance matrix. Set , and the two-channel covariance matrix is thus given by

 (13)

where  is the mathematical expectation operator, the  is the covariance matrix between the th and th channel. It is noted that for , a Toeplitz matrix is asymptotically equivalent to a circulant matrix if its elements are absolutely summable, giving in (13) as (14)

 (14)

where is the Fourier matrix defined with elements .The matrix contains elements corresponding to the frequency bins, which are formed from the Discrete Fourier Transform (DFT) of the first column of .Letting  be the auto- and cross-correlation coefficients for and , respectively, establishing the relationship between  and  as follows

 (15)

The covariance matrix R of stereo signal can be expressed as

 (16)

In this content, E-norm is used to represent the condition number  of matrix R. The E-norm is equivalent to the F-norm scaled by a factor . Through this transformation, the dependence of conditional number on L is eliminated. After a series of derivation, the condition number under E-norm is obtained as follows

 (17)

Use positive definite covariance matrix to diagonalize to calculate , giving

 (18)

According to (15), We may now see that the square of the inter-channel coherence function at F frequency point can be expressed by the frequency spectrum of the input signal as

 (19)

After a series of derivation, (18) can be further expressed as

 (20)

B.1.1.3 Summary de-correlation based method

According to the description, the square of the condition number of covariance matrix R increases with the increase of the inter-channel coherence function , which means that the greater the inter-channel coherence, the greater the condition number of covariance matrix R, the more ill-conditioned matrix R, the greater the sensitivity of matrix calculation to error signals, and the more difficult it is for adaptive filters to solve coefficients. Therefore, in the stereo AEC, de-correlation of channel signals is a key step to the solution of adaptive filter coefficients, but this operation will bring sound quality degradation which should be balanced in this method.

# ]

Editor’s note: this is basis for further work.

Annex <X> (informative):  
Change history

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Change history** | | | | | | | |
| **Date** | **Meeting** | **TDoc** | **CR** | **Rev** | **Cat** | **Subject/Comment** | **New version** |
| 2023-02 | SA4#122 | S4-230317 |  |  |  | Initial version | V0.0.1 |
| 2023-04 | SA4#123-e | S4-230551 |  |  |  | Updated version based on SA4-post 122 24,March ,2023 | V0.0.2 |
| 2023-04 | SA4#123-e | S4-230646 |  |  |  | Update style and include agreed content in S4-230522 and S4-230523 | V0.0.3 |
| 2023-05 | SA4#124 | S4-230971 |  |  |  | Binaural capture on UE (from S4- 230881) and some online updates in addition | 0.1.0 |
| 2023-07 | SA4#124-Post | S4aA230088 |  |  |  | Update contents in scope section | 0.1.1 |
| 2023-08 | SA4#125 | S4-231347 |  |  |  | Integrate content based on S4aA230088 during SA4-e (AH) Audio SWG post 124 31 July 2023. | 0.1.2 |
| 2023-08 | SA4#125 | S4-231496 |  |  |  | Integrate agreed contents of stereo capture principal from S4-231460 | 0.2.0 |
| 2023-11 | SA4#126 | S4-231775 |  |  |  | Integrate agreed content from S4aA230109 and S4aA230111 | 0.2.1 |
| 2023-11 | SA4#126 | S4-231944 |  |  |  | Integrate agreed contents from S4-231661, S4-231717 and S4-231850 | 0.3.0 |
| 2024-01 | SA4#127 | S4-240218 |  |  |  | Integrate agreed contents from S4aA230132 | 0.3.1 |
| 2024-01 | SA4#127 | S4-240358 |  |  |  | Integrated following updates:  contents from S4-240150(two paragraphs) as 6.2.3.5  Contents from S4-240201 as 7.2  Contents from S4-240153 as Annex B  Contents from S4-240243 as 4.1.2  Contents from S4-240241  Integrate S4-240231 in the bracket for further refining | 0.4.0 |