



**5G;  
Study on Diverse Audio Capturing system  
(3GPP TR 26.933 version 19.0.0 Release 19)**



---

**Reference**

---

DTR/TSGS-0426933vj00

---

---

**Keywords**

---

5G

---

**ETSI**

650 Route des Lucioles  
F-06921 Sophia Antipolis Cedex - FRANCE

---

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B  
Association à but non lucratif enregistrée à la  
Sous-Préfecture de Grasse (06) N° w061004871

---

**Important notice**

---

The present document can be downloaded from the  
[ETSI Search & Browse Standards application](#).

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format on [ETSI deliver repository](#).

Users should be aware that the present document may be revised or have its status changed,  
this information is available in the [Milestones listing](#).

If you find errors in the present document, please send your comments to  
the relevant service listed under [Committee Support Staff](#).

If you find a security vulnerability in the present document, please report it through our  
[Coordinated Vulnerability Disclosure \(CVD\)](#) program.

---

**Notice of disclaimer & limitation of liability**

---

The information provided in the present deliverable is directed solely to professionals who have the appropriate degree of experience to understand and interpret its content in accordance with generally accepted engineering or other professional standard and applicable regulations.

No recommendation as to products and services or vendors is made or should be implied.

No representation or warranty is made that this deliverable is technically accurate or sufficient or conforms to any law and/or governmental rule and/or regulation and further, no representation or warranty is made of merchantability or fitness for any particular purpose or against infringement of intellectual property rights.

In no event shall ETSI be held liable for loss of profits or any other incidental or consequential damages.

Any software contained in this deliverable is provided "AS IS" with no warranties, express or implied, including but not limited to, the warranties of merchantability, fitness for a particular purpose and non-infringement of intellectual property rights and ETSI shall not be held liable in any event for any damages whatsoever (including, without limitation, damages for loss of profits, business interruption, loss of information, or any other pecuniary loss) arising out of or related to the use of or inability to use the software.

---

**Copyright Notification**

---

No part may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm except as authorized by written permission of ETSI.

The content of the PDF version shall not be modified without the written authorization of ETSI.

The copyright and the foregoing restriction extend to reproduction in all media.

© ETSI 2026.  
All rights reserved.

---

# Intellectual Property Rights

## Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the [ETSI IPR online database](#).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

## Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

**DECT™**, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™**, **LTE™** and **5G™** logo are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM®** and the GSM logo are trademarks registered and owned by the GSM Association.

---

# Legal Notice

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

The present document may refer to technical specifications or reports using their 3GPP identities. These shall be interpreted as being references to the corresponding ETSI deliverables.

The cross reference between 3GPP and ETSI identities can be found at [3GPP to ETSI numbering cross-referencing](#).

---

# Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the [ETSI Drafting Rules](#) (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

# Contents

|                                                         |    |
|---------------------------------------------------------|----|
| Intellectual Property Rights .....                      | 2  |
| Legal Notice .....                                      | 2  |
| Modal verbs terminology.....                            | 2  |
| 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 .....                                 | 11 |
| 4 Size and structure evolution of UEs .....             | 11 |
| 4.1 Mobile phones .....                                 | 11 |
| 4.1.1 Modern Mobile Phones.....                         | 11 |
| 4.1.2 Foldable Mobile Phones .....                      | 12 |
| 4.1.2.2 Book-Style Foldables .....                      | 12 |
| 4.1.2.3 Clamshell-Style Foldables .....                 | 13 |
| 4.2 Headphones .....                                    | 14 |
| 4.3 Tablets .....                                       | 14 |
| 4.4 Laptops .....                                       | 15 |
| 4.5 Watches .....                                       | 15 |
| 4.6 XR devices .....                                    | 17 |
| 4.7 Cars .....                                          | 17 |
| 4.8 Conclusion.....                                     | 18 |
| 5 Microphones used in immersive audio capture.....      | 18 |
| 5.1 Introduction .....                                  | 18 |
| 5.2 Transducer type .....                               | 18 |
| 5.2.1 Dynamic microphone.....                           | 18 |
| 5.2.2 Condenser microphone .....                        | 18 |
| 5.2.3 MEMS microphone.....                              | 19 |
| 5.2.4 Contact microphone .....                          | 19 |
| 5.3 Directional Microphone .....                        | 19 |
| 5.3.1 Introduction.....                                 | 19 |
| 5.3.2 Directional microphone capsule .....              | 19 |
| 5.3.3 Interference tube .....                           | 19 |
| 5.4 Binaural acoustic simulation .....                  | 20 |
| 5.5 Conclusion.....                                     | 20 |
| 6 Immersive audio capture format.....                   | 20 |
| 6.1 Stereo capture .....                                | 20 |
| 6.1.1 Principle of stereo signal representation .....   | 20 |
| 6.1.2 Characteristic of stereo capture .....            | 20 |
| 6.1.3 Factors that affect stereo capture .....          | 20 |
| 6.1.4 Stereo microphone configurations .....            | 21 |
| 6.1.4.1 Introduction.....                               | 21 |
| 6.1.4.2 Near-Coincident .....                           | 21 |
| 6.1.4.3 Baffled.....                                    | 21 |
| 6.1.4.4 Coincident .....                                | 22 |
| 6.1.4.4.1 Introduction .....                            | 22 |
| 6.1.4.4.2 X/Y .....                                     | 22 |
| 6.1.4.4.3 Blumlein .....                                | 22 |
| 6.1.4.4.4 M/S .....                                     | 23 |
| 6.1.4.5 Spaced .....                                    | 23 |

|           |                                                                                                    |    |
|-----------|----------------------------------------------------------------------------------------------------|----|
| 6.2       | Spatial audio capture .....                                                                        | 24 |
| 6.2.1     | Introduction.....                                                                                  | 24 |
| 6.2.2     | Binaural capture .....                                                                             | 24 |
| 6.2.2.1   | Principle of binaural signal representation.....                                                   | 24 |
| 6.2.2.2   | Possible issues in binaural capture .....                                                          | 24 |
| 6.2.2.3   | Factors that affect binaural capture .....                                                         | 24 |
| 6.2.2.4   | Differences between binaural and stereo audio.....                                                 | 24 |
| 6.2.3     | Parametric spatial audio capture .....                                                             | 25 |
| 6.2.3.1   | Principle of parametric spatial audio representation .....                                         | 25 |
| 6.2.3.2   | Characteristics of parametric spatial audio capture and representation.....                        | 26 |
| 6.2.3.3   | Factors that affect parametric spatial audio capture .....                                         | 26 |
| 6.2.3.4   | Multi-microphone configurations in parametric spatial audio .....                                  | 26 |
| 6.2.4     | Non-parametric spatial audio capture .....                                                         | 27 |
| 6.2.4.1   | Principle of non-parametric spatial audio representation .....                                     | 27 |
| 6.2.4.2   | Characteristics of non-parametric spatial audio capture and representation .....                   | 27 |
| 6.2.4.3   | Factors that affect non-parametric spatial audio capture.....                                      | 27 |
| 6.2.4.4   | Multi-microphone configurations in non-parametric spatial audio .....                              | 27 |
| 6.2.4.4.1 | Immersive audio ORTF configuration [4] .....                                                       | 27 |
| 6.2.4.4.2 | Immersive audio M/S configuration.....                                                             | 28 |
| 6.2.4.4.3 | IRT-cross.....                                                                                     | 29 |
| 6.2.4.5   | Comparisons between non-parametric spatial audio capture and parametric spatial audio capture..... | 30 |
| 6.3       | Summary .....                                                                                      | 30 |
| 7         | Microphone integration in UEs .....                                                                | 30 |
| 7.1       | Microphone integration in mobile phones.....                                                       | 30 |
| 7.1.1     | Modern mobile phones .....                                                                         | 30 |
| 7.1.2     | Foldable mobile phones .....                                                                       | 31 |
| 7.2       | Microphone integration in headphones .....                                                         | 33 |
| 7.3       | Microphone integration in tablets.....                                                             | 33 |
| 7.4       | Microphone integration in laptops.....                                                             | 34 |
| 7.5       | Microphone integration in watches .....                                                            | 35 |
| 7.6       | Summary .....                                                                                      | 36 |
| 7.7       | Microphone integration and frequency response.....                                                 | 36 |
| 8         | Signal processing.....                                                                             | 40 |
| 8.1       | AEC.....                                                                                           | 40 |
| 8.1.1     | Principle of mono audio AEC.....                                                                   | 40 |
| 8.1.2     | Challenges for immersive audio AEC .....                                                           | 40 |
| 8.1.3     | The current status of the research .....                                                           | 41 |
| 8.1.4     | AEC for different UEs .....                                                                        | 41 |
| 8.2       | Microphone Array Signal Processing on device .....                                                 | 41 |
| 8.2.1     | Introduction.....                                                                                  | 41 |
| 8.2.2     | MASP for Channel-based .....                                                                       | 43 |
| 8.2.2.1   | MASP for Stereo .....                                                                              | 43 |
| 8.2.3     | MASP for Binaural .....                                                                            | 43 |
| 8.2.4     | MASP for Scene-based.....                                                                          | 43 |
| 8.2.4.1   | FOA.....                                                                                           | 43 |
| 8.2.4.2   | Matrix on current FOA microphones .....                                                            | 43 |
| 8.2.4.3   | HOA .....                                                                                          | 43 |
| 8.2.5     | MASP for MASA .....                                                                                | 43 |
| 8.2.6     | MASP for Object-based.....                                                                         | 44 |
| 8.2.6.1   | Mono object stream.....                                                                            | 44 |
| 8.2.6.2   | Associated Object Metadata.....                                                                    | 44 |
| 8.2.7     | MASP for OMASA .....                                                                               | 44 |
| 8.2.8     | MASP for OSBA .....                                                                                | 44 |
| 8.3       | Beamforming.....                                                                                   | 44 |
| 8.3.1     | Introduction.....                                                                                  | 44 |
| 8.3.2     | Delay-sum microphone.....                                                                          | 45 |
| 8.3.3     | Differential microphone array.....                                                                 | 46 |
| 8.4       | Noise reduction .....                                                                              | 46 |
| 8.5       | Conclusion.....                                                                                    | 47 |
| 9         | Example audio capture processing solutions.....                                                    | 47 |

|                 |                                                          |           |
|-----------------|----------------------------------------------------------|-----------|
| 9.1             | Capture Scenarios.....                                   | 47        |
| 9.1.1           | Telephony communications .....                           | 47        |
| 9.2             | End-user device (UE) characteristics/prerequisites ..... | 48        |
| 9.2.1           | General.....                                             | 48        |
| 9.2.2           | Number and placement of microphones .....                | 48        |
| 9.2.3           | Loudspeaker number and placement of loudspeakers .....   | 48        |
| 9.3             | Capture solution for end-user devices .....              | 49        |
| 9.3.1           | Overview .....                                           | 49        |
| 9.3.2           | Compensation .....                                       | 49        |
| 9.3.3           | Enhancement.....                                         | 49        |
| 9.3.1           | Introduction.....                                        | 49        |
| 9.3.3.2         | General considerations .....                             | 49        |
| 9.3.3.3         | AEC.....                                                 | 49        |
| 9.3.3.4         | Noise reduction .....                                    | 50        |
| 9.3.3.4.1       | Introduction .....                                       | 50        |
| 9.3.3.4.2       | Wind noise reduction.....                                | 50        |
| 9.3.3.4.3       | Microphone noise reduction .....                         | 50        |
| 9.3.3.4.4       | Background noise reduction .....                         | 50        |
| 9.3.3.4.5       | Audio focusing .....                                     | 50        |
| 9.3.4           | Audio format generation.....                             | 50        |
| 9.3.4.1         | Introduction.....                                        | 50        |
| 9.3.4.2         | Example of stereo processing.....                        | 50        |
| 9.3.4.3         | Example of scene-based audio processing .....            | 51        |
| 9.3.4.3.1       | Introduction .....                                       | 51        |
| 9.3.4.3.2       | Content based processing .....                           | 51        |
| 9.3.4.3.3       | Ambisonic upmixer .....                                  | 51        |
| 9.3.4.4         | Example of parametric spatial audio processing.....      | 51        |
| 9.3.4.4.1       | Introduction .....                                       | 51        |
| 9.3.4.4.2       | Downmixing.....                                          | 51        |
| 9.3.4.4.3       | Spatial analysis .....                                   | 51        |
| 9.3.5           | Post-processing.....                                     | 52        |
| 10              | Conclusions and Recommendations.....                     | 52        |
| <b>Annex A:</b> | <b>UE size.....</b>                                      | <b>52</b> |
| A.1             | Mobile phone size .....                                  | 52        |
| A.1.1           | Modern mobile phone structure size .....                 | 52        |
| A.1.1.1         | Introduction.....                                        | 52        |
| A.1.1.2         | Length.....                                              | 52        |
| A.1.1.3         | Width .....                                              | 53        |
| A.1.1.4         | Depth .....                                              | 54        |
| A.1.2           | Book-style foldable mobile phone structure size.....     | 56        |
| A.1.2.1         | Length.....                                              | 56        |
| A.1.2.2         | Width .....                                              | 56        |
| A.1.2.3         | Depth .....                                              | 57        |
| A.1.3           | Clamshell-style foldable structure size.....             | 58        |
| A.1.3.1         | Length.....                                              | 58        |
| A.1.3.2         | Width .....                                              | 59        |
| A.1.3.3         | Depth .....                                              | 60        |
| A.2             | Earbud size .....                                        | 60        |
| A.3             | Tablet size .....                                        | 61        |
| A.4             | Laptop size .....                                        | 62        |
| A.5             | Watch size .....                                         | 62        |
| A.6             | XR device size.....                                      | 63        |
| A.7             | Car exterior size .....                                  | 63        |
| A.8             | Car exterior and interior size .....                     | 64        |

|                               |                                                                                              |           |
|-------------------------------|----------------------------------------------------------------------------------------------|-----------|
| <b>Annex B:</b>               | <b>Stereo AEC .....</b>                                                                      | <b>65</b> |
| B.1                           | Introduction .....                                                                           | 65        |
| B.2                           | Intuitive understanding of Stereo AEC .....                                                  | 65        |
| B.2.1                         | De-correlation based method for stereo AEC.....                                              | 66        |
| B.2.1.1                       | Analysis of stereo audio covariance matrix .....                                             | 66        |
| B.2.1.2                       | Relationship between channel correlation and condition number of the covariance matrix ..... | 67        |
| B.2.1.3                       | Summary de-correlation based method.....                                                     | 68        |
| <b>Annex C (informative):</b> | <b>Change history .....</b>                                                                  | <b>69</b> |
| History .....                 |                                                                                              | 71        |

---

# 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, in 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, end-user devices are usually configured with varying numbers of microphones and different microphone setup configurations. Therefore, different audio capturing capabilities are expected. Based on this, the present document provides a diverse audio capturing system.

# 1 Scope

The goal of this technical report is to study diverse audio capturing methods and applicable audio formats for the end-user device, considering the current physical and software constraints. The scope of the work is shown in Figure 1-1.

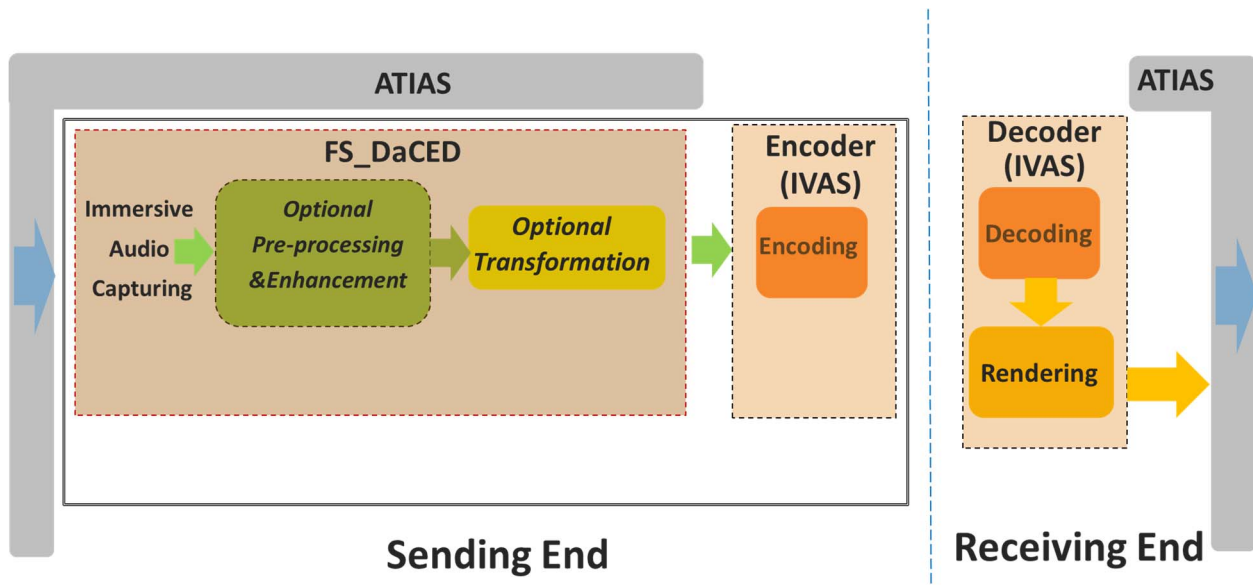


**Figure 1-1: Scope of TR 26.933 under the scope of the FS\_DaCED study item**

This document addresses audio capturing configurations for end-user devices, with the aim of equipping these devices with audio capturing capability to provide a truly immersive voice and audio service.

The document aims to study the following aspects:

- 1) Factors related to audio capture in different UE categories.
- 2) Components used in audio capture.
- 3) Acoustic design for audio capture.
- 4) Signal processing, such as microphone array beamforming processing, AEC processing etc.
- 5) Example of audio capture processing solutions.

# 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 TS 26.258: "Codec for Immersive Voice and Audio Services (IVAS); C code (floating point)"
- [4] 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.
- [5] Wittek, Haut, Keinath: "Double M/S – a Surround recording technique put to test", 24. Tonmeistertagung 2006
- [6] 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.
- [7] 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)
- [8] 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.
- [9] 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.
- [10] 3GPP TS 26.250: "Codec for Immersive Voice and Audio Services (IVAS); General overview
- [11] 3GPP TS 26.131: Terminal acoustic characteristics for telephony; Requirements
- [12] C. D. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, "Enhancement of spatial sound recordings by adding," J. Inf. Hiding Multimedia Signal Process., vol. 8, no. 6, pp. 1392–1404, 2017.
- [13] M. Geier, J. Ahrens, and S. Spors, "Enhancing binaural reconstruction from rigid circular microphone array recordings by using virtual microphones," in Proc. AES: Conf. Audio Virtual Augmented Reality, 2018, pp. 194–202.
- [14] Lübeck, T.; Arend, J.M.; Pörschmann, C. Spatial Upsampling of Sparse Spherical Microphone Array Signals. IEEE/ACM Trans. Audio Speech Lang. Process. 2023, 31, 1163–1174.
- [15] Chen, X.; Ma, F.; Bastine, A.; Samarasinghe, P.; Sun, H. Sound Field Estimation around a Rigid Sphere with Physics-informed Neural Network. arXiv 2023, arXiv:2307.14013.
- [16] 3GPP TS 26.253: "Codec for Immersive Voice and Audio Services (IVAS); Detailed Algorithmic Description including RTP payload format and SDP parameter definitions".

---

## 3 Definitions of terms, symbols and 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].

### 3.2 Symbols

Void

### 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].

|      |                                              |
|------|----------------------------------------------|
| ACN  | Ambisonic Channel Number                     |
| ADC  | Analog-to-Digital Converter                  |
| AEC  | Acoustical Echo Cancellation                 |
| AGC  | Automatic Gain Control                       |
| AR   | Augmented Reality                            |
| BLD  | Back-Left-Down                               |
| BRU  | Back-Right-Up                                |
| CI   | Confidence Interval                          |
| DMA  | Differential Microphone Array                |
| DOA  | Direction Of Arrival                         |
| FLU  | Front-Left-Up                                |
| FOA  | First Order Ambisonics                       |
| FRD  | Front-Right-Down                             |
| HOA  | High Order Ambisonics                        |
| HOA2 | Second Order Ambisonics                      |
| HOA3 | Third Order Ambisonics                       |
| IVAS | Immersive Voice and Audio Services           |
| IRT  | Institute for Radio Technology               |
| MASA | Metadata-Assisted Spatial Audio              |
| MASP | Microphone Array Signal Processing           |
| MEMS | Micro-Electro-Mechanical Systems             |
| M/S  | Mid-Side                                     |
| ORTF | Office de Radiodiffusion-Télévision Français |
| OSS  | Optimal Stereo System                        |
| SN3D | Spherical Harmonics Normalization 3D         |
| SNR  | Signal-to-Noise Ratio                        |
| TWS  | True Wireless Stereo                         |
| UE   | User Equipment                               |
| VR   | Virtual Reality                              |

---

## 4 Size and structure evolution of UEs

### 4.1 Mobile phones

#### 4.1.1 Modern Mobile Phones

Since 2012, the dimensions of mobile phones, including length, width, and depth, have been steadily increasing (refer to Annex A.1). This trend indicates a continuous growth in mobile phone size. This may be due 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 a full screen, which provides a market foundation for the increase in screen size. Detailed data on these dimensions are listed in Annex A.1

- The maximum dimensions for length, width, and depth are 168.78mm ,80.6mm and 9.92mm respectively.
- The minimum dimensions are 123.8mm (L), 58.6mm (W), and 6.4mm (D).
- The average dimensions are 152.65mm (L), 73.17mm (W) and 8.08mm (D).
- The 95% CI for length, width, and depth are (149.60 mm,155.69 mm), (71.92 mm,74.42 mm) and (7.85 mm,8.31 mm) respectively.

## 4.1.2 Foldable Mobile Phones

Foldable mobile phones are gaining significant growth in the market adoption, advances in flexible display technology and the introduction of new form factors are likely driving consumer interest. Foldable smartphones come in various dimensions, primarily it can be categorized into book-style and clamshell-style designs. Book-style foldables are generally larger in width when opened and bulkier in depth when folded while the clamshell style foldables are compact when folded while offering a substantial display when opened. The flexible design of foldable phones introduces considerable variation in audio capture.

### 4.1.2.2 Book-Style Foldables

Book-style foldables are generally offer larger display when opened suitable for multi-tasking and media consumption. These devices when closed looks bulkier (in-depth), however recent trends have shown decrease in tendency in the depth. Detailed data on these dimensions are listed in Annex A.1.2.

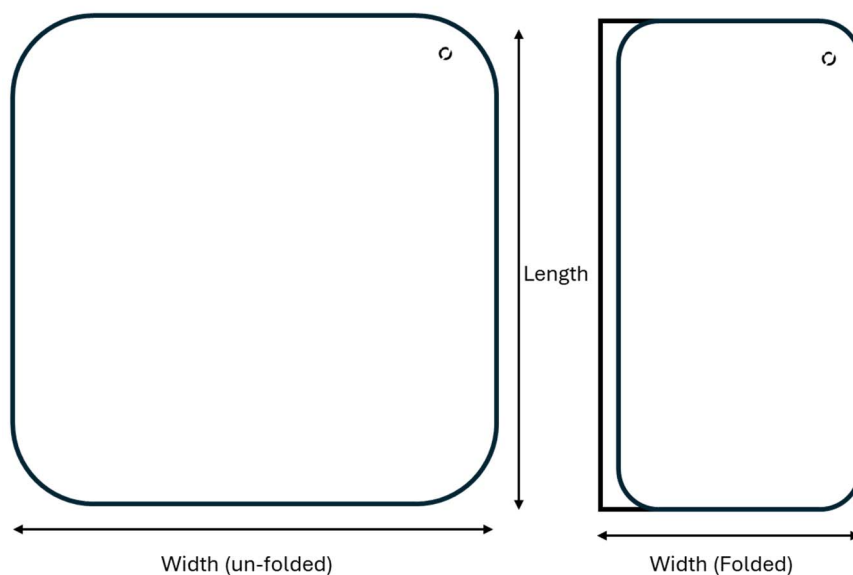


**Figure 4.1.2.2-1 Example of Book-style foldables**

#### Unfolded

- The maximum dimensions for length, width, and depth are 17.33 cm (L), 14.62 cm (W) and 0.8 cm (D) respectively.
- The minimum dimensions are 13.22 cm (L), 12.7 cm (W), and 0.435 cm (D).
- The average dimensions are 15.55 cm (L), 13.75 cm (W) and 0.59 cm (D).
- The 95% CI for length, width, and depth are (15.15 cm, 15.95 cm), (13.44 cm, 14.06 cm) and (0.54 cm, 0.64 cm) respectively.

#### Folded

- The maximum dimensions for length, width, and depth are 17.33 cm (L), 7.83 cm (W) and 1.72 cm (D) respectively.
- The minimum dimensions are 13.22 cm (L), 6.3 cm (W), and 0.92 cm (D).
- The average dimensions are 15.55 cm (L), 7.2 cm (W) and 1.29 cm (D).
- The 95% CI for length, width, and depth are (15.15 cm, 15.95 cm), (7.02 cm, 7.38 cm) and (1.18 cm, 1.4 cm) respectively.

#### 4.1.2.3 Clamshell-Style Foldables

Clamshell-style foldables are more compact in length when folded and lightweight, making them easier to carry while still offering a substantial display when opened in full-screen. Detailed data on these dimensions are listed in Annex A.1.3.

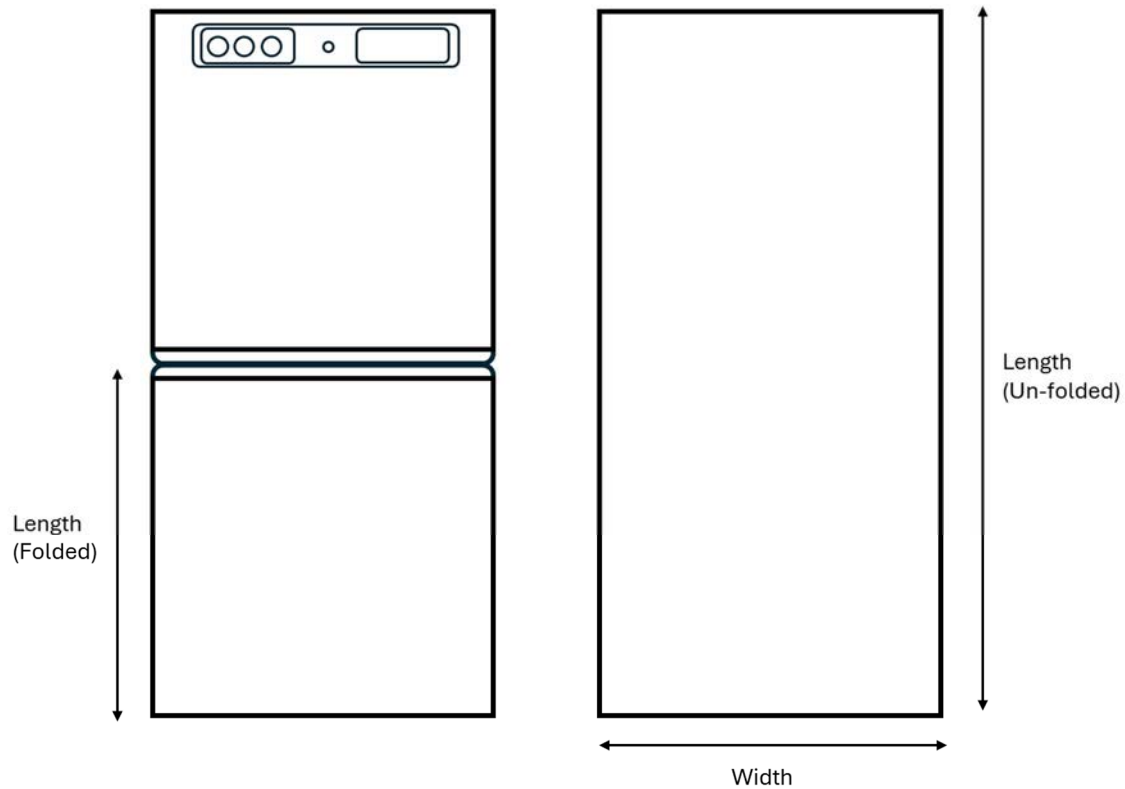


**Figure 4.1.2.3-1 Example of clamshell-style foldables**

##### Unfolded

- The maximum dimensions for length, width, and depth are 17.13 cm (L), 7.98 cm (W) and 0.79 cm (D) respectively.
- The minimum dimensions are 16.51 cm (L), 7.19 cm (W), and 0.69 cm (D).
- The average dimensions are 16.8 cm (L), 7.43 cm (W) and 0.73 cm (D).
- The 95% CI for length, width, and depth are (16.7 cm, 16.9 cm), (7.33 cm, 7.53 cm) and (0.71 cm, 0.75 cm) respectively.

##### Folded

- The maximum dimensions for length, width, and depth are 9.17 cm (L), 7.98 cm (W) and 1.73 cm (D) respectively.
- The minimum dimensions are 7.4 cm (L), 7.19 cm (W), and 1.49 cm (D).
- The average dimensions are 8.65 cm (L), 7.43 cm (W) and 1.58 cm (D).
- The 95% CI for length, width, and depth are (8.47 cm, 8.83 cm), (7.33 cm, 7.53 cm) and (1.54 cm, 1.62 cm) respectively.

## 4.2 Headphones

The mainstream headphones are TWS in-ear headphones. Therefore, some TWS in-ear headphones have been studied. Specifically, the length of all these TWS in-ear headphones is less than 4cm (refer to Annex A.2). Additionally, the depth and width of these in-ear headphones are around 2cm.

## 4.3 Tablets

Tablets, which are popular UE, are equipped with speakers and microphones. Annex A.3 lists 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, and the width is close to 20 cm. The models in the statistics range from 13.48 to 21.49 cm (refer to Annex A.3).

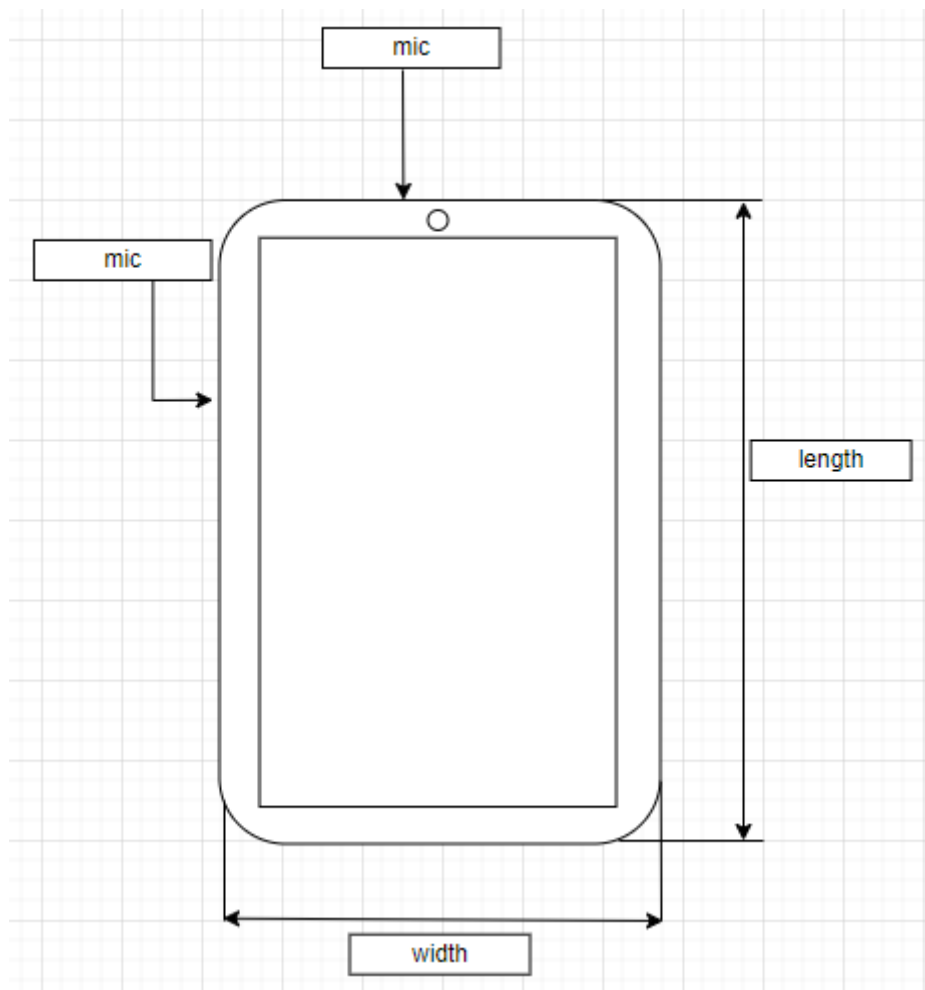


Figure 4.3-1 Tablet front view

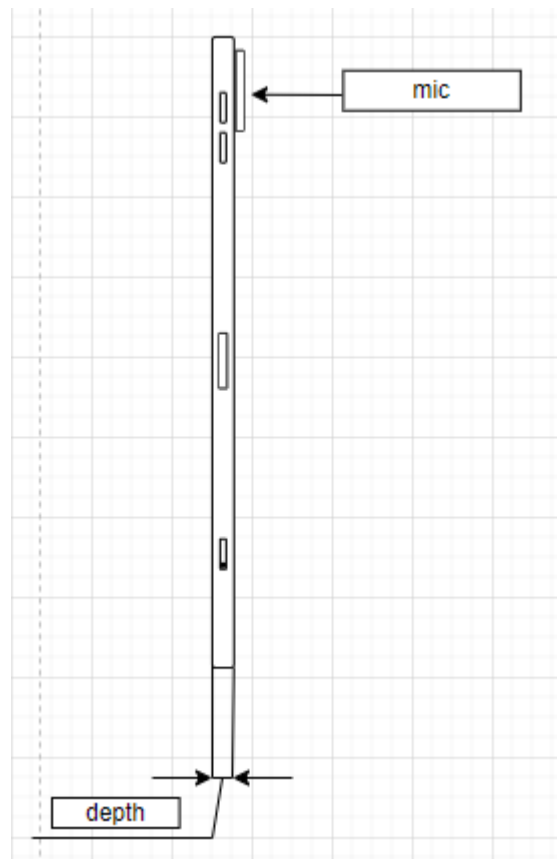


**Figure 4.3-2 Tablet side view**

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

## 4.4 Laptops

The primary distinction of laptops compared to tablets is their clamshell structure, which features a hinged screen and an attached keyboard. Additionally, most laptops have a larger structure size than mobile phones. The length of laptops ranges from approximately 28 cm to around 41 cm, resulting in a range of about 13 cm. Their widths span from about 18.5 cm to approximately 32 cm. The height of laptops extends from a few tenths of a centimetre (around 0.7 cm) to over 2.6 cm (refer to Annex A.4).

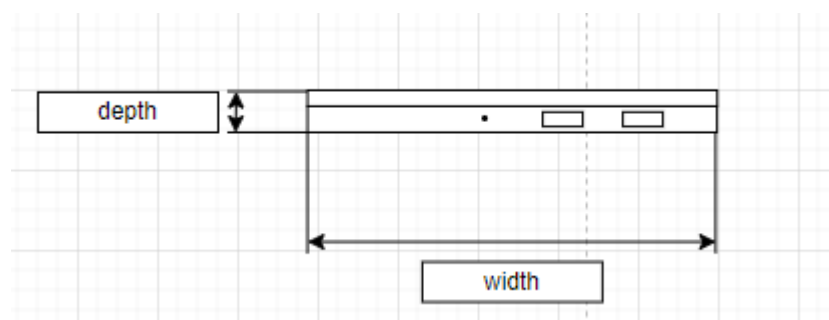


**Figure 4.4-1 Laptop side view**

## 4.5 Watches

Many watches possess calling capabilities. Primarily, smartwatches come in two main shapes: circular and rectangular. For the circular type, the diameter is around 4.7 cm, and the depth ranges from 10.9 to 13 cm (refer to Annex A.5).

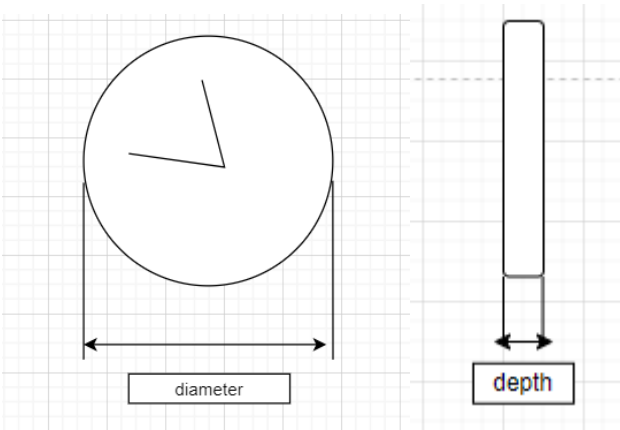


Figure 4.5-1 circular watch

For the rectangular 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 (refer to Annex A.5).

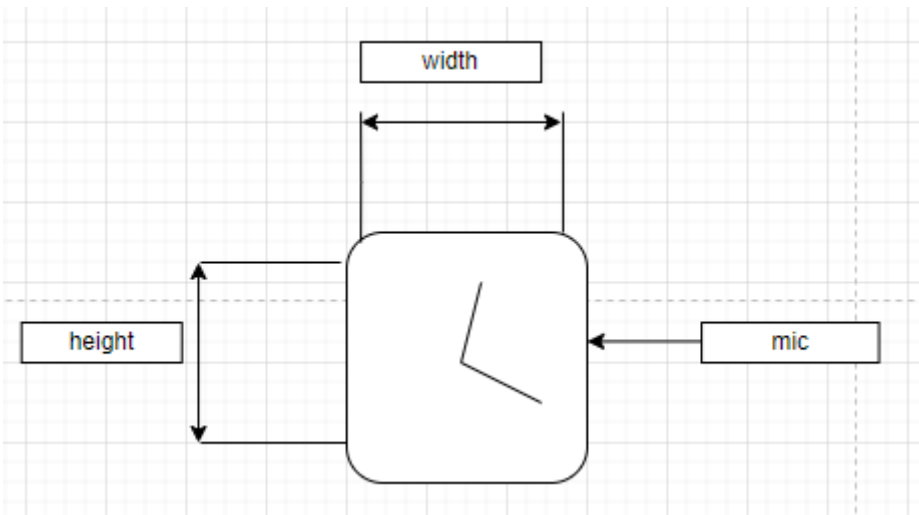


Figure 4.5-2 rectangular watch front view

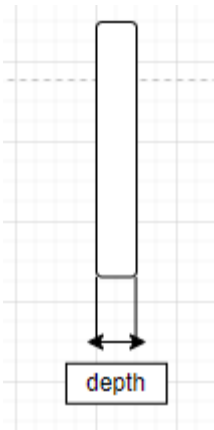


Figure 4.5-3 rectangular watch side view

## 4.6 XR devices

XR devices 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 mimic the appearance of regular eyewear. Therefore, the average size of the investigated model is approximately 157.48 mm in width, 130.92mm in length, 41.22mm in height respectively (refer to Annex A.6).

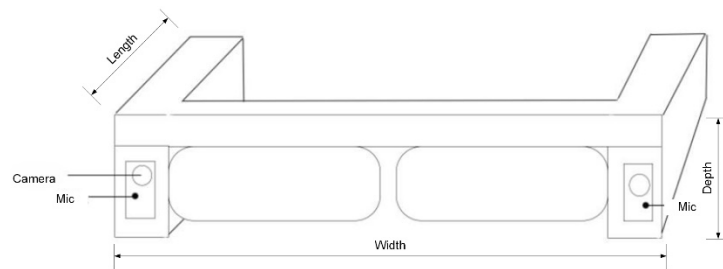


**Figure 4.6-1 XR device**

## 4.7 Cars

The dimensions of some mainstream civilian cars have been investigated. The lengths vary from 445.8cm to 532cm, widths range from 180.6cm to 208.9cm and heights fluctuate between 144.2cm and 180.0cm (refer to Annex A.7 and Annex A.8).

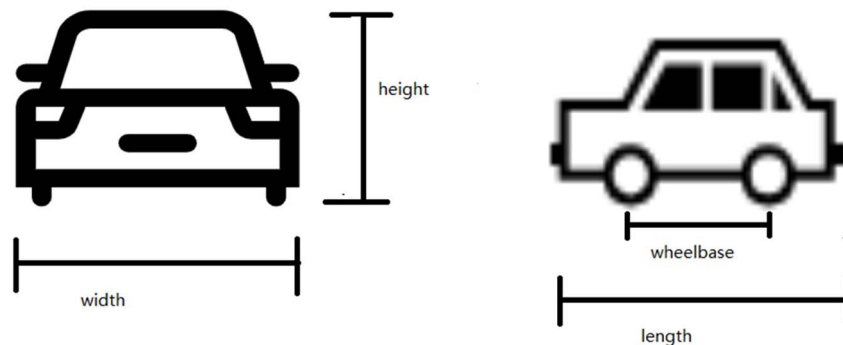


**Figure 4.7-1 car front and side view**

Due to the demand for voice services, modern cars are increasingly being equipped with more microphones, especially electric vehicles. One popular microphone placement strategy involves positioning microphone arrays on one side of the car roof. Other configurations include centralized placement and distributed placements.

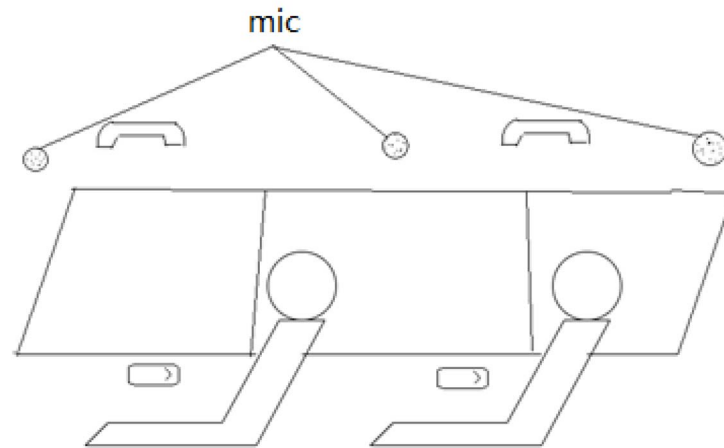


**Figure 4.7-2 microphone placement in car**

Some models with exterior and interior size are listed in Annexes A.7 and A.8.

## 4.8 Conclusion

Due to varying structures and sizes in each UE type, different microphone array designs (placement, distance between microphones) are possible. These various microphone integrations in UEs will be discussed in clause 7.

---

# 5 Microphones used in immersive audio capture

## 5.1 Introduction

The function of a microphone is to convert a sound pressure signal into an analog electrical signal within a circuit. This section describes four types of microphones that are popular in the market. These microphones have unique advantages for immersive audio capture in UE. They are classified as dynamic microphone, condenser microphone, MEMS and contact microphone.

## 5.2 Transducer type

### 5.2.1 Dynamic microphone

The dynamic microphone is one of the popular microphones on market. The main advantage of a dynamic microphone for UE is that it doesn't require external power; therefore, the entire recording system may become simpler. Another advantage is its durability, which makes it more suitable for loud and high-pressure situations. However, it usually has a disadvantage in that it is less sensitive to high frequencies.

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

### 5.2.2 Condenser microphone

The condenser microphone is another popular microphone on the market, especially for immersive audio. Most immersive systems use condenser microphones, such as ambisonic microphones and external stereo microphones for mobile phone. They are popular for their high sensitivity, wide frequency response, and 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 considering the channel number of immersive audio.

A condenser microphone uses a 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, which then changes the capacitance and influences the electrical signal.

### 5.2.3 MEMS microphone

Over the past few decades, the microphone for UE has changed from carbon microphones to electret condenser microphones. Recently, the MEMS microphone has spread rapidly, benefiting from its advantages of high stability and small volumes.

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

This means for manufacturers; it is much easier to build the capture system as the MEMS microphone can output the digital signal directly. On the other hand, it necessitates the careful selection of components. Since the microphone is much smaller and very uniform in its mechanical properties, it's suitable for UE and makes immersive audio possible for economically portable UE like mobile phones.

### 5.2.4 Contact microphone

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

Compared to the acoustic microphones, contact microphones have the advantage of not capturing sound waves in the air, but rather capturing mechanical vibrations of the target object. Hence, they are resistant to noise in air.

Nowadays, the 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.

## 5.3 Directional Microphone

### 5.3.1 Introduction

Directivity plays a very crucial role in immersive audio, and every immersive audio format has requirements for directivity. Even for object-based audio, it is necessary to consider the directivity to mitigate the impact of environment noise.

### 5.3.2 Directional microphone capsule

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

### 5.3.3 Interference tube

The interference tube, typically used on shotgun microphones, makes them more directional than a typical cardioid or supercardioid microphone. This long, narrow extended tube is placed in front of the microphone capsule and features multiple small holes along its length. It creates phase shift for sounds arriving from off-axis directions. Consequently, the off-axis sound arrives at the diaphragm with varying phase relationships, causing them partially cancel each other out.

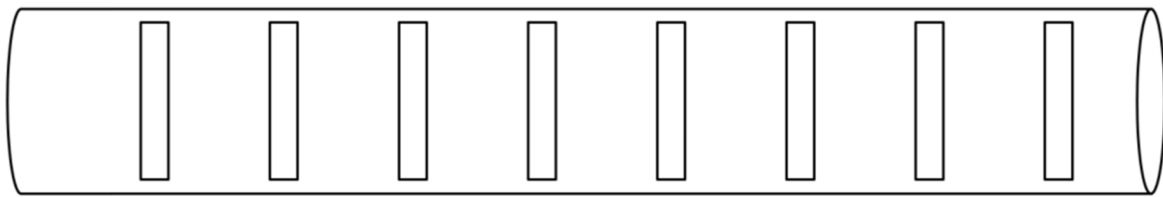


Figure 5.3.3-1 The schematic diagram of interference tube

## 5.4 Binaural acoustic simulation

According to the principle of binaural signal, the typical solution involves placing microphones on each ear of the user, or on a model of a human head or ears to capture the binaural cues. This model could be a head with a torso to simulate all influences, including those from the ear, head and reflections from the torso. Alternatively, it could be a single head or simply a model with a pair of ears.

## 5.5 Conclusion

Following advancements in microphone devices, more possibilities have been granted for UEs integrated with the miniature microphones to capture audio signals. Specifically, the digital MEMS microphone, which includes preamps, ADC, and clocks, can directly output digital signals, making audio capture significantly more convenient. This innovation is particularly beneficial for small-sized devices due to its size and consistent performance. These miniature microphones can also be used in combination with signal processing to generate immersive audio.

---

# 6 Immersive audio capture format

## 6.1 Stereo capture

### 6.1.1 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 a sound image with spaciousness, direction and a sense of depth for listeners. This can be reproduced through headphones or loudspeakers.

### 6.1.2 Characteristic of stereo capture

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

### 6.1.3 Factors that affect stereo capture

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

In the past, discussions regarding the factors that affect stereo capture primarily focused on microphone properties such as directionality and frequency range, as well as the placement of microphones.

With advancements in audio processing, there are now more methods available for controlling audio signals, which hold significant promise for stereo applications. This is particularly relevant since UE imposes strict restrictions on hardware due to space constraints. The ability to fine-tune audio signals through processing offers immense potential for enhancing stereo performance despite various limitations. However, it may also introduce more influence on audio

experience, which needs to be carefully analysed. Therefore, acoustic design also needs to take into account the characteristics of relevant processing.

## 6.1.4 Stereo microphone configurations

### 6.1.4.1 Introduction

Stereo microphones can generally be classified as spaced, near coincident, baffled, and coincident. This classification is based on the directional characteristics of the microphones, as well as their angle and distance from each other. The following sub-clause presents some of these configurations.

### 6.1.4.2 Near-Coincident

Near-Coincident involves using two directional microphones placed closely together at an angle 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 a number of near coincident configurations which are arranged at specific angles and distances. One of the well known near coincident stereo microphones is the ORTF configuration. In the ORTF configuration, a pair of microphones is arranged to mimic the placement of human ears. It uses two cardioid microphones that are 17cm apart and at a  $110^\circ$  angle from each other.

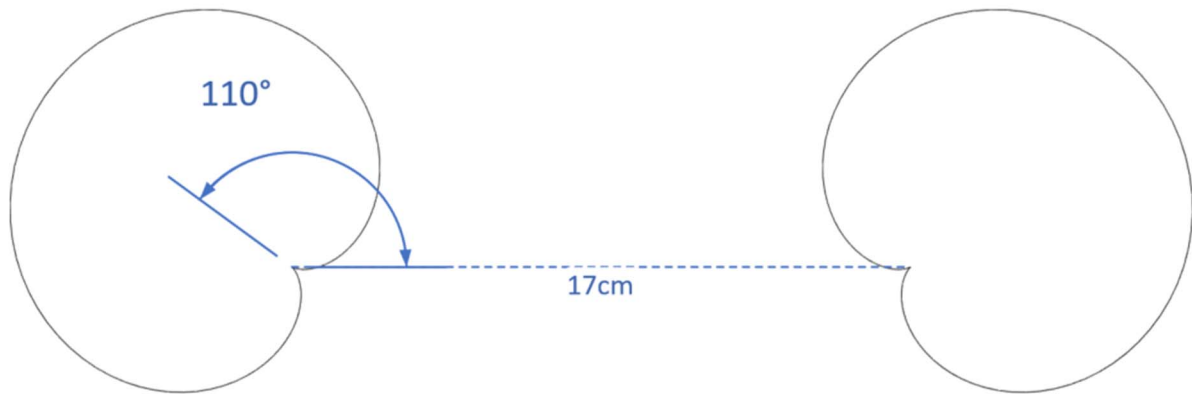


**Figure 6.1.4.2-1 The configuration of ORTF stereo microphone**

### 6.1.4.3 Baffled

A baffled configuration is a stereo recording technique that utilizes an acoustic baffle to enhance the separation between the left and right audio channels. The baffle is typically a physical barrier placed between the two microphones. One example of baffled microphone setup is the OSS method. This 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  $20^\circ$ . The capsules of the two microphones are positioned 16.5 cm apart.

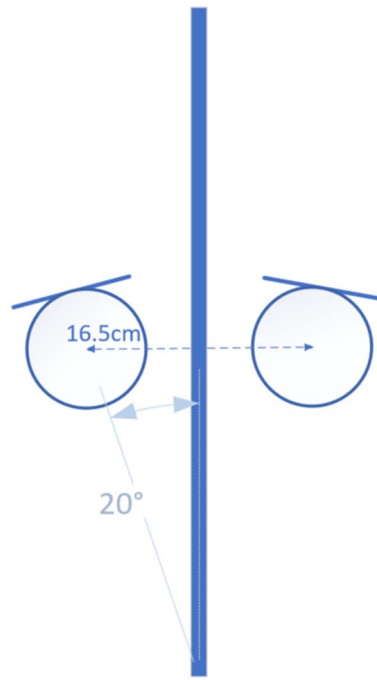


**Figure 6.1.4.3-1 The configuration of OSS stereo microphone**

#### 6.1.4.4 Coincident

##### 6.1.4.4.1 Introduction

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

##### 6.1.4.4.2 X/Y

X/Y stereo microphone commonly uses two cardioid microphones ranging from 90° to 135°.

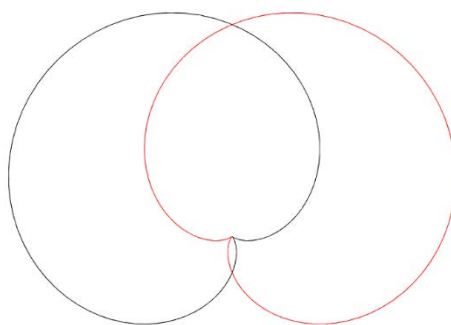


**Figure 6.1.4.4.2-1 The configuration of X/Y stereo microphone**

##### 6.1.4.4.3 Blumlein

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

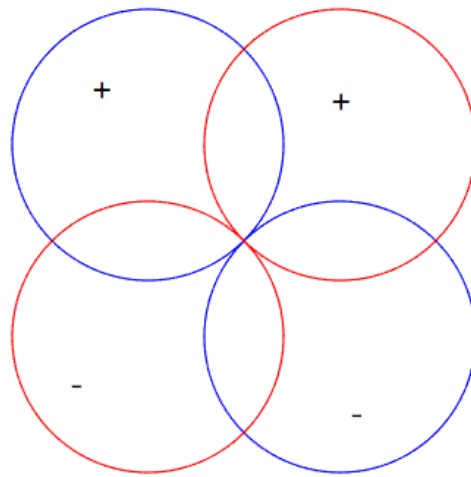


**Figure 6.1.4.4.3-1 The configuration of Blumlein stereo microphone**

#### 6.1.4.4.4 M/S

A M/S stereo microphone uses a forward-pointing microphone (usually a cardioid) and a bidirectional (figure-eight) microphone oriented perpendicular to the directional microphone. The figure-eight microphone captures the side signal, while the cardioid microphone captures the mid signal. Therefore, the left and right channel signals can be obtained through the simple addition and subtraction.

$$Left = mid + side$$

$$Right = mid - side$$

Additionally, by controlling the ratio of the two signals, different angles can be achieved.

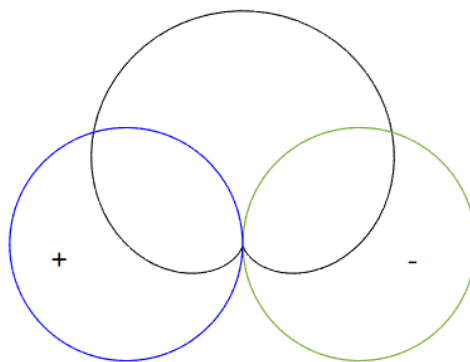


**Figure 6.1.4.4.4-1 The configuration of M/S stereo microphone**

#### 6.1.4.5 Spaced

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

This technique leverages the distance between two microphones to create a time difference and level difference between the left and right channels. This difference is due to the variation in the arrival time of sound waves at each microphone and the absorption of sound by the air between the microphones.

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

## 6.2 Spatial audio capture

### 6.2.1 Introduction

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

### 6.2.2 Binaural capture

#### 6.2.2.1 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 capturing these signals in the ears of a listener, it can retain both timbre and spatial aspects, even preserving the personal feature in binaural. And it can be accurately reproduced through headphones.

#### 6.2.2.2 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 cases, it is difficult to place the microphone precisely at the entrance of the ear canals. Therefore, it may be helpful to determine what factors influence binaural capture, so that a better signal can be obtained under limited conditions.

#### 6.2.2.3 Factors that affect binaural capture

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

Earbuds typically have transducers positioned at the entrance of ear canals for playback. This location is crucial for binaural recording, and the microphone needs to be set a few millimetres outside the ear canal entrance. The surface of the earbud may also cause reflection. It is evident that any reflection from the pinnae captured by the microphone will be affected.

#### 6.2.2.4 Differences between binaural and stereo audio

Both binaural and stereo formats consist of two channels, left and right. Several differences are outlined in Table 6.2.2.4-1.

**Table 6.2.2.4-1: 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 rendering |
|--------|------------------------------------------|--------------|------------------|--------------------------------------------------------------------------|-------------|--------------------|
|        |                                          |              |                  |                                                                          |             |                    |

|                                                                                                                                                                                   |                                |                                                                                                                      |             |                                                                                                                       |                                     |             |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------|----------------------------------------------------------------------------------------------------------------------|-------------|-----------------------------------------------------------------------------------------------------------------------|-------------------------------------|-------------|
| <b>Stereo</b>                                                                                                                                                                     | 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 |
| <b>Binaural</b>                                                                                                                                                                   | 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 is transformed to be positioned between left and right virtual loudspeakers. |                                |                                                                                                                      |             |                                                                                                                       |                                     |             |

## 6.2.3 Parametric spatial audio capture

### 6.2.3.1 Principle of parametric spatial audio representation

Spatial or immersive audio representations generally enable reproduction of audio scenes, allowing the listener experience to optimally correspond with the actual recorded situations and environments. This means, for example, 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 through headphones, including head-tracked binaural rendering, or a loudspeaker setup that provides sufficient spatial capability.

Parametric spatial audio describes a spatial sound field using a parametric representation. In a typical solution, a multi-microphone capture is compressed into a small number of audio channels and associated spatial parameters. These parameters define the perceptually relevant properties of the sound field, such as directional information and the degree of diffusion the sound field, while the audio channels define the actual energetic representation of the captured sound field.

The analysed 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 then applied to the transport audio channels. To enhance the quality of the reproduction, the parameters can be analysed based on the desired auditory frequency bands.

The typical processing flow of the parametric spatial audio capture is illustrated in the Figure 6.2.3.1.-1.

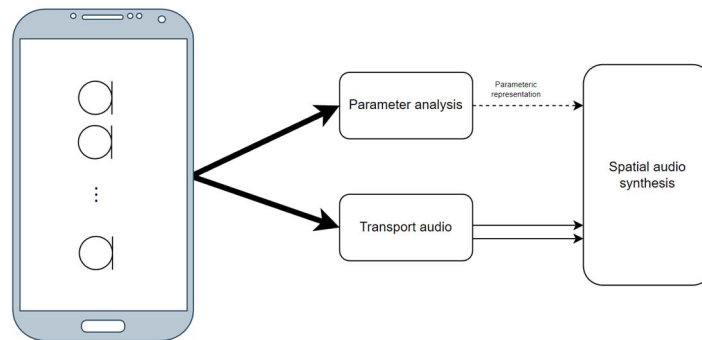


**Figure 6.2.3.1-1 Overview of parametric spatial audio capture**

A prominent example of a parametric spatial audio representation is the MASA format, as defined in 3GPP TS 26.258 [3]. Specifically, MASA consists of either one (mono) or two (stereo) transport audio signals, along with metadata.

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

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

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

### 6.2.3.3 Factors that affect parametric spatial audio capture

Multiple factors may affect the quality and accuracy of parametric spatial audio capture. These factors include:

- The dimensions of the capture device, specifically the distance and placement of the microphones
- The number and the characteristics of the microphones, such as directivity and frequency response.
- The signal processing techniques applied to the captured multi-microphone signals, for example, noise suppression and filtering
- The quality of the spatial parameter analysis algorithm and the device specific tuning

Moreover, due to these factors, the quality and the accuracy of the parametric spatial capture may not be consistent across all the capture directions. This inconsistency is highly dependent on factors such as the specific device form and its microphone placement/spacing.

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

The number and the configuration of the microphones can be arbitrary. However, in principle, the minimum 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, the minimum number for microphones in this case is 4.

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

## 6.2.4 Non-parametric spatial audio capture

### 6.2.4.1 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 a 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 challenging to generate the spatial audio representation directly from their embedded acquisition units or even from selected accessory devices. The common solution is to use a microphone array to catch raw signals and then perform mathematical processing to output the expected results.

An example processing flow of the non-parametric spatial audio capture is referred to in Figure 6.2.4.1-1

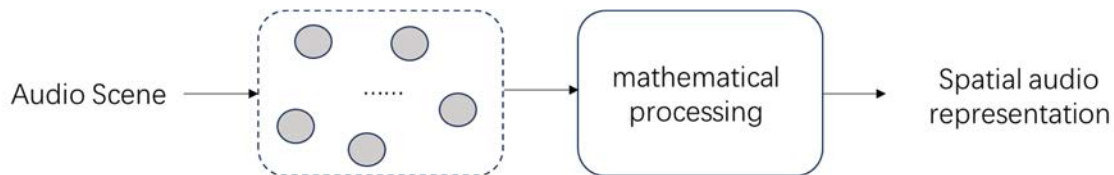


**Figure 6.2.4.1-1 Overview of non-parametric spatial audio capture**

### 6.2.4.2 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, which are used as IVAS codec input formats, include surround (5.1 and 7.1), surround + height (5.1+4 and 7.1+4), FOA, HOA2, HOA3, and Object-based audio. The standard format audio is a necessary part of the interoperable solution between different types of end-user devices.

Non-parametric spatial audio capture and representation serves as an important intermediate link between the originally captured raw signals at sending end and the rendered spatial signals at receiving end. This can allocate computational complexity of the end-to-end real-time spatial audio solution to both ends.

The accuracy of the non-parametric spatial audio representation can significantly vary due to the corresponding non-parametric spatial audio capture solution. Therefore, 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.

### 6.2.4.3 Factors that affect non-parametric spatial audio capture

They are the same as parametric spatial audio capture; refer to section 6.2.3.3

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

Immersive audio capture technology using microphones has been developed over decades; however, its corresponding microphone configuration is not suitable for current mobile phones.

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

#### 6.2.4.4.1 Immersive audio ORTF configuration [4]

##### 6.2.4.4.1.1 ORTF-surround

The "ORTF-surround" configuration evolved from the ORTF stereo technique and consists 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 80° and 100°. The output from each microphone is individually routed to the corresponding Left (L), Right (R), Left Surround (LS), and Right Surround (RS) speakers, creating an immersive audio experience. Refer to Figure 6.2.4.4.1.1-1.

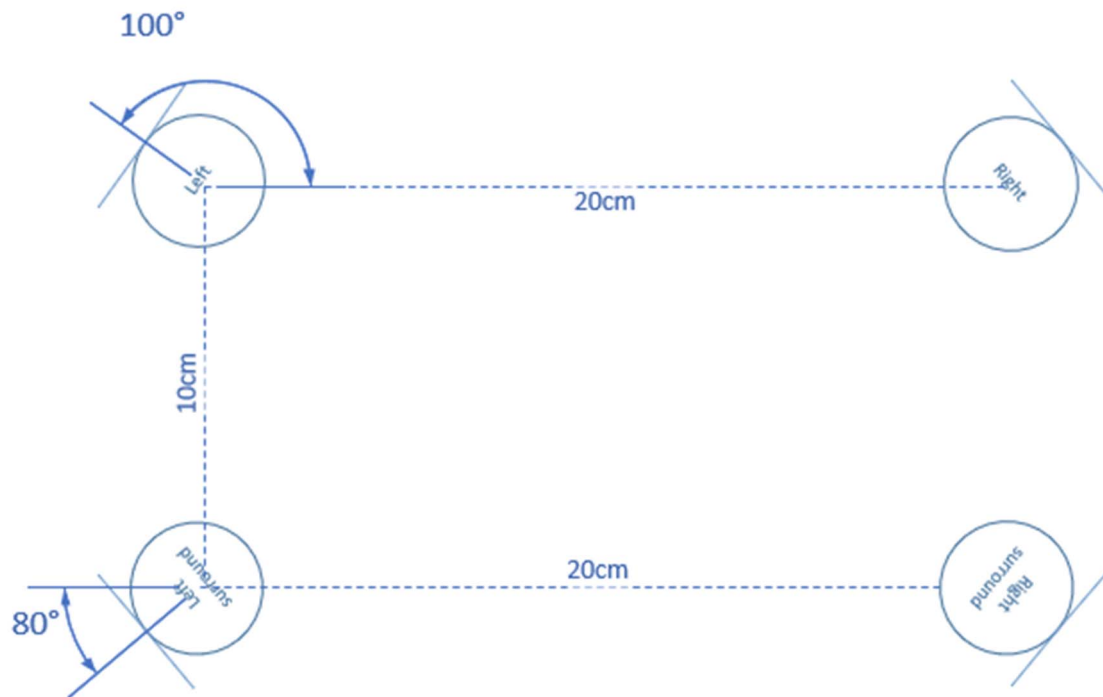


**Figure 6.2.4.4.1.1-1 The configuration of ORTF-surround microphone**

#### 6.2.4.4.1.2 ORTF-3D

The "ORTF-3D" consists of two "ORTF-surround" configuration, one of which is placed directly on top of one another with a 90° elevation angle. Refer to Figure 6.2.4.4.1.2-1.

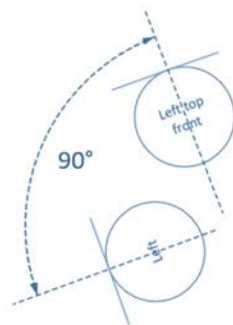


**Figure 6.2.4.4.1.2-1 Left view of ORTF-3D microphone front channels**

#### 6.2.4.4.2 Immersive audio M/S configuration

##### 6.2.4.4.2.1 Double-M/S[5]

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. The corresponding channel signal can be obtained through the following equation, as refer to in Figure 6.2.4.4.2.1-1.:

$$Center = mid$$

$$Left = mid + side$$

$$Right = mid - side$$

$$Left\ surround = midrear + side$$

$$Right\ surround = midrear - side$$

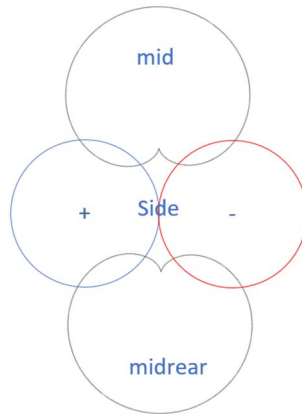


**Figure 6.2.4.4.2.1-1 The configuration of Double-M/S microphone**

#### 6.2.4.4.2.2 M/S-3D[6][7]

Incorporating a vertically oriented figure-8 microphone as a “Z” signal into the "Double-M/S" configuration allows the "M/S-3D" setup can capture the height channel, as referred to in Figure 6.2.4.4.2.2-1.:

$$Left\ top\ front = mid + side + Z$$

$$Right\ top\ front = mid - side + Z$$

$$Left\ top\ rear = midrear + side + Z$$

$$Right\ top\ rear = midrear - side + Z$$

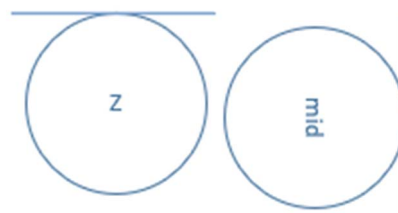


**Figure 6.2.4.4.2.2-1 Left view of M/S-3D microphone**

#### 6.2.4.4.3 IRT-cross

Another well-known configuration is the IRT Cross, which is an equally segmented 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 6.2.4.4.3-1.

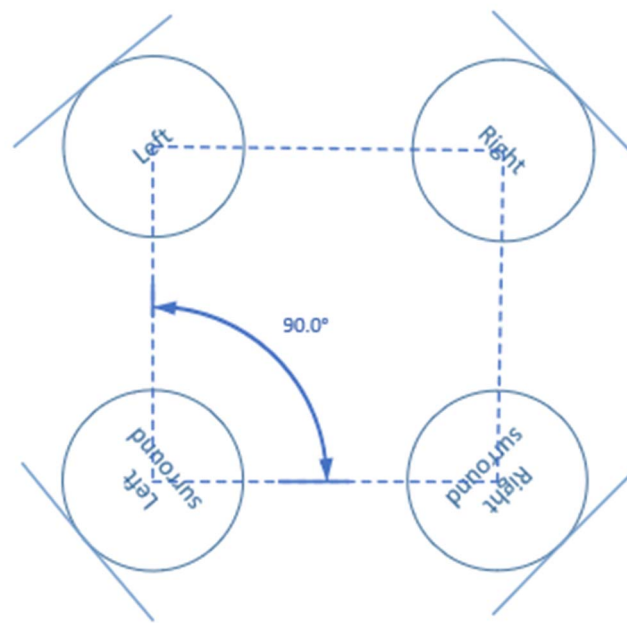


**Figure 6.2.4.3-1 The configuration of IRT-cross microphone**

#### 6.2.4.5 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. This is based on the analysis of the raw microphone signals, which are used to produce parameter metadata and potentially convert 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 through certain signal processing.

### 6.3 Summary

Principles, configurations, and audio quality factors for achieving immersive capture were discussed. Stereo capture involves recording two audio signals to create a sound image with spaciousness and direction. Spatial capture uses at least two microphones to accurately reproduce audio scenes with spatial cues, such as direction and elevation, for immersive listening.

---

## 7 Microphone integration in UEs

### 7.1 Microphone integration in mobile phones

#### 7.1.1 Modern mobile phones

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

The spacing between the microphones in a multi-array, their polar pattern, and the number of microphones varies from device to device. However, the spacing generally ranges from 3cm to 17cm. The following table 7.1-1 covers the specifications of these microphones in mobile phones.

Table 7.1-1: MEMS Microphone example characteristics for Multi-array microphones in mobile phone.

| 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 (on the left and right side of USB C-Port), while the larger spacing is achieved with a combination of top and bottom bezel microphone and/or rear facing microphones. As the user can operate and hold the device in various ways, a combination of these microphones can be activated for stereo voice communication. Efficient capturing configuration is achievable with advancements in audio processing. This involves the fine-tuning of audio signals through pre-processing, which offers great potential for enhancing stereo performance in different user end consumption scenarios, such as loudspeaker and/or headphones.

7.1.2 Foldable mobile phones

As stated in clause 5 MEMS microphones are the standard for mobile phones. The number and placement of microphones on foldable devices (including both book-style and clamshell designs) vary by manufacturer and model. Basic configurations in the foldable segment includes 3 microphones (positioned at the top, bottom bezel and rear), while more advanced models offer up to 6 microphones (with configurations of 2 microphones at the top, 2 at the bottom bezel, and up to 2 rear-facing). An increased number of microphones can significantly enhance audio quality during calls, contributing to a better overall user experience. Given the flexible design of foldable devices, various usage modes can be expected (example: full-screen, half-folded at a 90-degree angle, single-handed mode, desk mode, etc.), therefore more attention may be needed when processing captured signals from multi-microphones positioned strategically to support high-quality audio capture. Following figures provide example microphone configurations for book-style and clamshell-style design.

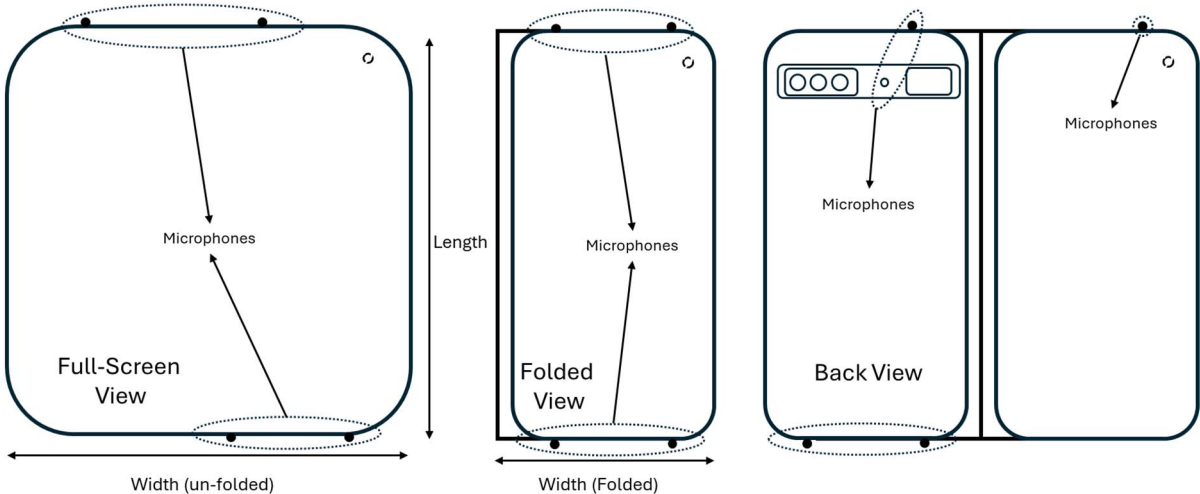


Figure 7.1.2-1: Example book-style foldable mobile device microphone configuration

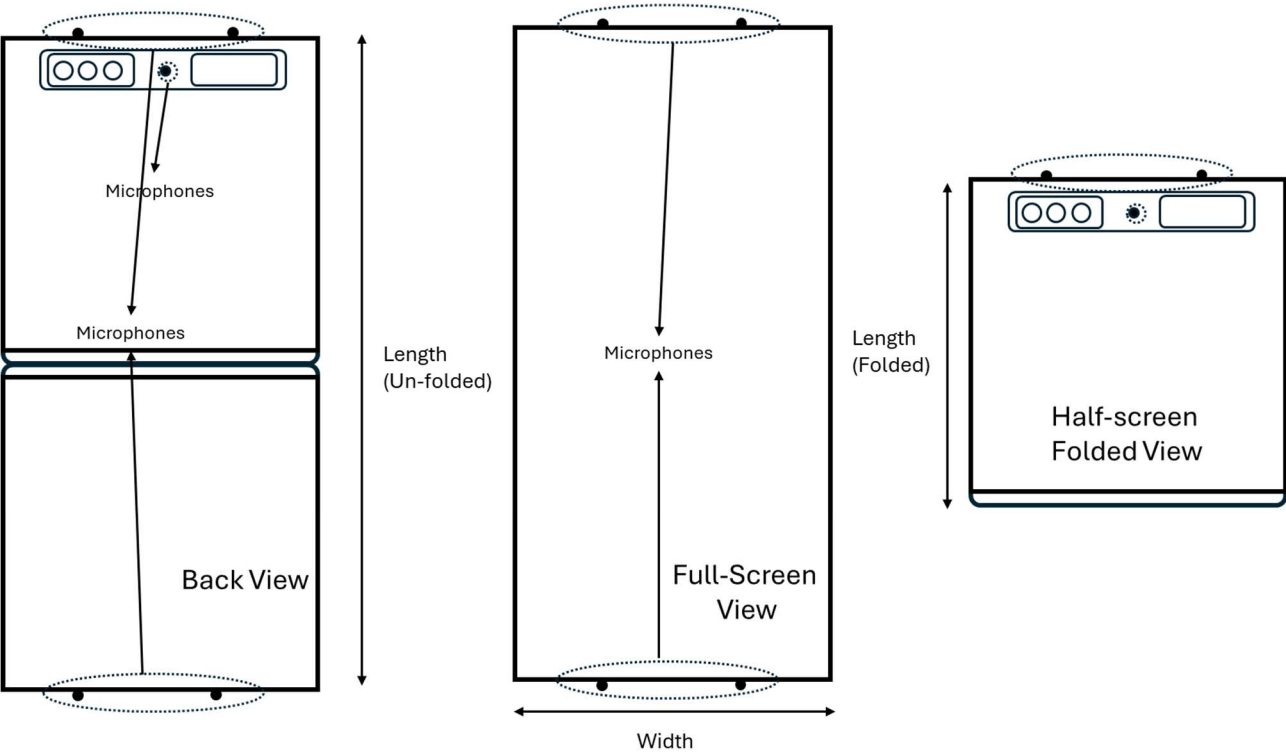


**Figure 7.1.2-2: Example clamshell-style foldable mobile device microphone configuration**

Table 7.1.2-1 presents example characteristics of multi-microphone configurations for foldable devices. Due to the flexible design of these devices, various usage modes may alter the spacing between microphones. As a result, processing solutions may need to adapt to the adjustable multi-microphones configuration to deliver high-quality immersive audio quality.

**Table 7.1.2-1: MEMS Microphone example characteristics for multi-microphones in foldable mobile phones.**

| Foldable Mobile Phone       | Number of Microphones | Spacing    | Polar Pattern    |
|-----------------------------|-----------------------|------------|------------------|
| Book-Style (folded)         | Up to 6 microphones   | 2 to 17 cm | Omni-Directional |
| Book-Style (un-folded)      | Up to 6 microphones   | 2 to 18 cm |                  |
| Clamshell-Style (folded)    | Up to 6 microphones   | 2 to 4 cm  |                  |
| Clamshell-Style (un-folded) | Up to 6 microphones   | 2 to 17 cm |                  |

## 7.2 Microphone integration in headphones

Generally, headphone form factors can be divided into three categories over-ear, on-ear, and in-ear. Nowadays, wireless headphones are quite popular. These devices are often connected to a mobile phone or a portable device via Bluetooth both for both speech communication and audio streaming. They are equipped with MEMS microphones, supporting more than one microphone. The number and placement of microphones per device category vary from manufacturer to manufacturer, with the goal of delivering better sound quality and spatial experiences.

Despite the form-factor for in-ear headphones being comparatively smaller compared to on and over-ear devices, in-ear devices incorporate up to 6 microphones (three per bud), on the other hand, on-ear devices incorporate up to nine microphones, with a major focus on improving user experience. Tables 7.2-1 and 7.2-2 cover the specifications of these microphones for on-ear, over-ear, and in-ear headphones.

**Table 7.2-1: MEMS microphone example characteristics for multi-array microphones for on-ear or over-ear devices**

| Number of Microphones per pair | Placement of Microphones                                                      | Polar Pattern   |
|--------------------------------|-------------------------------------------------------------------------------|-----------------|
| up to 9* microphones           | up to 3 external microphones per side distributed.<br>on the outside ear cup. | Omnidirectional |
|                                | 1 internal microphone per side                                                |                 |
|                                | w/wo 1 dedicated microphone for voice (right cup)                             |                 |

9\* For some headphones (on-ear), up to 9 microphones are available, while others may have fewer than 9 microphones.

**Table 7.2-2: MEMS microphone example characteristics for multi-array microphones for in-ear devices**

| Type | Number of Microphones per pair | Placement of Microphones per side                                                | Polar Pattern                |
|------|--------------------------------|----------------------------------------------------------------------------------|------------------------------|
| Pod  | Up to 6*                       | Bottom of the stem (1), Top of the stem (1)<br>w/wo Inward facing microphone (1) | Omnidirectional<br>/Cardioid |
| Bud  |                                | 2 mics located on the outer part of the bud<br>w/wo one inward facing microphone |                              |

6\* Some headphones may incorporate one or two additional bone conduction microphone(s).

The general tendency among wireless headphone manufacturers primarily focuses on delivering better noise cancellation. To achieve this, larger microphone arrays up to 4 per side are used.

## 7.3 Microphone integration in tablets

Tablets share a similar structure with mobile phones but are significantly larger in size. This allows for more flexibility in the placement of microphones, with some tablets accommodating up to five microphones. Small microphone arrays are also a common feature on tablets. However, due to the less common handheld mode in tablets, microphones near the USB port - the closest location to the mouth in handheld mode - are rarely seen. Instead, microphones are typically located near the front and back camera modules and along the edges of the tablet, as referred to in table 7.3-1.

**Table 7.3-1 MEMS microphone example characteristics for multi-array microphones for tablets**

| Number of Microphones | Placement of Microphones                              |
|-----------------------|-------------------------------------------------------|
| up to 5 microphones   | 1-3 on the top of the devices (near the front camera) |
|                       | 1 on near the back camera                             |

|  |                         |
|--|-------------------------|
|  | 1 on the edge of device |
|--|-------------------------|

## 7.4 Microphone integration in laptops

Due to the clamshell structure of a laptop, there are more options for microphone placement in current devices. Some are positioned on the screen, while others are located on the keyboard section. More potential microphone location result in a larger microphone distance and moveable microphone location. Some popular microphone placements on laptops are explained in following figures:

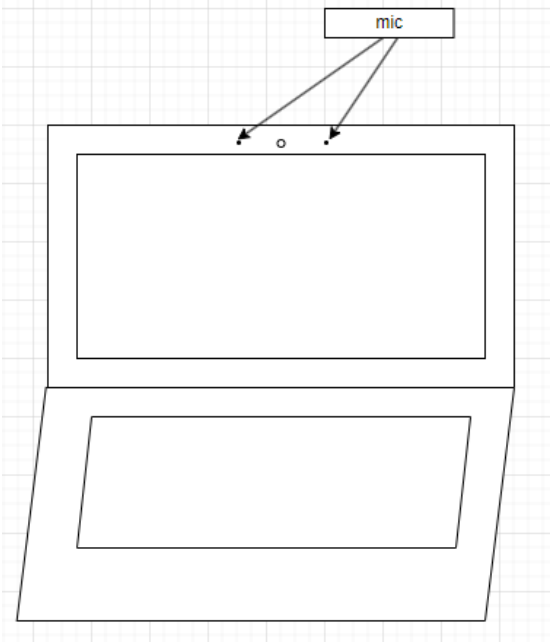


Figure 7.4-1 Laptop microphone placement 1

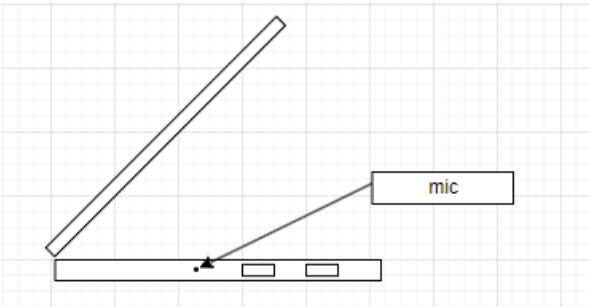


Figure 7.4-2 Laptop microphone placement 2

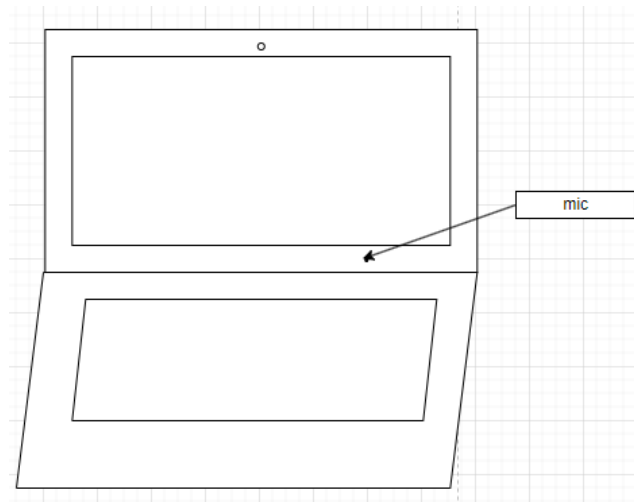


Figure 7.4-3 Laptop microphone placement 3

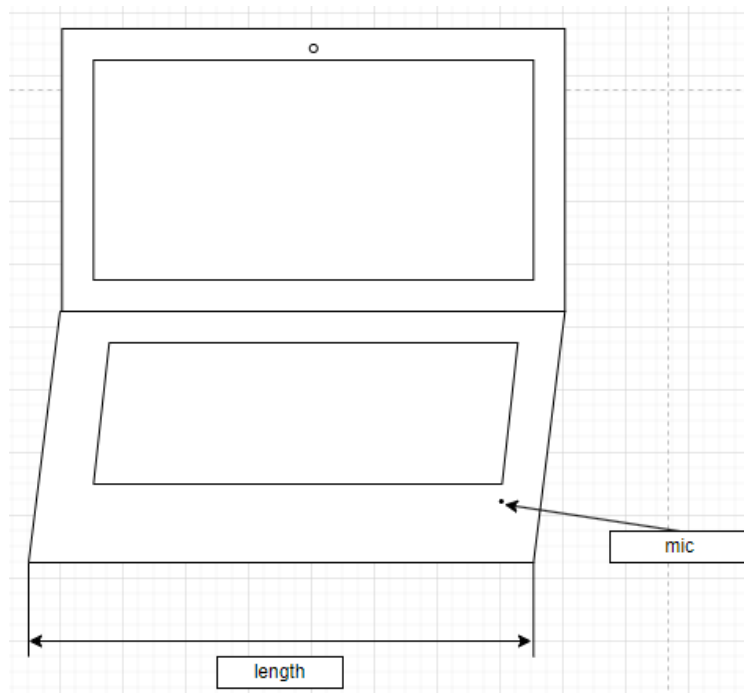


Figure 7.4-4 Laptop microphone placement 4

## 7.5 Microphone integration in watches

Modern watch devices are increasingly being equipped with microphones and speakers, enabling them to establish speech services independently. Some contemporary models often feature two built-in microphones. Typically, one microphone is located on one edge of the watch, while the second is situated on the opposite edge. The spacing between these two microphones generally correlates with the size of the watch, ranging from 3 to 5cm.

## 7.6 Summary

MEMS microphones are the default choice for the UEs discussed in the above sub-clauses. Each UE category typically includes a minimum of two microphones, enabling multi-channel capture, which may offer improved audio quality, better spatial awareness, enhanced noise suppression, and sound localization compared to a mono microphone. However, to achieve optimal sound quality with a multi-microphone setup, several factors may be considered, such as microphone placement, type, characteristics, user interaction and handling, etc.

Some of the UE designs have narrow spacing between microphones, which may result in poor spatial characteristics and may not provide the best sound quality for multi-channel capture. For UE categories such as mobile phones, users can hold and operate their devices in various ways, which can obstruct microphones and affect audio quality, stereo imaging, or spatial awareness. Additionally, variations in hardware quality across different devices and UEs can also lead to inconsistent audio performance, even within the devices of the same UE brand.

To fully benefit from 3GPP IVAS codec for the UEs of various categories, recommendations per UE category such as spacing between microphones, type, characteristics, etc., will be needed without restricting the manufacturer's flexibility in UE hardware design.

## 7.7 Microphone integration and frequency response

The MEMS microphone package includes a port that allows external sound to enter. The transducer is shielded by a can, which may be perceived as chambers. These components form a Helmholtz resonator, resulting in resonances. Numerous studies in this field have provided various formulas to calculate such resonators. A rise in specific frequencies can often be observed in numerous MEMS microphones, exhibiting a curve like resonance.

Microphone capture is defined by the electroacoustic properties of the microphone transducer and the acoustic properties of the microphone integration. Microphone integration on mobile devices is typically product-specific, and different products have variation between microphone assemblies due to differences in product mechanics and electronics. Mobile phones use MEMS microphone technologies that are typically either top port or bottom port type, depending on the product mechanics. The choice of MEMS microphone type influences the microphone port geometry. In mobile phone design, microphones in the top and bottom ends can be integrated more similarly to each other, as shown in figure 7.7-1. When microphones are integrated on other surfaces, such as next to the main camera, the design needs to use different mechanical structures.

In addition to differences in microphone integration, the acoustic diffraction from rigid body of the product will influence the overall frequency response of the microphone. Acoustic diffraction is highly dependent on the direction of the sound source due to the non-uniform shape of the mobile phone body.

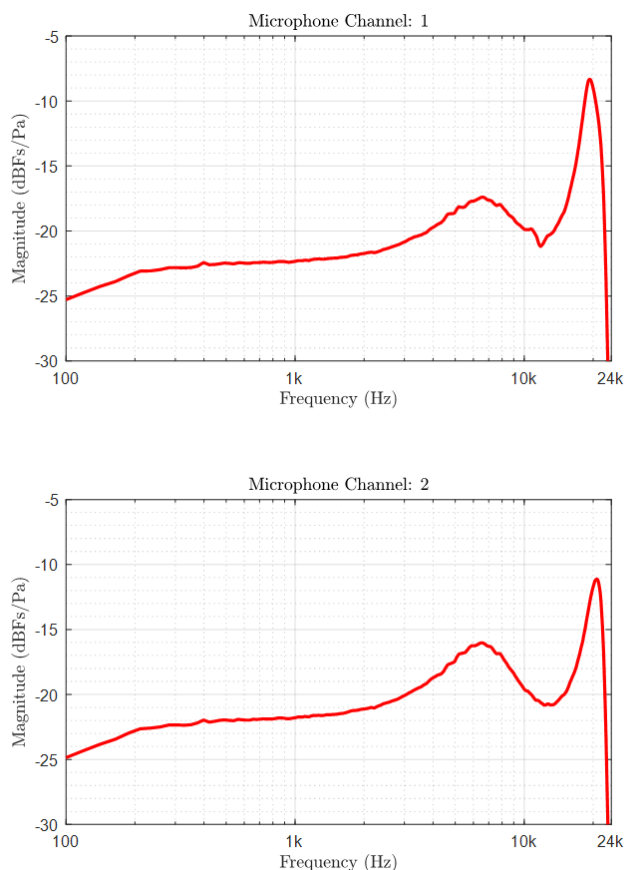


**Figure 7.7-1. Integrated microphone frequency responses (incl. port and transducer) for stereo microphone integration in both ends of a commercial mobile phone device.**

In general, it is unlikely for a mobile device, especially with more than 2 microphones, to have similar microphone frequency response characteristics for all microphones. Two mobile devices with 4 microphones were analyzed, indicating differences between typical mobile phone microphone integrations (see figure 7.7-2). Microphone frequency responses, illustrated in figure 7.7-2 and 7.7-3, provide examples where resonances are clearly audible and need to be compensated with microphone signal equalization to enable high-quality audio capture.

Accurate frequency response data for equalization filter design can be created using multiple techniques, such as acoustic measurements in an anechoic chamber, the use of lumped parameter component models, or the application of numerical acoustic simulation tools. The numerical acoustic methods require detailed acoustic port geometry, which can include information about dust and liquid protection filters or membranes. These filter materials have an acoustic impedance that is considered in the acoustic design process. Usually, the acoustic characterization of microphone integration is done, or at least verified, using acoustic measurements, because equalization relying on inaccurate frequency response data can lead to further issues.

Due to these practical limitations in product design, it would be unrealistic to assume that pressure on the device surface, specifically at microphone port inlet, could be used as an approximation of the pressure on the integrated microphone transducer diaphragm. For instance, if this would be the case, the frequency responses, as presented in Figure 7.7-1, Figure 7.7-2 and Figure 7.7-3, would all be flat.

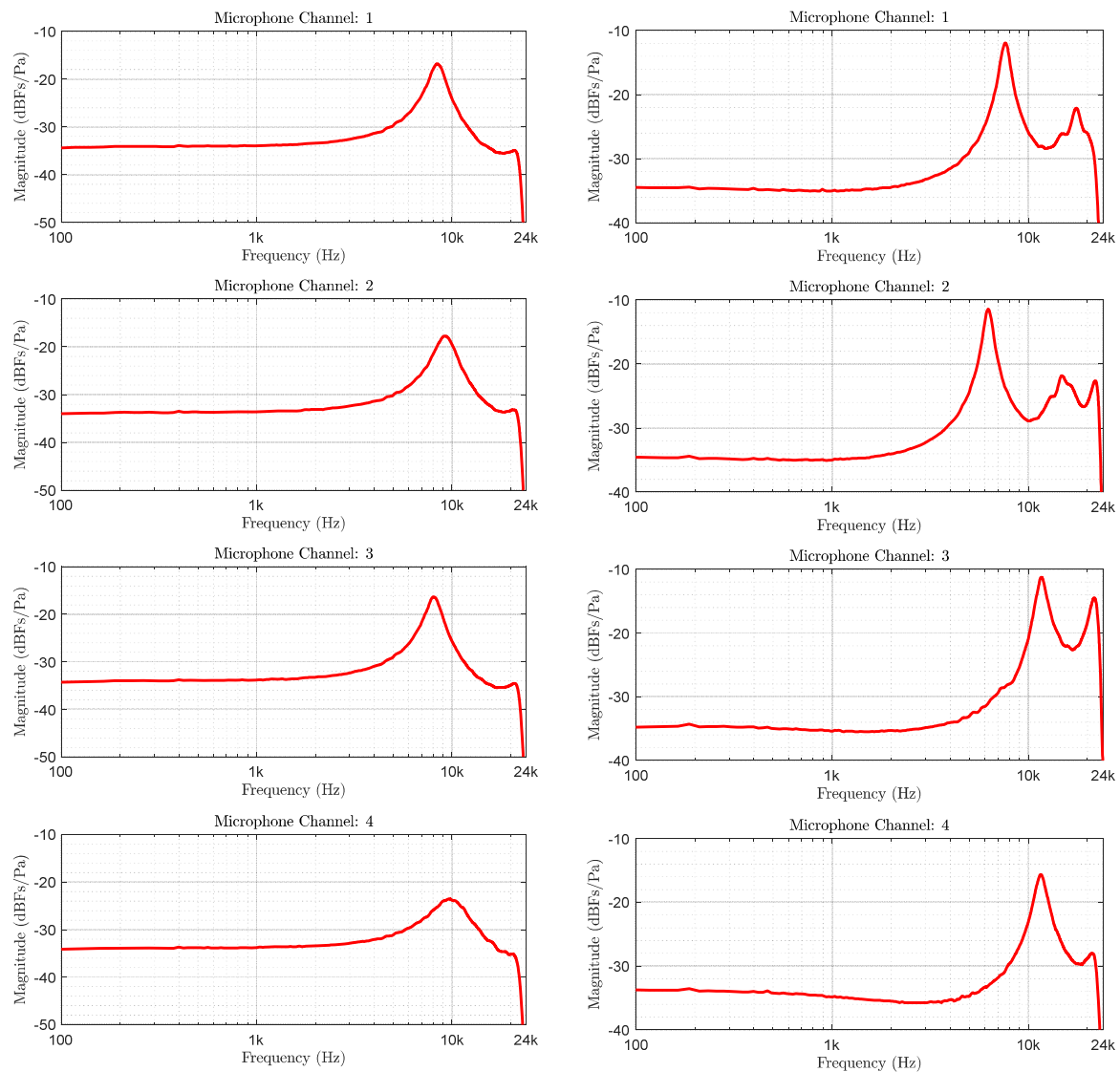


**Figure 7.7-2. Integrated microphone frequency responses (incl. port and transducer) for two commercial mobile phone devices with 4 microphones.**

In microphone integration it is generally beneficial to aim for simple mechanical structures that implement a uniform acoustic channel from the surface of the device to the microphone sensor, without unnecessary bending, cavities, and other variation in the cross-sectional area of the sound channel. This is commonly achieved when the microphone component is integrated close to the device's surface, minimizing the length of the sound port relative to its cross-sectional area. Since the hole in the device cover is typically desired to be either small or visually hidden, the acoustic targets of the design are generally best achieved when the length of the microphone port is kept short.

Robust microphone operation throughout the lifetime of a mobile device requires sealing of the acoustic channel. Sealing and IP protection filters are commonly integrated for simplified assembly. The acoustic impedance of various acoustic filter materials may be included in the numeric analysis of the acoustic channel, unless the impedance is low enough to be insignificant.

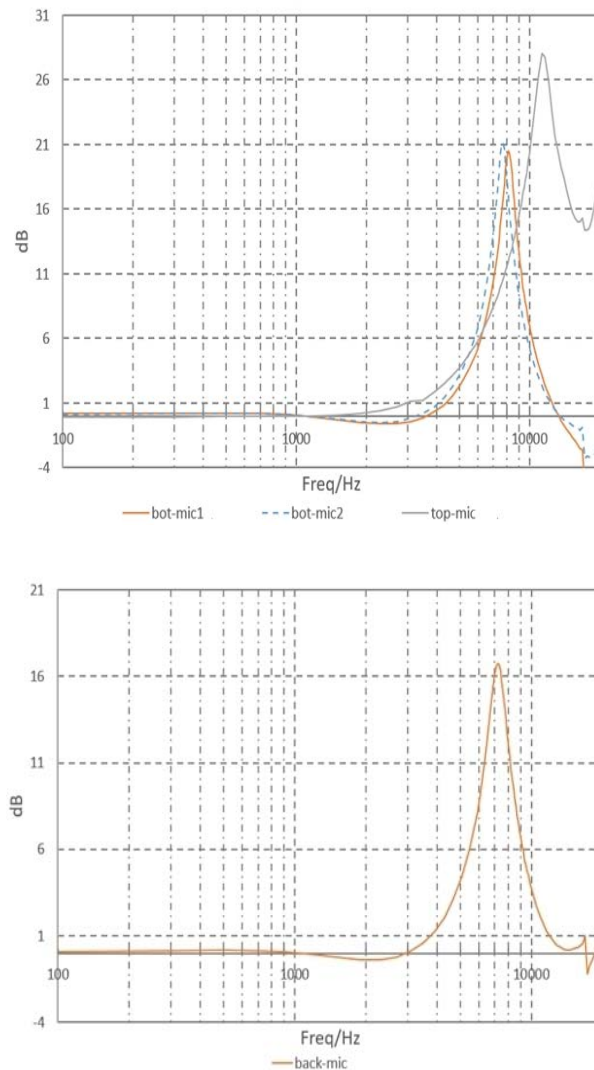


**Figure 7.7-3. Integrated microphone frequency responses (incl. port and transducer) for one commercial mobile phone devices with 4 microphones.**

All microphone integrations include some acoustic filtering effects, which typically include resonance. Favourable microphone integration can be characterized by parameters that may be similar for all microphones: microphone resonances occur at high frequencies ( $>10\text{kHz}$ ), and the resonance magnitude is small with low Q factor. When all microphones of the device are integrated using similar good design practices, it provides a good basis for multi-microphone capture algorithms.

Acoustic scattering from the rigid body of the device significantly impact spatial audio capture algorithms for multi-microphone devices. It can be said that free-field approximation using microphone positions will produce inaccurate results, particularly at high frequencies.

As part of audio design, acoustic measurements can be performed in an anechoic chamber to measure and analyse the frequency responses of each integrated microphone as a function of source direction. This is technically a relatively straightforward measurement task, for example, using a turntable to measure multiple source (azimuth) directions in a 360-degree plane. The 3D measurements, including azimuth and elevation directions, require a short series of measurements with different device orientations, or turntable measurements with multiple sound sources. Alternatively, frequency responses can be simulated using numerical acoustic methods.

The frequency response of the raw microphone signal typically appears mostly flat with a pronounced rise in certain high frequency bins, usually exceeding 10dB. To enhance sound quality, selecting the pronounced rise to higher frequency bins can minimize its perceptual distortion.

## 8 Signal processing

### 8.1 AEC

#### 8.1.1 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 illustrated in Figure 8.1.1-1, where a speaker uses a sending device with voice communication capability to play an audio clip (represented by 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 (represented by blue colour waveform signal) for the sending audio clip. If an adaptive filter can be designed such that its output signal is the exact opposite of the echo signal, as represented by the green colour waveform signal, the next step would be to superimpose the blue colour waveform signal with the green colour waveform signal. Consequently the echo signal is eliminated depending on the performance of the adaptive filter.

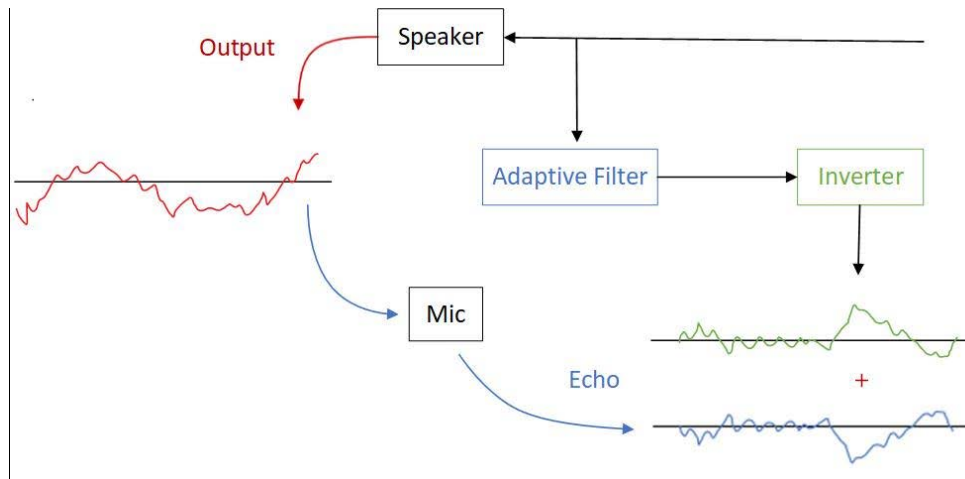


**Figure 8.1.1-1: diagram of AEC**

AEC utilizes 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:

$$e(n) = d(n) - y(n)$$

$$y(n) = \sum_{i=0}^{M-1} w(i)x(n-i)$$

$$w(n+1) = w(n) + \mu e(n)x(n)$$

Where,  $x(n)$  is the input signal,  $d(n)$  is the desired signal,  $y(n)$  is the estimated echo signal,  $e(n)$  is the error signal,  $w(n)$  is the coefficient of the adaptive filter,  $\mu$  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, ultimately reducing the error signal in an acceptable range.

Note: Terminal acoustic characteristics for telephony requirements were specified in TS 26.131[11]

#### 8.1.2 Challenges for immersive audio AEC

For immersive audio services, 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.,  $N \times M$ . This is several times the number of echo paths generated

by a mono audio communication system. Taking stereo audio as an example, a stereo audio device uses two speakers and two microphones to create a two-way audio service would need to estimate four echo paths in a Stereophonic Acoustic Echo Cancellation setup.. The significant increase 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 variations in speakers, moments, and positions of speech. The adaptive filtering algorithm for mono audio case is not suitable for this situation. Clearly, directly extending mono audio echo cancellation algorithm to multi-channel results in poor acoustic echo cancellation performance.

### 8.1.3 The current status of the research

The core of the multi-channel acoustic echo cancellation algorithm is the same as that of single-channel echo cancellation algorithm, that is adaptive filtering algorithm. However, multi-channel acoustic echo cancellation presents more challenges than single-channel AEC, primarily due to the non-uniqueness of the solution. The strong correlation of multi-channel input signals can 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 primarily 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 a faster convergence speed and lower computational complexity in multi-channel situations.

Example of stereo-AEC is referred to Annex B

### 8.1.4 AEC for different UEs

Theoretically, enhancing AEC performance can be achieved by increasing the distance between the microphone and loudspeakers, which reduces the echo level coupling from the loudspeakers to the microphones. However, the need for additional microphones and loudspeakers due to immersive audio significantly reduce the distance.

This reduction is particularly noticeable in smaller UE, such as mobile phones and in-ear headphones, where the compact size restricts the distance, resulting in a higher echo level and shorter echo delay. In some devices, the microphone and loudspeaker may even be placed in the same sound port.

On the other hand, larger UE, such as cars, may experience a longer echo delay. However, they provide a more stable acoustic environment as all the microphones and speakers for communication are installed within the cabin.

## 8.2 Microphone Array Signal Processing on device

### 8.2.1 Introduction

Clause 6 outlines various audio capture formats applicable to UEs listed in clause 4. However, generating suitable input signals for IVAS encoder modes from the UEs raw microphone signals necessitates MASP as a crucial step. This involves converting raw microphone signals into the desired audio representation, often with necessary enhancements, to align with IVAS encoding modes.

Evaluating the applicability of each IVAS encoding mode per UE category is essential to ensure optimal performance. Table 8.2.1-1 provides an overview of the 3GPP IVAS encoding modes, ranging from mono and stereo to advanced configurations like scene-based audio (SBA), metadata-assisted spatial audio (MASA), and independent streams with metadata (ISM). Most UE categories feature at least two microphones, enabling multi-channel audio capture. However, achieving high-quality results from such setups depends on factors like microphone placement, characteristics, and overall system complexity, which can significantly influence the deployment of IVAS encoding modes.

Although hardware quality varies across UE categories, the minimum number of raw microphone signals are as follows: mono requires one signal, stereo/binaural requires two, and advanced modes like MASA, ISM requires a minimum of X signals. The complexity of IVAS encoding modes was not taken into consideration for the consideration of advanced coding modes as the hardware quality even within the same category pose challenges to uniform deployment.

**Table 8.2.1-1 IVAS Encoding Modes**

| <b>Coding modes</b> | <b>Full Name</b>                                                   |
|---------------------|--------------------------------------------------------------------|
| Mono                | EVS/IVAS Mono                                                      |
| Stereo              | Stereo Operation                                                   |
| SBA                 | Scene-based Audio (SBA, Ambisonics) Operation (1-3)                |
| MASA                | Metadata-assisted Spatial Audio (MASA) Operation (1 or 2 channels) |
| ISM                 | Objects (Independent Streams with Metadata, ISM) Operation         |
| MC                  | Multi-Channel (MC) Operation (5_1, 7_1, 5_1_2, 5_1_4, 7_1_4)       |
| OMASA               | Combined Objects and MASA (OMASA) Operation                        |
| OSBA                | Combined Objects and SBA (OSBA) Operation                          |

Clause 7 discusses microphone configurations for various UE categories:

- Mobile phones and tablets: These UEs can feature up to six microphones, supporting configurations like mono, binaural, stereo, and multi-channel capture. They are capable of IVAS coding modes such as mono, stereo, ISM, SBA, and MASA. However, additional processing, such as metadata generation and defining encoding formats per capture mode (handsfree, handheld, half-folded at a 90-degree angle, etc.), is required for compatibility with IVAS. Identifying these processing steps is outside the scope of this study.
- Headphones: With up to nine microphones, headphones can support binaural or mono capture. The most suitable IVAS modes depend on the number and placement of microphones, typically resulting in binaural or mono encoding.
- Laptops: These devices may include up to three microphones, allowing mono (one microphone), stereo (two microphones), or multi-channel (three microphones) capture. With more than two microphones, laptops can support IVAS modes like ISM, SBA (2D), and MASA.
- Watches: Limited to two microphones, watches can support mono (one microphone), stereo, or binaural (two microphones) capture.

The scope of MASP discussed in subsequent sub-clauses focuses on channel-based, binaural, scene-based, MASA, and object-based capture formats, emphasizing their applicability to specific UE types trying to clarify some of the below list, while leaving implementation and detailed MASP steps out of scope.

- Spacing between microphones per UE category for achieving immersive audio capture.
- Number of raw audio microphone inputs needed to convert to 3GPP IVAS MASA format. Similarly for ISM, SBA.
- Given that each UE has different dimension and microphone spacing, clarification on suitable IVAS coding modes per UE category is needed. Signal processing steps (from UE's raw multi-channel capture to IVAS coding mode) required to produce MASA, ISM, SBA signals, etc.
- Steps to generate the metadata for the IVAS coding format.
- Suitable coding format per capture mode.

## 8.2.2 MASP for Channel-based

### 8.2.2.1 MASP for Stereo

In Clause 4.1, it is stated current mobile phones have 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 phone. Therefore, this subchapter will focus on how to complete microphone array signal processing for producing stereo signals based on such a microphone array.

### 8.2.3 MASP for Binaural

According to clause 6.2.2, 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.

NOTE: When a binaural microphone is used for capturing, support might be needed for binaural to stereo conversion if it for loudspeaker consumption at the receiver side by appropriate signaling information.

### 8.2.4 MASP for Scene-based

#### 8.2.4.1 FOA

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

#### 8.2.4.2 Matrix on current FOA microphones

The current FOA microphones utilize a tetrahedral configuration, which comprises four cardioid microphones oriented in the directions of FLU, FRD, BLD, and BRU. The W, X, Y, Z components are produced through matrix multiplication of the four cardioid signals, as refer to equation(1). They do not follow the SN3D and ACN channel order.

$$\begin{bmatrix} W \\ X \\ Y \\ Z \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} * \begin{bmatrix} FLU \\ FRD \\ BLD \\ BRU \end{bmatrix} \quad (1)$$

#### 8.2.4.3 HOA

In ambisonic microphones, a minimum of  $(N + 1)^2$  microphones is required to capture N order HOA. Consequently, a second-order HOA necessitates at least 9 channels of ambisonic microphones. The number of microphones on most mobile phones is limited to 6. Therefore, obtaining HOA microphone array on current microphones configuration of mobile phones may be difficult. However, there are still some relevant research on upmixing FOA signals to HOA signals. For instance, the first solution is to interpolate the FOA to HOA signal, as suggested by various studies refer to [12][13][14]. The second solution is to utilize a generative model capable of translating FOA into HOA signal, as referred to in [15]. Lastly, some signal-dependent technologies, such as parametric spatial audio, have the capability to analyse the captured signals and subsequently convert them into an excepted HOA format [16].

## 8.2.5 MASP for MASA

MASA format signals consist of both audio signals and metadata. The metadata, as referred to in 26.258[3], are derived from the analysis of microphone raw signals. Therefore, microphone array signal processing is a crucial module for producing MASA signal.

## 8.2.6 MASP for Object-based

According to 26.250 [10], Object-based audio consists of 1-4 individual mono object streams, each with associated metadata. Various existing technologies can be utilized to obtain object-based audio.

### 8.2.6.1 Mono object stream

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

### 8.2.6.2 Associated Object Metadata

A minimal set of associated object metadata is the object's position in the polar coordinate system, described using azimuth  $[-180^\circ, 180^\circ]$  and elevation  $[-90^\circ, 90^\circ]$  angles.

DOA is commonly utilized in current audio services 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.

## 8.2.7 MASP for OMASA

It is for further study

## 8.2.8 MASP for OSBA

It is for further study.

## 8.3 Beamforming

### 8.3.1 Introduction

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 for achieving target directivity. Although it's mostly used for mono speech now, it has great potential in immersive audio. There are also many studies in this area.

This proposal begins with two fundamental technologies: Delay-sum and differential. It aims to find a suitable solution for immersive audio on UE.

### 8.3.2 Delay-sum microphone

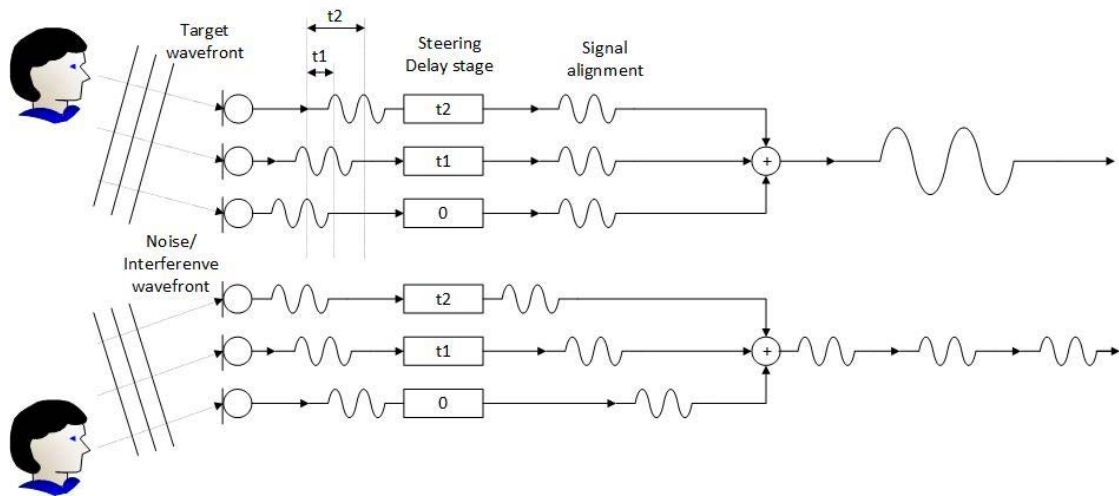


**Figure 8.3.2-1 The diagram of Delay-sum microphone array**

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 signals from the desired direction have the same phase so that they can be reinforced.

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

### 8.3.3 Differential microphone array

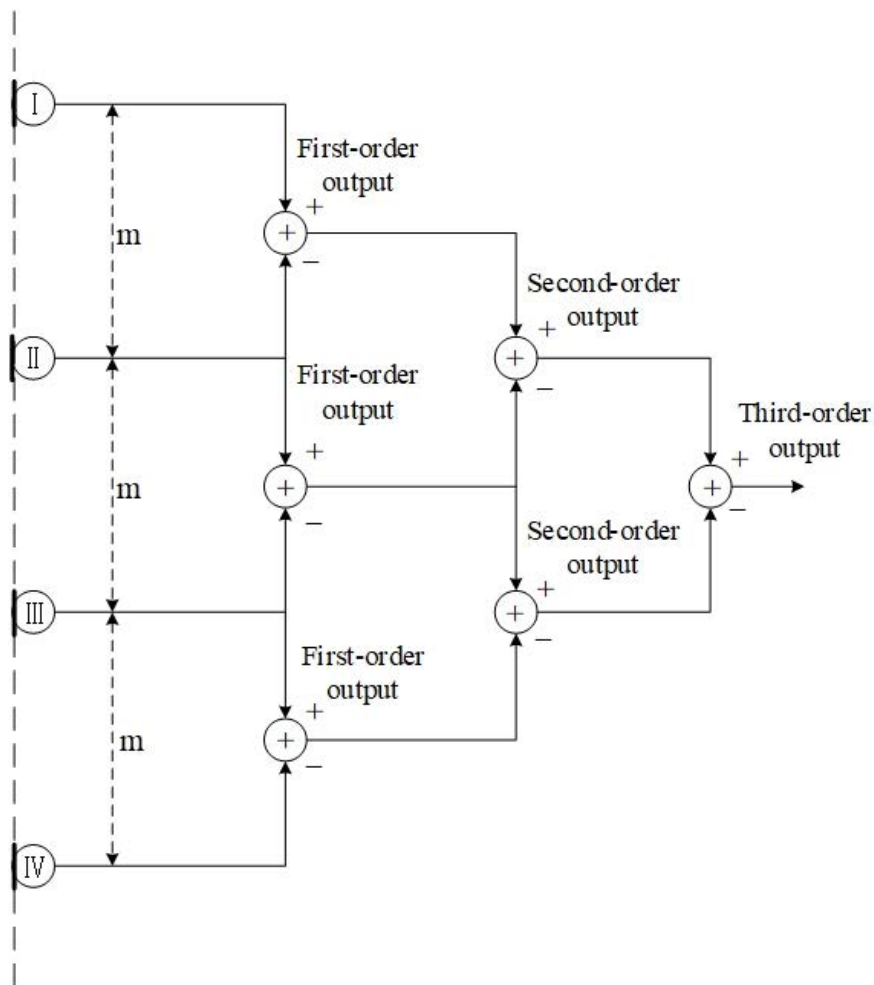


Figure 8.3.3-1 The diagram of Differential microphone array

In DMA, signals from two or more microphones are subtracted from each other to create a unique directivity. The traditional directional microphone can be viewed as a special type of differential beamforming.

By adjusting the weight and phase of the differential signal, various directivities can be achieved, such as: cardioid, bidirectional (Figure-8), supercardioid, hypercardioid, and subcardioid (wide cardioid).

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

## 8.4 Noise reduction

The needs of noise reduction depend on certain usage scenario or sometimes specific solutions. For example, in an office scenario, which mainly consists of speech components, noise reduction is highly necessary. However, in a concert application scenario, which includes various music instruments and other atmospheric audio components, noise reduction may not be necessary. In this case, moreover, audio scene recognition may be also needed to identify specific scenarios

## 8.5 Conclusion

The raw microphone signals are used to generate the expected format audio signals using MASP to get satisfactory audio signal quality. Some signal processing modules, such as AEC and NS, may be needed for achieving minimum quality of experience.

NOTE: Headphones with higher capabilities often come with built-in processing capabilities. These process capabilities may also be combined with UE processing. As a result, the UE terminal may still need to perform some audio processing tasks to ensure that the audio quality meets certain requirements. Regardless of the capabilities of the equipment, a minimum quality may need to be defined for ensuring suitable captured signals.

---

# 9 Example audio capture processing solutions

## 9.1 Capture Scenarios

### 9.1.1 Telephony communications

**Capturing Type:** Multi-Microphone Capturing

**Summary**

A Call was established between Tom and Harry.

Tom's device has a 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. He changes the orientation of the portable communication device from portrait to landscape.

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

**User Story:**

Tom planned vacation with his friends, and they are now at the Bhutan airport. One of Tom's friends, Harry, who was supposed to be part of the travel group, had to drop out of the vacation at the last minute. Harry felt devastated when he called Tom to break the news. To lighten up Harry's mood and ensure he didn't feel left out, Tom came up with an idea of sharing their vacation experiences 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 rest of the friends stood at the observatory deck of the airport, he decided to share his experiences from day-0 (flights landing and takeoff, airport ambience, picturesque mountains in the backdrop etc..) with Harry. Tom extended his hand, holding his communication device in landscape mode towards the flight landing and taking off with beautiful mountain view in the background. Harry could now view and listen to the airport observatory deck scene clearly, which brought a smile to his face, as he felt he was 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 (Mobile phone, 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 becoming more common on mobile phones.

**Potential Processing Solutions**

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

## 9.2 End-user device (UE) characteristics/prerequisites

### 9.2.1 General

Support of immersive communication necessitates the appropriate integration of electroacoustic transducers, specifically microphones and loudspeakers. Current mobile phones, characterized by a large display and a slim body, present a challenging shape for spatial audio capture devices. The shape of a device influences its spatial capture characteristics, and spatial capture typically necessitates parametric audio capture techniques or suitable modeling to ensure uniform performance from various sound source directions.

Modern mobile phones typically possess at least 2 microphones, which are sufficient to facilitate immersive voice and audio capture for a broad range of use cases and applications.

Immersive voice and audio playback can be based, for example, on rendering for headphones or external loudspeaker systems. However, stereo and immersive playback can also be supported also on the mobile device itself. The number of loudspeakers and their locations within the device, in conjunction with the rendering processing, significantly impact the perceived immersion of the loudspeaker playback. Loudspeaker placement relative to the microphone locations influences both the (spatial) audio capture performance and requirements of the AEC solution.

### 9.2.2 Number and placement of microphones

Spatial audio capture requires at least two microphones, but for a rich capture of audio scenes, using more microphones is recommended. Two microphones can be used to support stereo recording, and it may be advisable to implement, for example, symmetrical microphone integration and placement at the ends of the device to achieve a symmetric stereo image.

Spatial audio capture, supporting 360-degree spatial capture in a 2D plane, can be supported with 3 microphones. It is common for stereo recording microphones to be complemented with a third microphone integrated on the back cover of the device. This supports spatial audio capture aligned with the optical axis of the main camera.

Practical aspects related to microphone placement include the detailed mobile device form factor (geometry), which is related to how the device is used, and how user's hands might interfere audio capture, for example, by blocking or disturbing any of the microphone inputs. The acoustic performance of 360-degree capture is dependent on microphone placement to ensure robust direction analysis for different sound source directions.

For immersive voice and audio use cases that require higher spatial fidelity and flexibility, 4 or more microphones may be considered for supporting a full 3D spatial audio capture in both landscape and portrait device orientations. Multiple microphones can improve the spatial resolution of audio capture and increase robustness in different usage environments, e.g., they can effectively reduce disturbing wind noise.

Maintaining a sufficient distance between microphones and loudspeakers can improve AEC performance by reducing the echo level coupling from loudspeakers to the microphones. More details on loudspeaker placement are provided in the following.

### 9.2.3 Loudspeaker number and placement of loudspeakers

In order to produce immersive (stereo or spatial) playback over loudspeakers integrated into a mobile device, at least two loudspeakers are required. When playback is active, The number and placement of the loudspeakers also affect the spatial audio capture when playback is active, e.g., due to interaction with AEC.

Two loudspeakers in a mobile device allow for direct stereo content playback and, with suitable rendering processing, can provide an adequate level of immersive experience for the listener's front. It is straightforward to achieve a wider stereo image by extending the physical loudspeaker locations, but creating a real feeling of immersion around listener's head, so that sounds are perceived as coming from behind, is very challenging.

Using similar speaker components and placing the speaker outlets symmetrically in the device generally provides the best basis for a high-quality playback. It is beneficial to have the loudspeaker outlets at the device ends so that when listened to in landscape mode, the stereo image will be at its widest. If the loudspeaker ports are facing the user, the spatial audio quality can be further improved, and the audio performance is more consistent and independent of device handling, specifically interaction effects with user's hands and grip. More than two loudspeakers can be used, especially

in larger mobile devices, e.g., in tablets and foldable mobile phones, to provide favourable stereo loudspeaker pairs at different usage orientations and device configurations.

## 9.3 Capture solution for end-user devices

### 9.3.1 Overview

End-user device's support for immersive voice and audio services requires the successful combination of several audio technologies and appropriate product design for taking full advantage of the new immersive capabilities. Relevant devices (UE) can vary in many shapes and sizes (form factors) for different use cases. Even traditionally dominant UE form factors, such as mobile devices, are expected to be utilized in new ways, for instance, multiple orientations (landscape and portrait) for various use cases and applications that offer immersive voice and audio communication.

Multi-microphone capture is used to get expected audio signals for end-user devices. The following figure 9.3.1-1 illustrates an example process of generating audio signals.

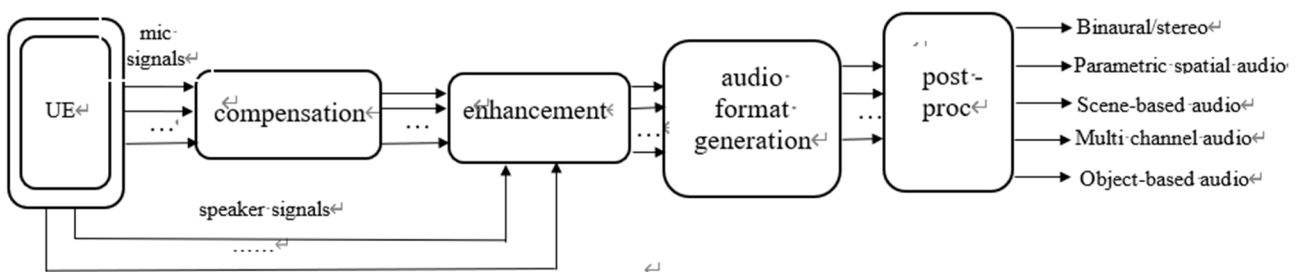


Figure 9.3.1-1 example process of generating audio signals.

### 9.3.2 Compensation

The compensation block is used to improve the quality of microphone signals, specifically to smooth the frequency responses and minimize the mismatch within microphone arrays. One of the classic solutions involves using the EQ filter to modify the responses of the microphone signals. Generally, it smooths the microphone response of each channel according to the measured microphone response, which makes the microphone signals comparable to each other.

### 9.3.3 Enhancement

#### 9.3.1 Introduction

Enhancement block consists of various operations aimed at improving audio quality, considering the targeted audio source or sources. These operations can include, for example, AEC, noise reduction, audio focusing, etc.

#### 9.3.3.2 General considerations

Modern UEs, such as mobile phones, typically come with a minimum of two microphones, which may be spatially placed to support stereo and/or spatial audio capture. Users may hold and operate their UEs in landscape (horizontal) or portrait (vertical) mode. Additionally, a user's hands might sometimes interfere audio capture. Based on user interaction with the UE, the audio capture settings can be adjusted accordingly to optimize the multichannel capture performance based on the intended use.

#### 9.3.3.3 AEC

AEC is typically performed separately for each microphone signal, but alternative solutions may also be considered. The goal of AEC processing is to remove the loudspeaker signal component from the microphone signals based on a reference signal. In the case of stereo playback, both stereo channels are needed as reference inputs for the AEC.

The baseline approach involves applying traditional Acoustic Echo Cancellation on the individual microphone channels. The traditional AEC solution has been to use a linear AEC filter, followed by residual echo suppression (RES).

Nowadays, RES is often implemented using DNN. Recently, solutions in which the entire AEC is managed with a single DNN have been introduced.

#### 9.3.3.4 Noise reduction

##### 9.3.3.4.1 Introduction

In monaural speech audio, the main emphasis is on capturing the speech signal. However, spatial audio offers the additional advantage of presenting the ambience. As a result, certain ambient sounds, which might be dismissed as background noise in monaural speech audio, are effectively conveyed in spatial audio. The interpretation of noise varies between monaural and spatial audio. Although device floor noise, such as microphone noise and wind noise, is still classified as noise, other non-speech sounds originating from the actual environment enhance the atmosphere and may be considered effective signals in spatial audio, depending on the scenarios.

##### 9.3.3.4.2 Wind noise reduction

The baseline approach involves applying wind noise reduction to the individual microphone channel through non-linear signal processing. A DNN based noise reduction/speech isolation, trained on wind noise, has also proven effective in reducing wind noise. While such a system performs well to maintain speech, it incurs higher processing costs and latency.

##### 9.3.3.4.3 Microphone noise reduction

Microphone noise reduction aims to remove self-noise generated by the microphones. It is generally based on defining the noise floor of the microphones.

##### 9.3.3.4.4 Background noise reduction

Background noise reduction can be used to enhance intelligibility of the captured speech. It generally makes audio more pleasant to listen to and can be used to remove irrelevant components from the captured audio. Depending on the use case, background noise reduction may only eliminate continuous noise such as air-conditioner or traffic noise, or it may aim to remove everything but speech from the captured audio. Content or context-based classification and processing can be utilized if different noise reduction processing is desired, for instance, for speech and music. For immersive voice and audio, background noise reduction settings may be more contextual than for traditional mono voice and audio. Noise reduction can also be necessary for accessibility purposes. DNN-based solutions are commonly used for noise reduction.

##### 9.3.3.4.5 Audio focusing

Audio focusing can be used to concentrate spatial audio capture in a preferred direction. The focus direction can be defined either automatically or manually, depending on the use case. Good focus performance can be achieved with suitably selected microphone locations using beamforming. Audio focusing may be used for content creation or accessibility purposes.

### 9.3.4 Audio format generation

#### 9.3.4.1 Introduction

The audio format conversion is used to convert the enhanced microphone signals into an expected audio signal format. The following clauses give example processing for mobile phone devices.

#### 9.3.4.2 Example of stereo processing

The stereo processing transforms the microphone signals into standard stereo audio. It employs two or more microphones on the mobile phone to create left and right channels. In the stereo audio, sounds coming from different angles present varied ICTD and ICLD. Concurrently, sounds from various angles present a phase difference between the microphones, which can be leveraged to modify ICTD and ICLD to different degrees. A transformation matrix could be constructed based on this information, and technologies such as beamforming may be utilized.

### 9.3.4.3 Example of scene-based audio processing

#### 9.3.4.3.1 Introduction

The example solution uses a content-based processing module and an ambisonics upmixer module to obtain scene-based audio signals.

#### 9.3.4.3.2 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 a 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.

#### 9.3.4.3.3 Ambisonic upmixer

This block is responsible for mapping the multi-microphone capture to first order ambisonics or higher. Among 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. These simulations 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, assuming a simpler 3D geometry of the device and microphones as point sinks.

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

### 9.3.4.4 Example of parametric spatial audio processing

#### 9.3.4.4.1 Introduction

The example solution uses a downmixing module and a spatial analysis module to obtain parametric spatial audio signals (e.g., MASA).

#### 9.3.4.4.2 Downmixing

With parametric spatial audio capture, spatial audio output can be synthesized from a number of audio channels (e.g., one or two) and metadata. Therefore, the number of audio channels can generally be reduced after the audio capture. This can be achieved, for example, by selecting one or two representative audio channels from the output of audio enhancement block.

#### 9.3.4.4.3 Spatial analysis

The aim of spatial analysis in parametric spatial audio capture is to estimate spatial properties of the captured audio signal (audio scene) and generate metadata. They can be later utilized in spatial synthesis or rendering. Spatial analysis uses information about the device shape and the microphone locations. It is common to conduct spatial analysis for frequency domain sub-bands, with analysis may be based on coherence and level analysis between the microphone signals. To ensure the effectiveness of spatial analysis, it is important that echo is properly removed from the captured signal (refer to AEC in clause 9.3.4.2). Additionally, it is important that audio enhancement processing (refer to clause 9.3.4) has preserved relevant spatial cues in the signals.

In an example spatial analysis, the direction of one or more dominant sound sources is analyzed for every sub-band. Furthermore, the directional energy and other spatial parameters are estimated. For example, the directional energy can be a ratio relative to the total energy of the corresponding sub-band. The analyzed parameters form the metadata for the parametric spatial audio format. The selected audio signals (e.g., mono or stereo) and the analyzed metadata can then be provided as input to spatial synthesis or used, for example, in audio transmission (as codec input).

### 9.3.5 Post-processing

Post-processing, in which typically the AGC is utilized, adjusts the signals to a suitable level. AGC typically also includes a limiter, which is designed to prevent signal saturation.

---

## 10 Conclusions and Recommendations

The TR covers various UE categories including mobile phones, headphones, tablets, laptops, watches, XR devices and cars. On the playback side, most UE categories can support stereo, and/or binaural audio through on-device loudspeakers or headphones. On the capture side, multiple microphones are available in the modern UEs and can also support immersive audio capture by deploying features, such as:

- 1) Equalization filter for all raw microphone signals
- 2) Microphone selection and/or configuration depending on capture mode.
- 3) AEC, noise reduction, MASP, beamforming.

Widespread on-device support of immersive audio capture functionality may help to accelerate the deployment of 3GPP IVAS. To address the needs for immersive audio capture, including solving key challenges in that operation, it is recommended to launch work to specify minimum performance requirements or performance objectives on raw microphone signals for immersive audio capture. The work is expected to be aligned with TS 26.260 and 26.261 that are related to sending side terminal audio quality performance requirements/objectives for Immersive Audio Services. Example solutions of converting raw and/or compensated microphone signals into at least one IVAS encoder input format are expected to be developed meeting these requirements/objectives.

NOTE: Mechanisms for negotiating UE immersive audio capture capabilities have not been considered and are left for further study.

---

## Annex A: UE size

### A.1 Mobile phone size

#### A.1.1 Modern mobile phone structure size

##### A.1.1.1 Introduction

Since 2012, the dimensions of mobile phones, including length, width, and thickness, have been increasing, indicating a continuous increase in mobile phone size. This may be due 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 basis for the increase in screen size.

##### A.1.1.2 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, indicating an upward trend. With the development of mobile phone models, some phones are no longer confined to the 16:9 aspect ratio. For instance, there are now models with 18.5:9 and 19.5:9 aspect ratios. Although high aspect ratio screens can display more information, most video content is still in the traditional 16:9 format. Therefore, an excessively high aspect ratio is not beneficial for video display.

The trend in mobile phone length in recent years is shown in Figure A.1.1.2-1

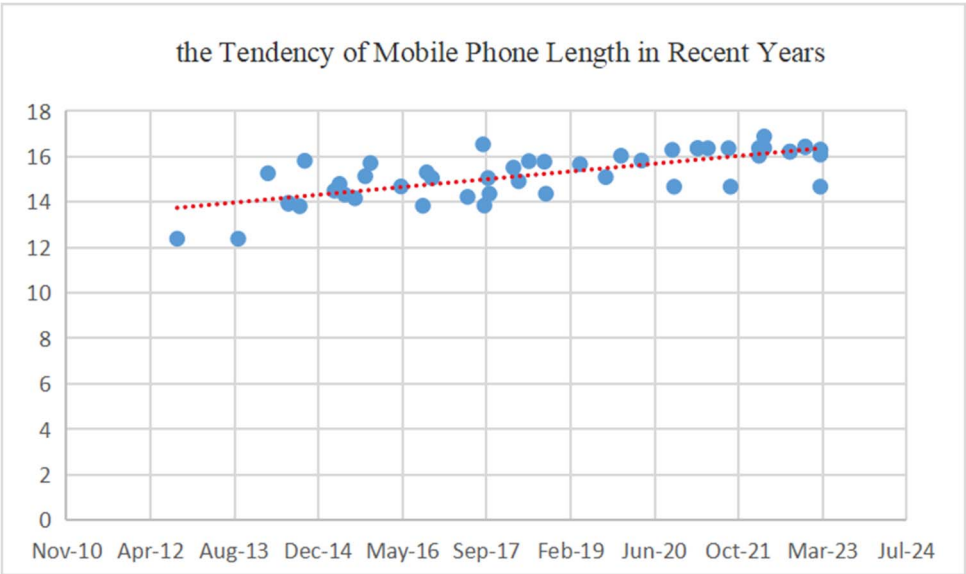


Figure A.1.1.2-1 The tendency of mobile phone length

A.1.1.3 Width

According to the surveyed mobile phone data, the width of mobile phones was around 5.86cm in 2012. By 2022, the width of mobile phones had increased to around 7.55cm, with a maximum value of 8.06cm. In recent years, the average width of mobile phones has been 7.32cm. Generally, an increase in the length of mobile phones corresponds with an increase in width. This trend is a reasonable evolution given the purpose of function requirements and aesthetic preferences.

The trend in mobile phone width in recent years is shown in Figure A.1.1.3-1

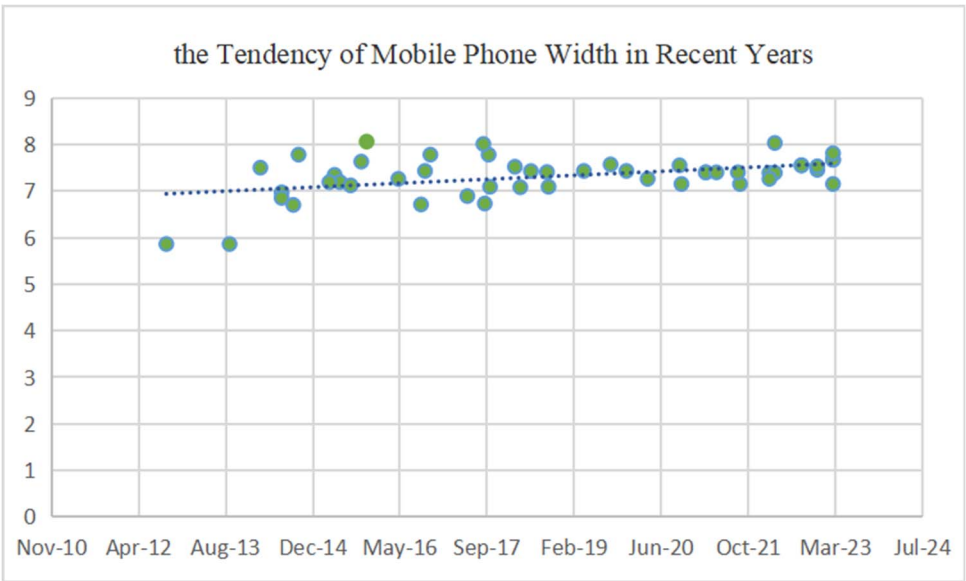


Figure A.1.1.3-1 The tendency of mobile phone width

A.1.1.4 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 in in mobile phone width in recent years is shown in Figure A.1.1.4-1

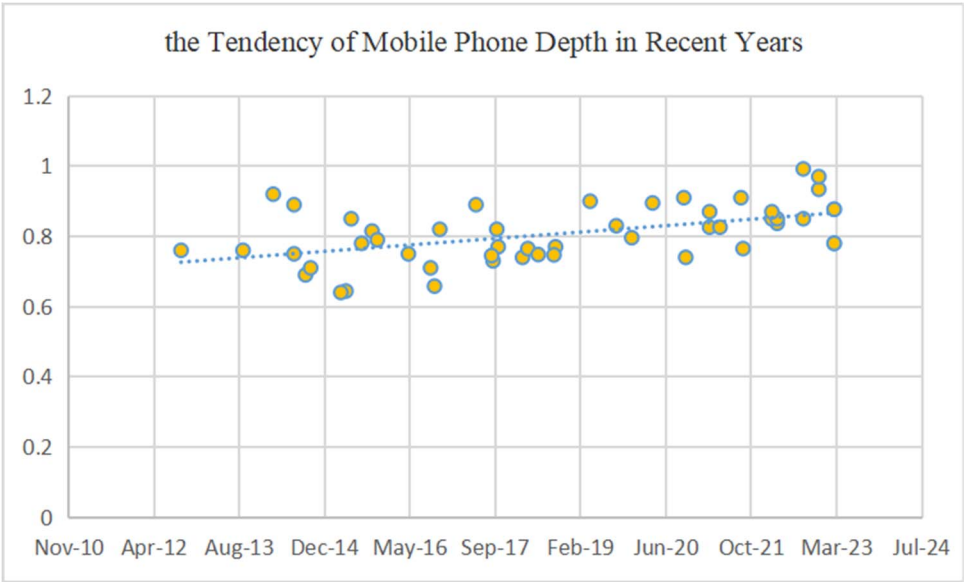


Figure A.1.1.4-1 The tendency of mobile phone depth

Table A.1.1: Modern mobile phone size

| Model | Date<br>(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.1.2 Book-style foldable mobile phone structure size

### A.1.2.1 Length

The trend in book-style foldable design length (fold/unfold) in recent years is shown in the following figure

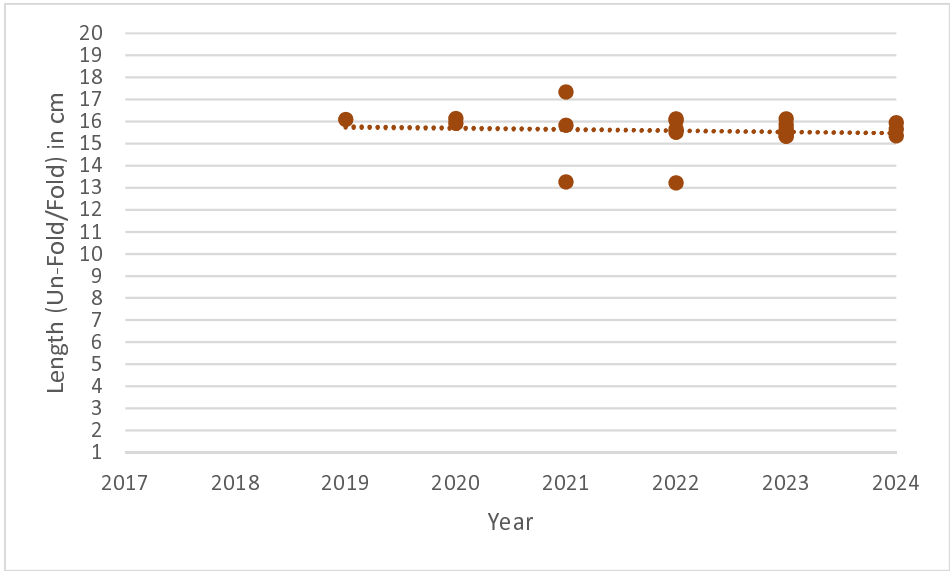


Figure A.1.2.1.1: Tendency of Book-Style Mobile Phone Length (un-fold/fold)

### A.1.2.2 Width

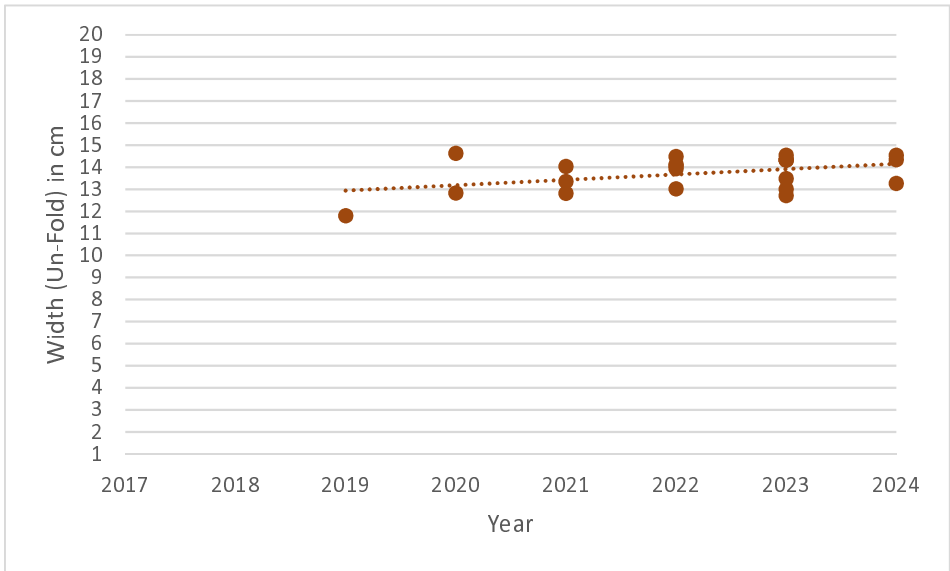


Figure A.1.2.2.1: Tendency of Book-Style Mobile Phone Width (Un-fold)

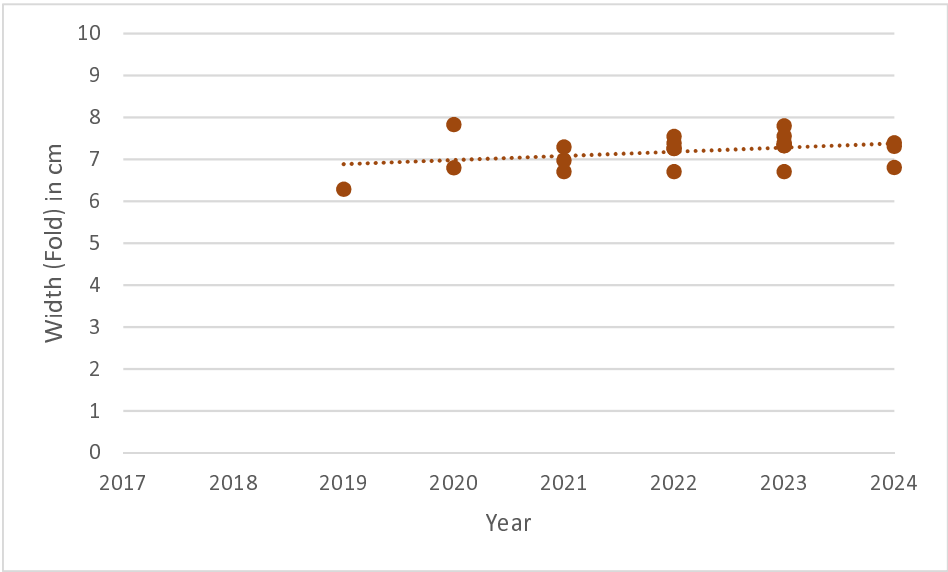


Figure A.1.2.2.2: Tendency of Book-Style Mobile Phone Width (Fold)

A.1.2.3 Depth

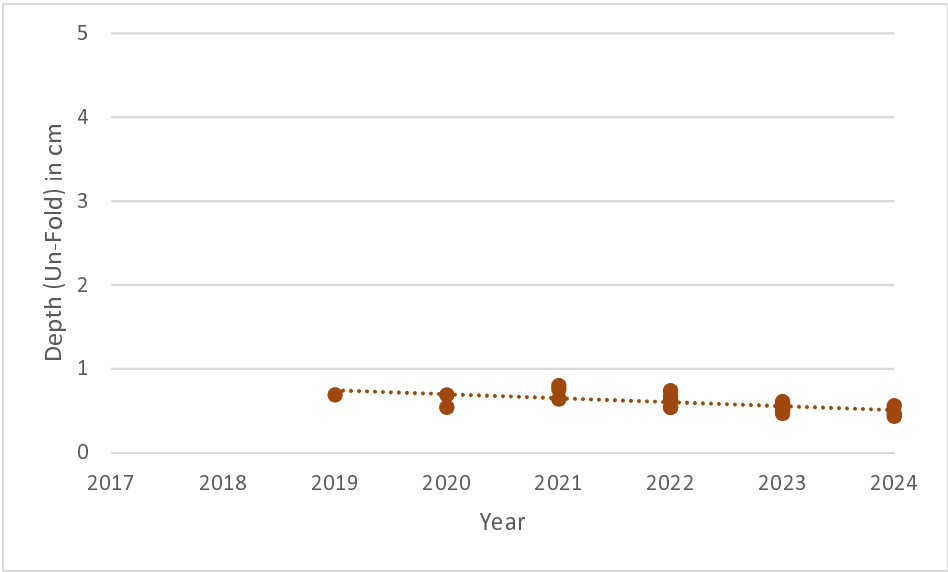


Figure A.1.2.3.1: Tendency of Book-Style Mobile Phone Depth (Un-fold)

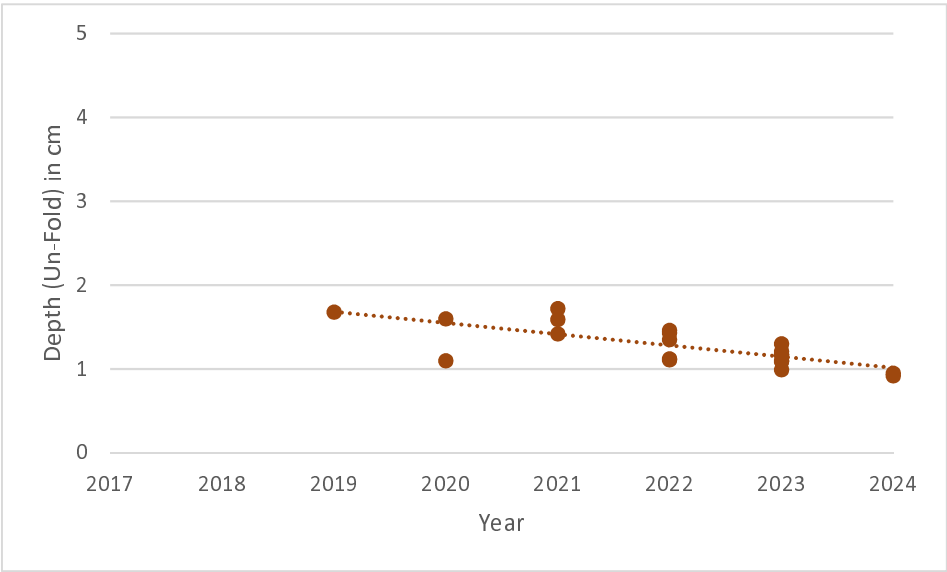


Figure A.1.2.3.2: Tendency of Book-Style Mobile Phone Depth (fold)

A.1.3 Clamshell-style foldable structure size

A.1.3.1 Length

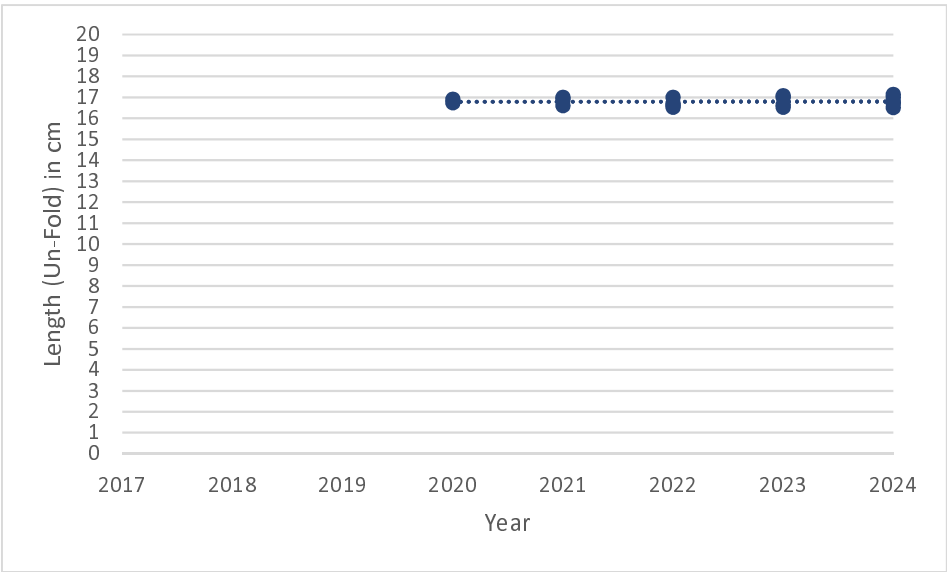


Figure A.1.3.1.1: Tendency of Clamshell-Style Mobile Phone Length (Un-fold)

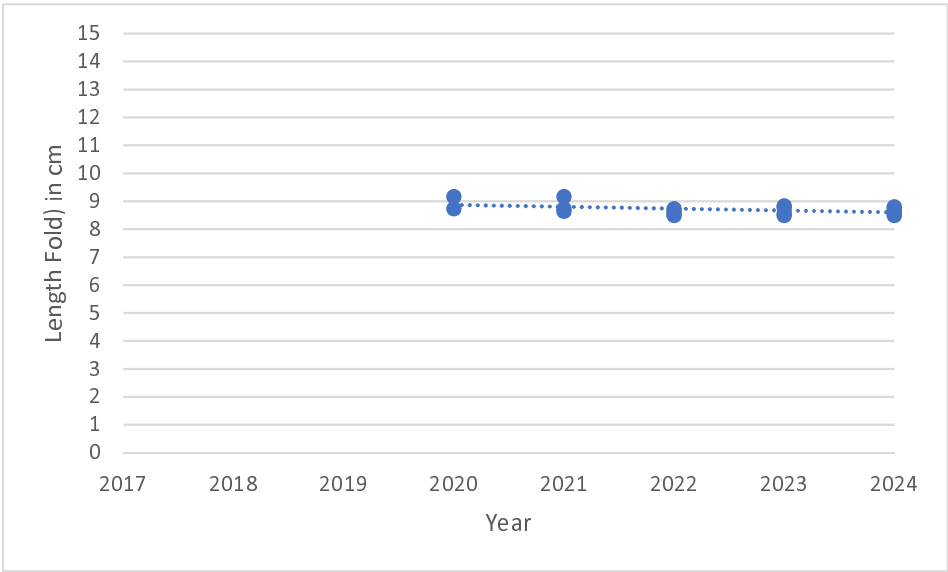


Figure A.1.3.1.2: Tendency of Clamshell-Style Mobile Phone Length (fold)

A.1.3.2 Width

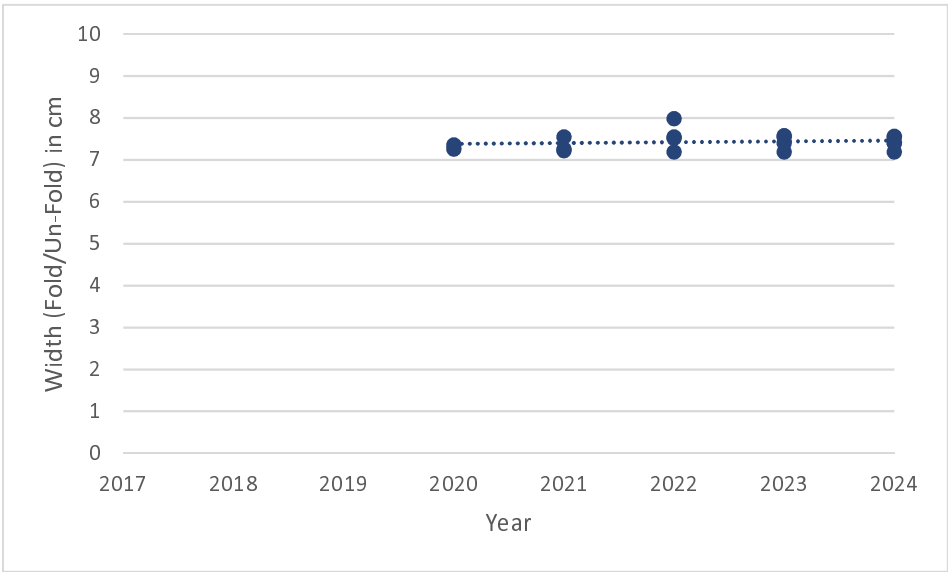


Figure A.1.3.2.1: Tendency of Clamshell -Style Mobile Phone width (Fold/Un-fold)

A.1.3.3 Depth

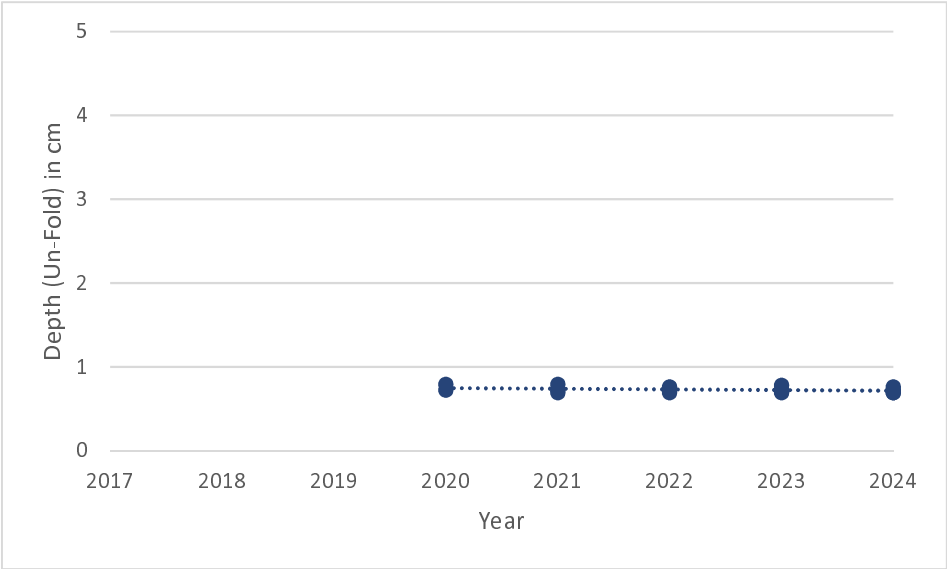


Figure A.1.3.3.1: Tendency of Clamshell-Style Mobile Phone Depth (Un-fold)

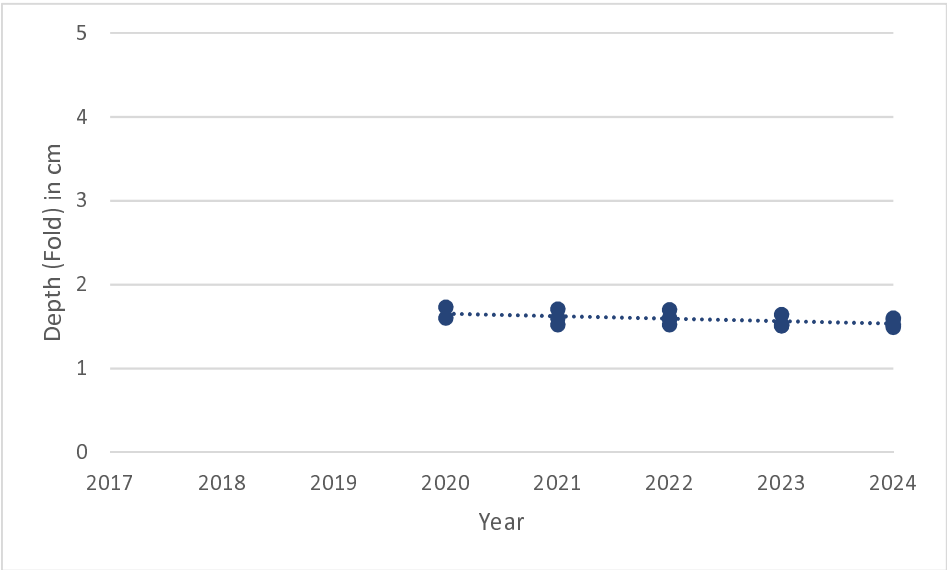


Figure A.1.3.3.2: Tendency of Clamshell-Style Mobile Phone Depth (Fold)

A.2 Earbud size

Table 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.3 Tablet size

**Table A.3: Tablet 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

**Table 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

**Table A.5-1: 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    |

**Table A.5-1: Watch size (circular type)**

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

---

## A.6 XR device size

**Table A.6: XR device 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

**Table 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

Table 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

### B.1 Introduction

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

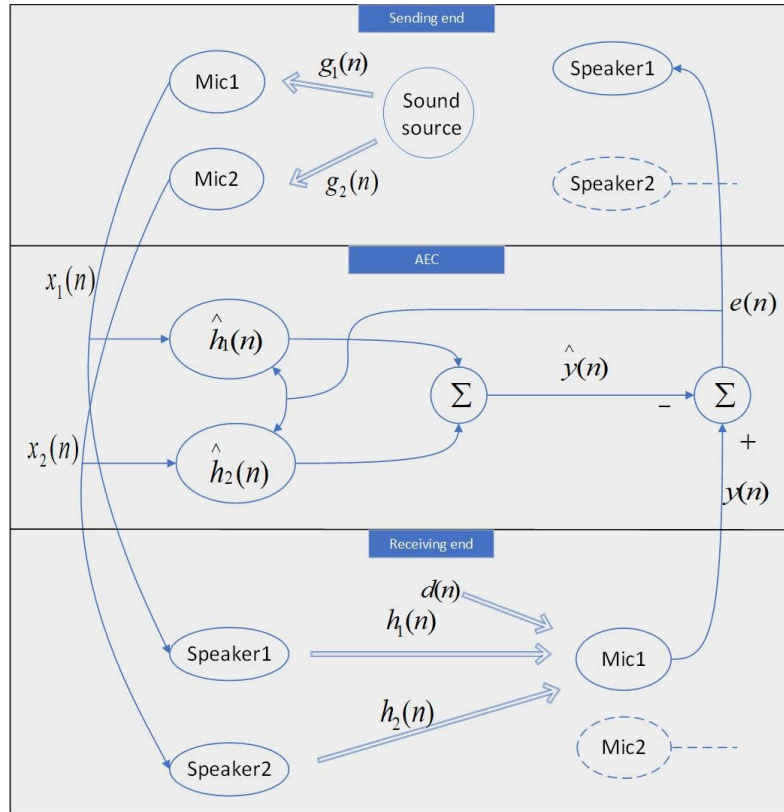


**Fig. B.1-1. Stereo AEC system.**

The two input signals  $x_1(n)$  and  $x_2(n)$  are derived from the same sound source convolving with two distinct impulse responses  $g_1(n)$  and  $g_2(n)$ . As such, 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,  $h_1(n)$  and  $h_2(n)$ , of the near-end room to a single microphone, forming the desired signal  $y(n)$  together with the noise signal  $d(n)$ . Subsequently, two adaptive FIR filters  $\hat{h}_1(n)$  and  $\hat{h}_2(n)$  are employed to estimate the impulse responses  $h_1(n)$  and  $h_2(n)$  of the near-end room, yielding the estimated desired signal  $\hat{y}(n)$ . Finally, the difference between the desired signal and the estimated desired signal, termed as the error signal  $e(n)$ , 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, with the other microphone following a similar process.

### B.2 Intuitive understanding of Stereo AEC

As illustrated in Fig.B.1-1, assuming  $h = [h_1^T, h_2^T]^T$  is the truncated vector of the actual echo path impulse response in the near-end room, and  $\hat{h} = [\hat{h}_1^T, \hat{h}_2^T]^T$  is the coefficient vector of the FIR filter. The vector  $\tilde{h} = h - \hat{h}$  is defined as the filter misalignment vector. Ideally,

$$\tilde{h}_1 * x_1 + \tilde{h}_2 * x_2 = 0 \quad (1)$$

The impulse responses of the distant room are denoted as  $g_1$  and  $g_2$ , and then substitute  $x_1 = g_1 * s$  and  $x_2 = g_2 * s$  into equation (1), and perform a Fourier transformation as follows:

$$[\tilde{H}_1(jw)G_1(jw) + \tilde{H}_2(jw)G_2(jw)]S(jw) = 0 \quad (2)$$

$\tilde{H}_1(jw)$  and  $\tilde{H}_2(jw)$  are the Fourier transforms of  $\tilde{h}_1$  and  $\tilde{h}_2$ ,  $G_1(jw)$ , and  $G_2(jw)$  are the Fourier transforms of  $g_1$  and  $g_2$ ,  $S(jw)$  represents the Fourier transform of the signal source  $s$ . For the single-channel AEC,  $G_2(jw)$  is zero, so as long as  $G_1(jw)S(jw)$  is not zero at the frequency point of interest, ensuring  $\tilde{H}_1(jw)$  to zero enables the filter estimating the path perfectly. However, for stereo AEC,  $G_2(jw)$  must not be zero, and  $S(jw)$  as the distant sound source is not zero either. Hence, it can be simplified (2), giving

$$\tilde{H}_1(jw)G_1(jw) + \tilde{H}_2(jw)G_2(jw) = 0 \quad (3)$$

It can be seen from (3) that it is obviously impossible to deduce the conclusion of  $\tilde{H}_1(jw) = 0$  and  $\tilde{H}_2(jw) = 0$ .

## B.2.1 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 clauses.

### B.2.1.1 Analysis of stereo audio covariance matrix

The following paragraphs refer to [8]. In practical scenarios, the length of the actual room impulse response is infinite. however, 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, while the impulse response length of the near-end room is set as N. The the adaptive filter coefficient length is set as L. The error signal at time n, which is the difference between the output signal of the adaptive filter and the desired signal, can be expressed as

$$e(n) = y(n) - \hat{y}(n) \quad (4)$$

Use recursive least square error formula, given

$$J(n) = \sum_{i=0}^n \lambda^{n-i} e^2(i) \quad (5)$$

where  $\lambda$  ( $0 < \lambda < 1$ ) is an exponential forgetting factor. The minimization of (5) leads to the normal equation

$$R(n) \begin{bmatrix} \hat{h}_{1,L}(n) \\ \hat{h}_{2,L}(n) \end{bmatrix} = r(n) \quad (6)$$

Where

$$R(n) = \sum_{i=0}^{n-L} \lambda^{n-i} \begin{bmatrix} X_{1,L}(i) \\ X_{2,L}(i) \end{bmatrix} \begin{bmatrix} X_{1,L}^T(i) & X_{2,L}^T(i) \end{bmatrix} = \begin{bmatrix} R_{11}(n) & R_{12}(n) \\ R_{21}(n) & R_{22}(n) \end{bmatrix} \quad (7)$$

is an estimate of the input signal covariance matrix and

$$r(n) = \sum_{i=0}^{n-L} \lambda^{n-i} y(i) \begin{bmatrix} X_{1,L}(i) \\ X_{2,L}(i) \end{bmatrix} \quad (8)$$

is an estimation of the cross-correlation vector between the input and output signals. In this scenario, if  $R(n)$  is not full rank, the normal equation may not have a unique solution, potentially causing the solution to the adaptive filter convergence deviate from the actual room impulse response. Here, the problem is examined in terms of the sizes of the adaptive filter coefficient length (L) and the impulse response length of the distant room (M), in order 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  $u = [g_{2,L}^T, -g_{1,L}^T]^T$  of length 2L, where  $g_{i,L}$  is the truncated vector of  $g_{i,M}$ , giving

$$\begin{bmatrix} X_{1,L}^T(i) & X_{2,L}^T(i) \end{bmatrix} u = X_{1,M}^T(i)g_{2,M} - X_{2,M}^T(i)g_{1,M} - t_1(n-L) + t_2(n-L) \neq 0 \quad (9)$$

where

$$t_1(n-L) = \sum_{i=M-1}^{i=L} x_1(k-i)g_2(i) \quad (10)$$

$$t_2(n-L) = \sum_{i=M-1}^{i=L} x_2(k-i)g_1(i) \quad (11)$$

Hence, matrix  $R(n)$  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  $t_1(n-L)$  and  $t_2(n-L)$ , the covariance matrix  $R(n)$  is very ill-conditioned, exhibiting significant divergence in eigenvalues. This results in a slow convergence rate of the adaptive filter, a phenomenon commonly referred to as the “non-uniqueness” issue in the multi-channel acoustic echo problems. Under the premise of  $L < M$ , 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  $h_{i,N}$  is divided into two parts: one is the vector  $h_{i,L}$  matching the first  $L$  points of the adaptive filter length, and the other is the trailing vector  $h_{i,t}$  with the length of  $N-L$ . Simultaneously, repeated supplementary definitions for the input signal are defined to match the room impulse response length. After a series of derivation, the Wiener solution can be got as follows:

$$\begin{bmatrix} \hat{h}_{1,L}(n) \\ \hat{h}_{2,L}(n) \end{bmatrix} = R^{-1}(n)r(n) = \begin{bmatrix} h_{1,L} \\ h_{2,L} \end{bmatrix} + R^{-1}(n)R_t(n) \begin{bmatrix} h_{1,t} \\ h_{2,t} \end{bmatrix} \quad (12)$$

It can be concluded that, from the point of view of Wiener solution, the estimated adaptive solution is closely related to  $R^{-1}(n)$ .

### B.2.1.2 Relationship between channel correlation and condition number of the covariance matrix

The following paragraphs refer to [9]. From the above description, it is evident that the coefficients of the adaptive filter are closely related to the norm of covariance matrix  $R$ . It 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, the condition number is utilized to establish the relationship between the correlation of stereo signals and the covariance matrix. Let  $\mathbf{x}(n) = [\mathbf{x}_1^T(n)\mathbf{x}_2^T(n)]^T$ , and the two-channel covariance matrix is thus given by

$$\mathbf{R} = E\{\mathbf{x}(n)\mathbf{x}^T(n)\} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix}_{2L \times 2L} \quad (13)$$

where  $E\{\cdot\}$  is the mathematical expectation operator, the  $\mathbf{R}_{jk}$  is the covariance matrix between the  $j$ th and  $k$ th channel. It should be noted that for  $L \rightarrow \infty$ , a Toeplitz matrix is asymptotically equivalent to a circulant matrix, provided its elements are absolutely summable, given in (13) as (14)

$$\mathbf{R}_{jk} = \mathbf{F}^{-1} \mathbf{S}_{jk} \mathbf{F}, \quad j, k = 1, 2 \quad (14)$$

where  $\mathbf{F}$  is the Fourier matrix defined with elements  $F_{pq} = e^{-i2\pi pq/L}$ ,  $p, q = 0, 1, \dots, L-1$ . The  $L \times L$  matrix  $\mathbf{S}_{jk} = \text{diag}\{S_{jk}(0), S_{jk}(1), \dots, S_{jk}(L-1)\}$  contains elements corresponding to the frequency bins. These are formed from the Discrete Fourier Transform (DFT) of the first column of  $\mathbf{R}_{jk}$ . Let  $r_{jk}(l)$  be the auto- and cross-correlation coefficients for  $j = k$  and  $j \neq k$ , respectively. This establishes the relationship between  $\mathbf{S}_{jk}$  and  $r_{jk}(l)$  as follows

$$S_{jk}(f) = \sum_{l=-\infty}^{\infty} r_{jk}(l) e^{-i2\pi fl}, \quad f = 0, 1, \dots, L-1 \quad (15)$$

The covariance matrix  $R$  of stereo signal can be expressed as

$$\mathbf{R} = \begin{bmatrix} \mathbf{F}^{-1} & \mathbf{0}_{L \times L} \\ \mathbf{0}_{L \times L} & \mathbf{F}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{F} & \mathbf{0}_{L \times L} \\ \mathbf{0}_{L \times L} & \mathbf{F} \end{bmatrix} \quad (16)$$

In this content, E-norm is used to represent the condition number  $\chi[R]$  of matrix  $R$ . The E-norm is equivalent to the F-norm scaled by a factor  $1/\sqrt{2L}$ . 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

$$\chi_E \left[ \mathbf{R}^{\frac{1}{2}} \right] = \left\| \mathbf{R}^{\frac{1}{2}} \right\|_E \left\| \mathbf{R}^{-\frac{1}{2}} \right\|_E \quad (17)$$

Use positive definite covariance matrix to diagonalize to calculate  $\chi_E[\mathbf{R}^{1/2}]$ , giving

$$\chi_E^2[\mathbf{R}^{\frac{1}{2}}] = \frac{1}{4L^2} \text{tr}(\mathbf{R}) \text{tr}(\mathbf{S}^{-1}) \quad (18)$$

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

$$|\gamma(f)|^2 = \frac{|s_{12}(f)|^2}{s_{11}(f)s_{22}(f)}, \quad f = 0, 1, \dots, L-1 \quad (19)$$

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

$$\chi_E^2 \left[ \mathbf{R}^{\frac{1}{2}} \right] = \frac{1}{4L^2} \left[ \sum_{l=1}^{L-1} [S_{11}(l) + S_{22}(l)] \right] \times \sum_{l=1}^{L-1} [1 - |\gamma(l)|^2]^{-1} [S_{11}^{-1}(l) + S_{22}^{-1}(l)] \quad (20)$$

### B.2.1.3 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  $\gamma$ . This 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 stereo AEC, the de-correlation of channel signals is a key step to the solution of adaptive filter coefficients. However, this operation may cause sound quality degradation, which may be balanced using this method.

---

## Annex C (informative): Change history

| Change history |                                                   |             |        |     |     |                                                                                                                                                                                                                                                                                                                                          |                |
|----------------|---------------------------------------------------|-------------|--------|-----|-----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|
| Date           | Meeting                                           | TDoc        | C<br>R | Rev | Cat | Subject/Comment                                                                                                                                                                                                                                                                                                                          | New<br>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-<br>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:<br>contents from S4-240150(two paragraphs) as 6.2.3.5<br>Contents from S4-240201 as 7.2<br>Contents from S4-240153 as Annex B<br>Contents from S4-240243 as 4.1.2<br>Contents from S4-240241<br>Integrate S4-240231 in the bracket with editor's note                                                      | 0.4.0          |
| 2024-03        | SA4#127-<br>post                                  | S4aA240014  |        |     |     | Editorial corrections: fixed hanging paragraphs, chapter/ reference number, etc.                                                                                                                                                                                                                                                         | 0.4.1          |
| 2024-04        | SA4#127-<br>bis-e                                 | S4-240740   |        |     |     | Integrated following updates:<br>Agreed contents from S4aA240015<br>Agreed contents from S4aA240014                                                                                                                                                                                                                                      | 0.5.0          |
| 2024-04        | SA4#127-<br>bis-e                                 | S4-240764   |        |     |     | Integrated following updates:<br>1:Agreed contents from S4-240731<br>2:Agreed contents from S4-240733<br>3:Section 2 from S4-240665 in bracket for further discussion                                                                                                                                                                    | 0.6.0          |
| 2024-05        | SA4#128                                           | S4-241038   |        |     |     | Editorial updates and remove bracket for the completed parts                                                                                                                                                                                                                                                                             | 0.6.1          |
| 2024-05        | SA4#128                                           | S4-241173   |        |     |     | Integrated the following contents:<br>1: Agreed contents from S4-241192<br>2: Agreed contents from S4-241038<br>3: Agreed contents from S4-241039 in brackets                                                                                                                                                                            | 0.7.0          |
| 2024-06        |                                                   |             |        |     |     | Version 1.0.0 created by MCC, presentation for information to TSG                                                                                                                                                                                                                                                                        | 1.0.0          |
| 2024-06        | <a href="#">3GPPSA4-e (AH) Audio SWG post 128</a> | SA4aA240044 |        |     |     | Remove automatic-numbering and some other editorial updates                                                                                                                                                                                                                                                                              | 1.0.1          |
| 2024-08        | SA4#129-e                                         | S4-241524   |        |     |     | Some editorial updates and integrating contents from SA4aA240040, SA4aA240042, SA4aA240043, SA4aA240045.                                                                                                                                                                                                                                 | 1.0.2          |
| 2024-08        | SA4#129-e                                         | S4-241666   |        |     |     | Integrated agreed contents from S4-241511 and change from "microphone design for UEs" to "microphone integration in UEs" in clause 7.<br>Integrated agreed contents "2.2 Proposed updates in clause 7.2.4.3 HOA" from S4-241511 and made reference for third solution with IVAS.<br>Remove 2 editor's note.<br>Replace "should" by "may" | 1.1.0          |
| 2024-08        | SA4#129-e                                         | S4-241787   |        |     |     | Integrated clause "conclusion and recommendation" from offline discussion, the clause is put in brackets.                                                                                                                                                                                                                                | 1.2.0          |
| 2024-09        | 3GPPSA4-e (AH) Audio SWG post 129-e               | S4aA240055  |        |     |     | Integrated the contents from online editing during Sep2 telco.                                                                                                                                                                                                                                                                           | 1.3.0          |
| 2024-11        | SA4#130                                           | S4-241946   |        |     |     | Made several editorial updates                                                                                                                                                                                                                                                                                                           | 1.4.0          |
| 2024-11        | SA4#130                                           | S4-242096   |        |     |     | Integrated Foldable mobile phone description and add several editorial updates, update clause 8.2.1, clause 10 and some online edits.                                                                                                                                                                                                    | 1.5.0          |
| 2024-11        | SA4#130                                           | S4-242139   |        |     |     | Update Figure style to Arial 10. In clause 10: replace word "should" to "is expected to", added reference of "TS 26.260 and 26.261", added "NOTE".                                                                                                                                                                                       | 1.6.0          |
| 2024-12        | SA#106                                            |             |        |     |     | Version created by MCC to send to TSG SA for approval                                                                                                                                                                                                                                                                                    | 2.0.0          |
| 2025-01        |                                                   |             |        |     |     | Version created by MCC upon approval of v2.0.0 in TSG SA#106                                                                                                                                                                                                                                                                             | 19.0.0         |

---

# History

| Document history |              |             |
|------------------|--------------|-------------|
| V19.0.0          | January 2026 | Publication |
|                  |              |             |
|                  |              |             |
|                  |              |             |
|                  |              |             |