

Discontinuous Transmission and Comfort Noise Generation for Discrete Stereo Audio Codecs

Jan Frederik Kiene
Fraunhofer Institute for Integrated Circuits IIS
Erlangen, Germany
jan.kiene@iis.fraunhofer.de

Eleni Fotopoulou
DSP Solutions GmbH
Regensburg, Germany
eleni.fotopoulou@dspsolutions.de

Abstract—In communication codecs, discontinuous transmission (DTX) is used to reduce the transmission bitrate in speech pauses. Instead of coding the actual audio signal, only a parametric representation of the background noise is transmitted and on the receiver side, so-called comfort noise with similar spectral characteristics is synthesized. Audio codecs for mobile communication traditionally only support transmission of one-channel (mono) audio. The 3GPP codec for Immersive Voice and Audio Services (IVAS) extends this capability to also support stereo and immersive audio formats. To enable DTX for stereo signals, not only the spectral characteristics of the background noise need to be captured and transmitted, but also its spatial properties. This paper describes a system for DTX and Comfort Noise Generation (CNG) for discrete stereo codecs which was developed as part of the IVAS codec standard. It comprises joint coding of the spectral noise shape parameters for the stereo signal and a spatial CNG algorithm based on modelling the inter-channel similarity of the stereo signal by mixing different uncorrelated noise sources. A listening test evaluation shows benefits of this system over a baseline stereo DTX implementation with separate CNG for the two stereo channels.

Index Terms—communication, speech coding, DTX, CNG, IVAS

I. INTRODUCTION

Communication codecs apply discontinuous transmission (DTX) techniques to lower the bitrate in voice conversations during speech pauses [1]. DTX takes advantage of the fact that in a typical conversation, usually not all participants are actively talking at all times. Voice activity detection (VAD) is used in the encoder to detect whether there is active speech present in the input signal [1]. If no active speech is detected, it is assumed that the signal consists of background noise only. Instead of coding the background noise with the same bitrate as the active speech signal portions, a parametric description of the background noise is derived at the encoder and transmitted as a so-called silence descriptor (SID) with a much lower bitrate [1]. These SIDs are usually transmitted less frequently than frames containing active signal content. In between the SIDs, transmission can be completely switched off to lower the overall bitrate even more, and to help reduce energy consumption of the user end device by lowering the activity of the radio transmission modem [2] and reduce energy consumption in radio base stations [3].

On the decoder side, the parametric noise description transmitted in the SIDs is used to synthesize so-called comfort

noise that ideally matches the characteristics of the original background noise closely enough such that the transitions between full-bitrate coding and CNG are not disturbing for the listener or – ideally – not even easily noticeable. Outputting comfort noise instead of simply muting the signal avoids giving the impression to the listener that the call might have been dropped.

Most currently deployed state-of-the-art communication codecs such as the Enhanced Voice Services codec (EVS)[4] support DTX, but typically only support coding of mono audio signals. On the other hand, existing stereo-capable communication codecs or stereo extensions of mono codecs lack support for DTX. Stereo operation with EVS is only specified as dual-mono operation in [5]. However, no stereo-specific techniques are applied for DTX in that case, as the two mono channels are processed independently with EVS. A predecessor of EVS, the Adaptive Multi-Rate - Wideband (AMR-WB) codec, also supports DTX[6] and is available in an extended version called Extended Adaptive Multi-Rate - Wideband (AMR-WB+) which adds stereo support [7]. AMR-WB+, however, does not allow for combining stereo and DTX operation, supporting DTX only in an AMR-WB backwards-compatible mode. Stereo extensions for ITU-T codecs such as G.722 and G.711 exist[8], but lack DTX support already in their mono versions. Another fully stereo-capable codec targeting communication scenarios is the stereo-extended Enhanced Low Delay Advanced Audio Coding (AAC-ELDv2)[9]. Again, this codec does not support DTX. Opus[10] offers full support for stereo signals and also includes a DTX mode that does not require additional information about the background signals like SID frames. Instead, the bitrate is simply reduced during silence or background noise parts by only encoding one frame in a predetermined interval. The missing frames are filled using packet loss concealment (PLC) in the decoder, which is an optional decoder-side feature [10]. Any CNG mechanism would therefore be part of the PLC functionality of the respective decoder and depends on the specific implementation including any stereo-specific aspects. However, as no SID frames are specified for Opus, it would always be a decoder-only solution without taking into account any information derived from the original background noise.

The mentioned existing codecs do not provide a complete stereo DTX system including both stereo-aware VAD and

TABLE I
STEREO AND DTX CAPABILITIES OF STATE-OF-THE-ART COMMUNICATION CODECS.

Codec	Stereo	DTX	Remarks
EVS [4], [5]	(✓)	✓	Stereo only as dual-mono, no stereo-specific DTX
AMR-WB [6]	—	✓	Mono only
AMR-WB+ [7]	✓	(✓)	DTX only in AMR-WB backwards-compatible (mono) mode
G.711/G.722 stereo ext. [8]	✓	—	No DTX, not even in mono versions
AAC-ELDv2 [9]	✓	—	Fully stereo-capable, no DTX
Opus [10]	✓	(✓)	DTX without SID frames; CNG via optional decoder-side PLC

CNG, and explicit signaling of DTX operation and transmission of background noise information. DTX and stereo operation together are only supported by dual-mono EVS and Opus, but both lack aspects of a complete stereo DTX system. Table I summarizes the comparison of state-of-the-art communication codecs regarding stereo and DTX capability. The proposed system, as part of the recently standardized 3GPP codec for Immersive Voice and Audio Services (IVAS) [11], aims to fill the existing gap by providing a complete stereo-aware DTX mechanism in the framework of a modern communication audio codec.

IVAS is the first fully immersive communication codec, extending the EVS codec towards spatial audio formats such as stereo, multi-channel, object-based audio, scene-based audio and metadata-assisted spatial audio while addressing the low-latency and low-complexity requirements of a voice communication codec [11]. Along with its core use cases in immersive communication – such as immersive telephony with experience sharing and ad-hoc and multi-party conferencing – IVAS could also play a supporting role for applications in the field of the Internet of Sounds (IoS) [12] and Internet of Musical Things (IoMusT) [13] such as immersive audio experiences or installations, or virtual/augmented reality applications [14]. IVAS was especially designed for use in fifth-generation (5G) networks, which were identified as a key enabler for the IoMusT paradigm [15] and offers functionality such as jitter buffer management (JBM) and packet loss concealment (PLC), which are also vital for IoMusT applications [16]. Supporting DTX for spatial and immersive audio formats can also contribute to the ongoing work towards a sustainable IoS [17].

Stereo capability in a communication codec comes with challenges for the DTX operation. VAD classification needs to take two signals into account and generate a common decision for both channels of the stereo signal. The speech signal will usually be present in both channels with possible level and time differences, so both channels have to be considered to ensure no parts of the active speech content are lost and that transitions to and from DTX operation happen at the appropriate times.

Regarding CNG, background noise will be present in both channels, so the generated comfort noise needs to match not only its spectral characteristics, but also its spatial characteristics. Stereo inputs might come from stereo microphones, microphone array processing, or be artificially created by panning mono sources with a common stereo ambience signal. Back-

ground noise can be diffuse (e.g. traffic or babble noise) or dominated by a prominent directional source (e.g. a ventilation duct or car motor), and its perceived width is further influenced by the recording setup, especially the microphone distance. Since perceived source width is inversely proportional to the absolute value of the inter-channel cross-correlation coefficient [18], a stereo CNG system should generate comfort noise with a similar level of inter-channel correlation as the input background noise to accurately model the width of the original background noise in the stereo image.

Finally, a stereo DTX mechanism needs to integrate well with the applied active signal stereo coding paradigm. Two main approaches exist, namely parametric stereo coding methods and joint or discrete stereo coding methods [19]. While parametric stereo coders downmix the stereo signal into a single channel along with some parameters which are used to upmix the single channel again in the decoder, discrete stereo coding methods code both stereo channels after transforming them into a more compact representation. Adapting the DTX mechanism to the active stereo coding method allows for an efficient implementation and smooth transitions from and to CNG operation.

A straightforward solution for supporting stereo DTX could be achieved simply by two separate mono DTX coders running independently on the two stereo channels, as e.g. in the EVS dual-mono operation [5]. Based on the mentioned challenges for stereo DTX, we argue that such an implementation will not be able to correctly model and synthesize spatial comfort noise and will be outperformed by a stereo-aware DTX system. Hence, in this paper, a method for implementing DTX and CNG for stereo signals is proposed. Based on the mono DTX/CNG system described in [20], three modifications are introduced: synchronization of the VAD between the two stereo channels, joint encoding of the noise estimates, and a stereo CNG mechanism based on estimating the inter-channel coherence of the stereo signal. The system is used in the high bitrate discrete stereo mode [21] of the IVAS codec and is therefore designed to be compatible with a discrete stereo coder.

The rest of the paper is organized as follows: Section II describes the proposed stereo DTX system and how it extends existing techniques for single-channel DTX to address the mentioned challenges for DTX operation for stereo signals. Section III presents both subjective and objective evaluation results for the system. Section IV presents the conclusions.

II. PROPOSED STEREO DTX/CNG SYSTEM

A. Overview of the basic mono DTX/CNG system

The proposed DTX/CNG system is based on the existing DTX algorithm that is employed in the EVS codec[20].

In EVS, the input signal is processed frame-wise with a frame length of 20 ms. For every frame, VAD is performed and the signal is classified as either active or inactive to generate a VAD classification result $v \in \{0, 1\}$, where a value of zero corresponds to no voice activity being detected and a value of one corresponds to active voice activity being detected. The VAD classification mechanism evaluates sub-band SNR and spectral features; details can be found in [4].

The background noise estimation operates in the frequency domain and uses a hybrid spectral analysis where Fast Fourier Transform (FFT) analysis is used for low frequencies up to 6.4 kHz and a Complex Low-Delay Filter Bank (CLDFB) with lower spectral resolution is used for the remaining higher frequencies [20]. For every frame, the spectrum is divided into critical bands [22] and the energy for each band is computed. To estimate the background noise band-wise energies, a modified version of the minimum statistics algorithm[23][20] is used, which transforms the energy values to the logarithmic domain and quantizes them before generating the estimate. This preprocessing reduces the energies' dynamic range, which is favorable for fixed-point implementation[20]. The resulting noise estimates are converted back to the linear domain to produce the final encoder-side noise estimates $\sigma_{\text{est}}^2(k), k \in \{0, \dots, L_{\text{enc}} - 1\}$, where L_{enc} denotes the number of bands used for noise estimation and depends on the sampling rate the codec operates in [20].

For transmission in the SID, the noise estimates are converted into dB and quantized using a Multi-Stage Vector Quantizer (MSVQ) with separately coded gain[20]. The global gain is quantized using seven bits and the MSVQ uses six stages with 37 bits in total. Candidate vectors are selected using an M-best search algorithm with 24 survivors in each stage[24].

The noise estimates transmitted in the SID allow for adaptation to changes in the background noise during inactive signal parts, but due to the low-bitrate constraint of the SID, their spectral resolution is not enough to capture the fine spectral structure of the background noise. Hence, the same noise estimation algorithm as in the encoder is run in the decoder, but with a higher spectral resolution by using narrower frequency bands. This decoder-side noise estimate is combined with the noise estimate transmitted in the SID to derive an energy spectral density $\sigma_{\text{CNG}}^2(m)$ over the whole frequency range of the hybrid filterbank[20].

CNG is then performed in the frequency domain. Let $n(m) \sim \mathcal{N}(0, 1)$, then the complex comfort noise spectrum is defined as

$$N(m) = \sqrt{\frac{\sigma_{\text{CNG}}^2(m)}{2}}(n(2m) + jn(2m + 1)), \quad (1)$$

where j is the imaginary unit and m is the index of the frequency bins. Note that noise is only generated above 50 Hz,

so $N(m)$ is set to zero for the corresponding lower values of m and for the value of m corresponding to the Nyquist frequency, the imaginary part is set to zero to ensure that the corresponding time-domain signal stays real. This shapes the white spectrum of the signal from the noise source to adjust the spectral envelope to the shape estimated from the background noise represented by $\sigma_{\text{CNG}}^2(m)$. The resulting comfort noise spectrum is transformed back to the time domain to generate the final output comfort noise signal. Transitions between active and inactive frames are handled by a windowing and overlap-add mechanism and decoder signal memories are updated in inactive frames with the time-domain comfort noise signal to ensure a continuous output signal[20].

A SID frame in the mono DTX system as implemented in EVS consists of 48 bits which is equivalent to a bitrate of 2.4 kbps. In summary, the mono SID frame consists of:

- 44 bits for the MSVQ-coded noise estimate, with 7 bits for the global gain and 37 bits for the residual shape.
- Four bits of EVS-specific core-coder side information [25].

B. Proposed modifications for stereo operation

An overview of the proposed system is shown in Figure 1.

To generate a global, synchronized VAD classification, first two separate classifications $v_l \in \{0, 1\}, v_r \in \{0, 1\}$ for the left and right channel, respectively, are obtained. The combined VAD classification is then generated via

$$v_{\text{stereo}} = v_l \vee v_r, \quad (2)$$

where $\vee$ denotes the logical OR of the two single-channel classifications.

For encoding the two noise estimates, spectral analysis and noise estimation are performed separately on each channel as for mono, yielding the two noise estimates $\sigma_{\text{est},l}^2(k)$ and $\sigma_{\text{est},r}^2(k)$. To keep the SID bitrate low, the two noise estimates are then encoded jointly. Joint encoding using mid/side (M/S) transform is a common approach for efficient coding of stereo signals that exploits redundancies between the two channel signals[19]. Instead of coding the left and right channel signals directly, their sum (mid) and difference (side) signals are coded. High energy compaction into the mid signal can be achieved for stereo signals with similar channels, especially for near-monophonic signals[19]. This technique is applied to the noise estimates. Since the background noise is expected to be present in both channels, the noise estimates, which represent the background noise spectral shape, are expected to include some redundancies and M/S encoding can efficiently compress them to lower the SID size.

The noise estimates are converted to dB via

$$\sigma_{\text{est},l,\text{dB}}^2(k) = 10 \log_{10}(\sigma_{\text{est},l}^2(k) + \epsilon), \quad (3)$$

$$\sigma_{\text{est},r,\text{dB}}^2(k) = 10 \log_{10}(\sigma_{\text{est},r}^2(k) + \epsilon), \quad (4)$$

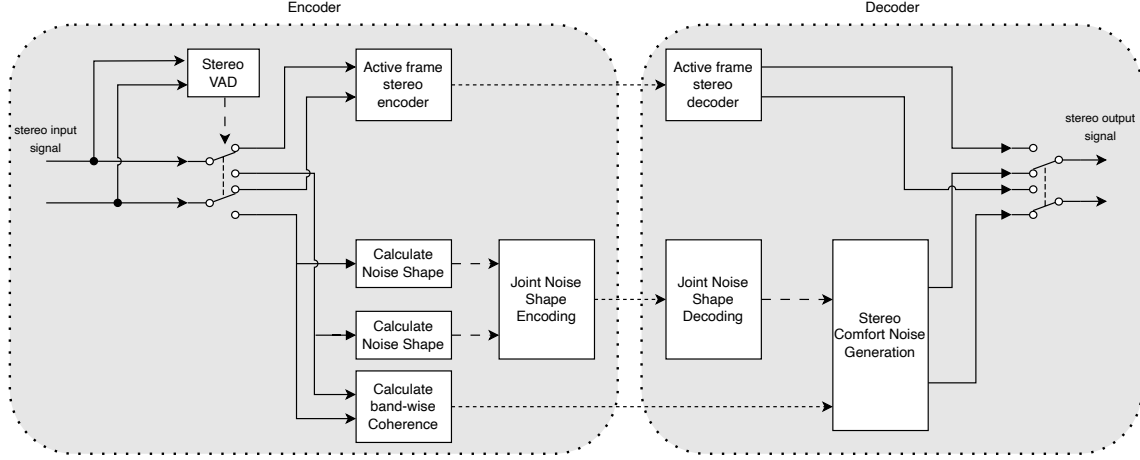


Fig. 1. High-level diagram of the proposed Stereo DTX/CNG system

TABLE II
ANALYSIS BANDS FOR THE COHERENCE MEASURE CALCULATION.

b	start_b	stop_b	Frequency range
0	2	7	50 Hz - 200 Hz
1	8	15	200 Hz - 400 Hz
2	16	31	400 Hz - 800 Hz
3	32	47	800 Hz - 1200 Hz
4	48	128	1200 Hz - 3200 Hz

with $\epsilon = 10^{-15}$, and then transformed to their M/S representation:

$$\sigma_{\text{est,mid}}^2(k) = \frac{\sigma_{\text{est,l,dB}}^2(k) + \sigma_{\text{est,r,dB}}^2(k)}{2}, \quad (5)$$

$$\sigma_{\text{est,side}}^2(k) = \frac{\sigma_{\text{est,l,dB}}^2(k) - \sigma_{\text{est,r,dB}}^2(k)}{2}. \quad (6)$$

The same quantization strategy as for mono is applied, with the modification that $\sigma_{\text{est,side}}^2(k)$ is quantized using only four stages of the MSVQ with only 25 bits in total. That value was chosen based on informal experiments and given the constraints of the existing MSVQ structure and the fact that in the IVAS framework, the SID frame size was harmonized across all formats to be 104 bits.

The stereo CNG mechanism produces comfort noise with a varying degree of inter-channel correlation based on a frequency-dependent encoder-side estimate of the coherence of the background noise channels. In preliminary experiments, it was found that a broadband estimate is not sufficient to achieve convincing spatial comfort noise synthesis as the inter-channel relationship of the background noise can vary with frequency. Therefore, a frequency-domain-based measure is used, re-using the FFT part of the spectral analysis which is split into $N = 5$ bands, to achieve a compromise between SID bitrate and necessary spectral selectivity. The frequency bands are derived by combining multiple critical bands with smaller bands at low frequencies and wider bands at higher frequencies[25]. Let $i \in \mathbb{N}$ denote the frame number and

L_i, R_i the spectra of the left and right channel in that frame, respectively, then the coherence measure $c_b[i]$ for each band $b \in \{0, \dots, N-1\}$ is computed as

$$c_b[i] = \sqrt{\frac{c_{\Re,b}^2[i] + c_{\Im,b}^2[i]}{e_{l,b}[i] \cdot e_{r,b}[i] + \epsilon}} \in [0, 1], \quad (7)$$

with

$$c_{\Re,b}[i] = \sum_{m=\text{start}_b}^{\text{stop}_b} \Re(L_i(m) \cdot R_i^*(m)), \quad (8)$$

$$c_{\Im,b}[i] = \sum_{m=\text{start}_b}^{\text{stop}_b} \Im(L_i(m) \cdot R_i^*(m)), \quad (9)$$

$$e_{l,b}[i] = \sum_{m=\text{start}_b}^{\text{stop}_b} L_i(m) \cdot L_i^*(m), \quad (10)$$

$$e_{r,b}[i] = \sum_{m=\text{start}_b}^{\text{stop}_b} R_i(m) \cdot R_i^*(m). \quad (11)$$

Here, $e_{l,b}, e_{r,b}$ denote the energies of the left and right channel per band, m is the index of the spectral bins, $\text{start}_b, \text{stop}_b$ denote the start and stop indices of band b , see Table II, and $\epsilon = 10^{-15}$, which is added to the denominator for numerical stability. For low values of c_b , the background noise in the corresponding frequency band is more diffuse and has a rather wide stereo width. Higher values of c_b are obtained for background noise which is more correlated between the channels and sounds less wide. As the background noise statistics can fluctuate from frame to frame, smoothing the intermediate values used in calculating c_b over time is beneficial to achieve a stable comfort noise synthesis. Therefore,

a first-order Infinite Impulse Response (IIR) filter is used to smooth the values over the frame index $i \in \mathbb{N}$:

$$\bar{c}_{\mathcal{R},b}[i] = \alpha \bar{c}_{\mathcal{R},b}[i-1] + (1-\alpha) c_{\mathcal{R},b}[i], \quad (12)$$

$$\bar{c}_{\mathcal{I},b}[i] = \alpha \bar{c}_{\mathcal{I},b}[i-1] + (1-\alpha) c_{\mathcal{I},b}[i], \quad (13)$$

$$\bar{e}_{1,b}[i] = \alpha \bar{e}_{1,b}[i-1] + (1-\alpha) e_{1,b}[i], \quad (14)$$

$$\bar{e}_{r,b}[i] = \alpha \bar{e}_{r,b}[i-1] + (1-\alpha) e_{r,b}[i], \quad (15)$$

with $\bar{c}_{\mathcal{R},b}[0] = \bar{c}_{\mathcal{I},b}[0] = \bar{e}_{1,b}[0] = \bar{e}_{r,b}[0] = \epsilon \forall b$. Here, $\alpha \in [0, 1]$ is the smoothing factor of the filter with $\alpha = 0.95$ in the following. The recursively smoothed coherence measure is then given by

$$\bar{c}_b[i] = \sqrt{\frac{\bar{c}_{\mathcal{R},b}^2[i] + \bar{c}_{\mathcal{I},b}^2[i]}{\bar{e}_{1,b}[i] \cdot \bar{e}_{r,b}[i] + \epsilon}}. \quad (16)$$

The coherence measure values $\bar{c}_b[i]$ are uniformly quantized using 4 bits per band, i.e., the quantized coherence value $\hat{c}_b[i]$ is given by

$$\hat{c}_b[i] = \frac{\min(15, \lfloor 15 \cdot \bar{c}_b[i] + 0.5 \rfloor)}{15}, \quad (17)$$

where $\lfloor \cdot \rfloor$ denotes rounding down to the next-lowest integer number.

At the decoder, the spatial comfort noise signal is then generated by mixing three uncorrelated Gaussian noise signals per frequency band based on the decoded coherence measure. Let $n_1(k) \sim \mathcal{N}(0, 1)$, $n_2(k) \sim \mathcal{N}(0, 1)$, $n_3(k) \sim \mathcal{N}(0, 1)$, then the mixed output noise signals for the left and right channels, respectively, are given as

$$n_{1,i}(k) = n_1(k) \cdot \sqrt{1 - \hat{c}_b[i]} + n_2(k) \cdot \sqrt{\hat{c}_b[i]}, \quad (18)$$

$$n_{r,i}(k) = n_3(k) \cdot \sqrt{1 - \hat{c}_b[i]} + n_2(k) \cdot \sqrt{\hat{c}_b[i]}. \quad (19)$$

The jointly encoded noise estimates are dequantized and the inverse M/S transform is applied to obtain the dequantized noise estimates in their original left/right representation:

$$\hat{\sigma}_{\text{est},l,\text{dB}}(k) = \hat{\sigma}_{\text{est},\text{mid}}^2(k) + \hat{\sigma}_{\text{est},\text{side}}^2(k), \quad (20)$$

$$\hat{\sigma}_{\text{est},r,\text{dB}}(k) = \hat{\sigma}_{\text{est},\text{mid}}^2(k) - \hat{\sigma}_{\text{est},\text{side}}^2(k), \quad (21)$$

where $\hat{\sigma}_{\text{est},\text{mid}}^2(k)$ and $\hat{\sigma}_{\text{est},\text{side}}^2(k)$ are the vector-quantized jointly-coded noise estimates obtained in (5) and (6). The reconstructed left and right noise shape parameter vectors are then converted back to the linear domain via

$$\hat{\sigma}_{\text{est},l}^2(k) = 10^{\hat{\sigma}_{\text{est},l,\text{dB}}(k)/10}, \quad (22)$$

$$\hat{\sigma}_{\text{est},r}^2(k) = 10^{\hat{\sigma}_{\text{est},r,\text{dB}}(k)/10}. \quad (23)$$

Afterwards, encoder- and decoder-side estimates are combined and energy spectral densities $\sigma_{l,\text{CNG}}^2, \sigma_{r,\text{CNG}}^2$ are derived from the combined noise estimates, as in the mono case. The left and right channel comfort noise spectra are defined as

$$N_l(m) = \sqrt{\frac{\sigma_{l,\text{CNG}}^2(m)}{2}} (n_{1,i}(2m) + j n_{1,i}(2m+1)), \quad (24)$$

$$N_r(m) = \sqrt{\frac{\sigma_{r,\text{CNG}}^2(m)}{2}} (n_{r,i}(2m) + j n_{r,i}(2m+1)), \quad (25)$$

analogously to (1). Similarly to the mono case, frequency bins below 50 Hz and the imaginary part at nyquist frequency are set to zero. Both channels are then transformed back to the time domain as in the mono case to generate the stereo comfort noise signal.

Transmitting two sets of noise estimates (in a joint representation) and generating two separate comfort noise signals directly without relying on any upmix parameters are design choices that allow for easy integration into the high-bitrate stereo coder of IVAS [21], which applies joint stereo coding techniques and does not rely on a downmix/upmix scheme for coding the stereo signal. Following similar paradigms for coding of the active signal and the background noise signal allow for reusing already existing functionality and avoid the need for any adaptation processing when switching between the two coding modes.

The stereo SID frame for the proposed implementation consists of 104 bits, corresponding to a bitrate of 5.2 kbps. Specifically, the SID frame contains:

- 8 bits of IVAS-specific side information[25].
- 44 bits for the MSVQ-coded mid noise estimates (7 bits for the global gain, 37 bits for the residual shape).
- 32 bits for the MSVQ-coded side noise estimates (7 bits for the global gain, 25 bits for the residual shape).
- 20 bits for the five coherence values (4 bits per value).

III. EVALUATION

A. Listening Test

The overall performance of the proposed stereo DTX system as part of the IVAS stereo coder was assessed during the characterization of the codec, where it achieved very good quality [26]. Specifically, as can be seen in the plots of the characterization results for experiment P800-21 in [26], the quality of the stereo coder for 48 kbps with DTX enabled is very close to the score of the uncoded signal. However, this overall score is influenced by the active speech parts of the signal, so it does not isolate the quality of the DTX system itself. To specifically assess the performance of the proposed stereo DTX system, a listening test was conducted where it was compared with a baseline DTX system that does not take into account spatial characteristics of the background noise when generating comfort noise. The results of this listening test are described in the following.

The test compares the stereo DTX system with a baseline “dual-mono” DTX system. This system can be seen as a “naïve” implementation of a stereo DTX system, where the VADs of the two channels are synchronized with the test system, and through two separate SID frames, the comfort noise of each channel is generated independently. In relation to the proposed system, this means that no joint coding of the noise shape parameters is applied, both channels’ noise estimates are vector-quantized as in the mono case (using the full-stage, 37-bit MSVQ) and comfort noise generation is done according to (19), but with $\hat{c}_b[i] = 0 \forall b, i$. The baseline was implemented in the same IVAS framework, which

TABLE III
BACKGROUND NOISE USED IN THE EVALUATION

Noise Type	Description
office	Recorded in a large office room with people working, including typing noises and conversations in the background.
traffic	Signals obtained from DEMAND database [28], specifically t1 from TMETRO and t2 from STRAFFIC. Converted to two-channel format by selecting microphones 5 and 7.
nature	Signals obtained from DEMAND database [28], specifically n1 from NRIVER and n2 from NFIELD. Converted to two-channel format by selecting microphones 5 and 7.
car	Recorded in a VW Passat driving on a highway.
artificial	Brownian noise generated using a computer program.

allowed for comparison of the CNG mechanism only, since this way coding of the active speech parts of the signal is identical between the two systems. Both conditions used an active coding bitrate of 48 kbps. Also, the synchronized VAD classification ensures active coding and CNG to be done at the same time in both systems under test, which makes comparing the two systems easier for the participants.

The listening test was conducted as a paired comparison test with explicit reference. In each trial, subjects were presented with the open uncoded reference signal and the two signals under test, which were presented blindly and in random order under the labels “A” and “B”. The task was to rate signal A compared to signal B on a bipolar discrete seven-grade comparison scale[27] with respect to which of them sounds more similar to the reference signal. The trial order was randomized per subject and the subjects could repeatedly listen to all presented signals and set loops around parts of the signal. Signals were presented on headphones.

In total, 12 subjects took part in the listening test. All of them were professionals working in the field of audio signal processing or audio coding and were familiar with the test methodology and critical assessment of audio signals in general. Before conducting the actual experiment, every subject completed a short training session. The subjects were instructed to focus primarily on the background noise parts of the presented signals; exact instructions can be found in the appendix.

1) *Test material:* The test signals used in the listening test should cover different telecommunication scenarios. Possible end devices for immersive audio communication can vary from smartphones to larger conferencing systems. The former will record signals with a rather small microphone distance while for the latter microphones might be significantly further apart. The type of the background noise itself will differ depending on the recording environment. To account for all of this, different types of recorded and artificially generated background noise signals were used as listed in Table III. Half of the recorded items used a microphone distance of 100 cm while the other half used a microphone distance of 10 cm. Two items used artificially created noise, one of them with the same noise signal in both channels and one with two different, mutually uncorrelated noise signals in each channel.

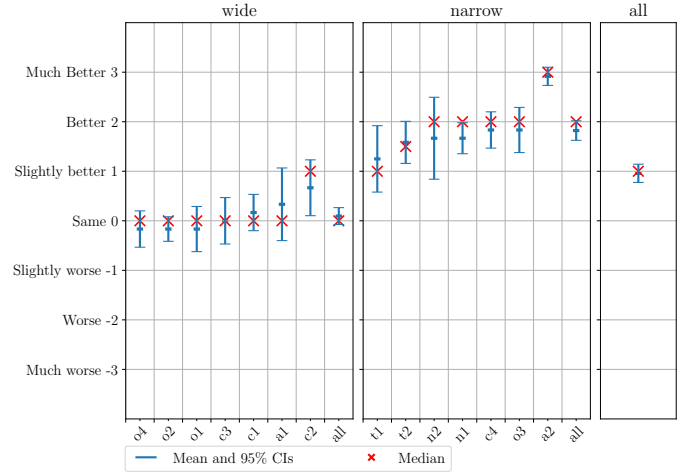


Fig. 2. Mean values and 95% confidence intervals assuming Student’s-T distribution (blue) and median values (red) for subject’s ratings split by microphone distance. Positive ratings indicate preference for the stereo DTX system. Microphone distance suffix omitted from the x-axis labels as already noted in subplot titles.

The artificial test items were included to serve as extreme cases with very high and very low inter-channel correlation. The listening test signals were created by combining clean speech recordings of both female and male speakers with the different background noise signals, where speech and noise were recorded using the same microphone setup. The final combined signals were 16 seconds long and contained two short sections of active speech (both about 3.5 seconds long) with sections of only background noise at the start and the end of the signal as well as in between the active parts. Signal-to-Noise-Ratio (SNR) was between 5 dB and 10 dB, keeping the background noise clearly audible relative to the overlaid speech, so that subjects could easily follow and focus on the background noise and on possible differences between CNG and actively coded signal parts. It was checked that DTX was triggered for both conditions in all background noise sections of all items for substantial portions of the signal, especially around the parts containing active speech. This way, it was ensured that there actually were comfort noise sections and transition to and from actively coded parts for the subjects to assess.

2) *Results:* Figure 2 shows mean values and 95% confidence intervals (CIs) and medians of the subjects’ ratings for the individual items as well as aggregated over all items. The figure is split by microphone distance of the respective items – the left subfigure only includes items with wide microphone distance (i.e., 100 cm and the artificial background noise with uncorrelated channels), the mid subfigure only includes the items with narrow microphone distance (i.e., 10 cm and artificial background with same noise signal in both channels) and the right subfigure shows scores aggregated over all items. Higher values indicate preference for the stereo DTX condition over the baseline condition.

Over all items (Figure 2, right), a clear tendency towards

preference for the stereo DTX condition can be observed. This is also backed by the result of a one-sided Wilcoxon signed-rank test with results $T = 6614$ and $p \approx 7.5 \cdot 10^{-16}$, indicating a statistically significant difference between the two tested conditions. An overall mean difference rating of 0.95 and a median of 1 are observed, corresponding to the label "Slightly better" on the rating scale.

When examining the data split into the two groups of wide and narrow microphone distance (Figure 2, left and mid), one can observe that the two groups of items were rated differently by the subjects. The stereo DTX condition was clearly preferred for the narrow items with an overall median of 2 and a mean of 1.82. For the wide items group, a median of 0 and a mean of 0.095 were observed, indicating no preference towards one of the two systems for this group of items. Only for the item *c2-wide*, a slight preference towards the stereo DTX system can be observed with a median of 1 and a mean of 0.67. A Mann-Whitney U test was performed between the two groups of microphone distances and the result was significant ($U = 553, p \approx 2.5 \cdot 10^{-22}$).

B. Coherence Analysis

For the signals used in the listening test, the inter-channel coherence in the output signals was analyzed. Only DTX frames were included in the analysis. In each DTX frame, the broadband inter-channel coherence was calculated in the same way as given by Equation 7, but in a broadband fashion, i.e. computed over the whole frequency range without splitting into bands. Also, no smoothing is applied.

Figure 3 shows the difference between the measured per-frame coherence for the two systems and the original input signals. As for the listening test results, the data is split into two groups based on the microphone distance. For the narrow items group (Figure 3, mid), the stereo system shows a smaller coherence difference than the baseline system for all items in the group. The data for the stereo system also consistently shows less spread than the data for the baseline. For the wide items group (Figure 3, left), coherence differences are much more similarly distributed between the two systems. The stereo system shows smaller absolute medians for all items and overall in this group, but by a much smaller extent than seen in the narrow items group.

C. Computational Complexity

Complexity was assessed in terms of Weighted Mega-Operations per Second (WMOPS) using the WMC tool from the ITU-T standard library[29]. The measurement was run on a long concatenated audio file consisting of the background noise files used for generating the listening test items. The whole background noise database amounted to a complete length of about one hour and 45 minutes. Measurements were run at 48 kbps for the stereo system and the baseline system and at 24.4 kbps per channel for dual-mono EVS (48.8 kbps in total, closely matching the stereo bitrate), both using 48 kHz input signals. Figure 4 shows average complexity numbers for the three systems, split by frame type (active and inactive).

The inactive category comprises both SID frames and non-transmitted frames ("NO_DATA").

The stereo system and the baseline system show very similar complexity numbers. For active frames, they are indeed identical, which is expected since the active frame processing is identical between the two systems. Both for the stereo system and for EVS, DTX operation enables a significant average complexity reduction. At the encoder, inactive frames are only about half as complex as active frames (50.8% for the stereo system and 47.2% for EVS). EVS also shows a reduction for the decoder for which inactive frames are only 67.4% as complex as active frames. The stereo system does not show a relevant reduction at the decoder, but is already less complex in active frames than EVS.

D. Discussion

The subjective listening test results indicate that the stereo DTX system overall outperforms the baseline system and can handle different types of background noise with varying degrees of inter-channel correlation better than the baseline. For stereo signals with wider microphone distance, the background noise signal is in general expected to be less coherent between the channels. This scenario can apparently be handled similarly well by the baseline system – which always produces comfort noise with low inter-channel correlation – and the stereo DTX system – which can adapt the inter-channel correlation of its output based on the coherence measure derived at the encoder. However, for stereo signals recorded with a narrower microphone setup, the baseline system is unable to model the inter-channel correlation of the input background noise, while the stereo DTX system can adapt to this situation and generate stereo comfort noise of similar coherence as the input background noise. This is reflected in the statistically significant tendency towards the stereo DTX system for the narrow items group. Overall, the stereo DTX system always performs at least on par with the baseline system (for stereo signals with low inter-channel correlation) and achieves significantly better results for certain types of signals (with higher inter-channel correlation), leading to an overall slight preference over the baseline system.

Results from inter-channel coherence measurements correspond well with the subjective listening test results. For the items with narrow microphone distance, the stereo system shows significantly less coherence error than the baseline, which explains the subjective preference seen in the listening test. For the wide items group, in contrast, median and spread of the coherence error are mostly comparable between the two systems; accordingly, the subjects' ratings do not favor either system for this group. The generally higher variance in that group indicates that for such items neither approach reliably matches the target coherence in all cases; the plots reveal a common tendency towards underestimation (negative Δ coherence).

In communication scenarios, different stereo input sources can be encountered. Smartphones have become ubiquitous and recent devices are capable of recording stereo and spatial au-

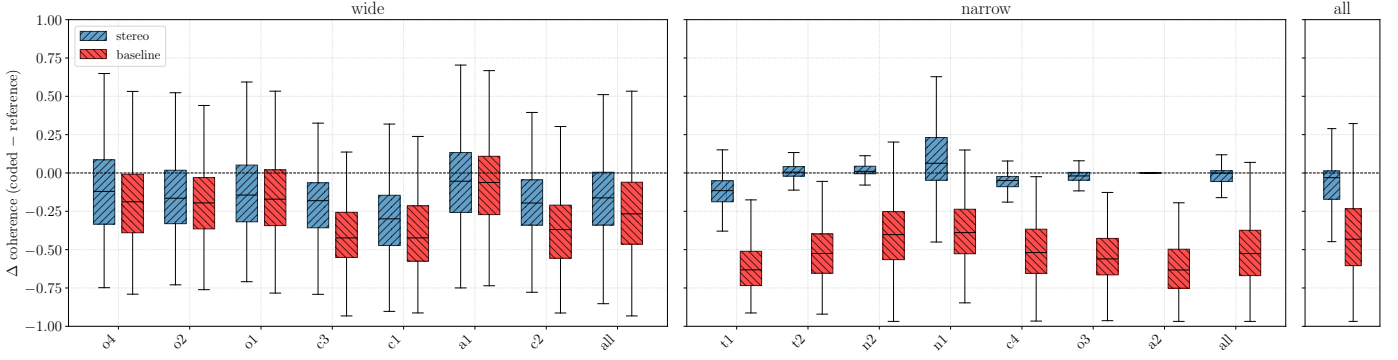


Fig. 3. Box plots of the difference in measured broadband inter-channel coherence between each system's output and the original input signal (Δ coherence), shown separately for the proposed stereo and the baseline system, for the signals used in the listening test. Only DTX frames were considered in the analysis. Boxes span from first (Q_1) to third (Q_3) quartile. Horizontal whiskers span from $Q_1 - 1.5IQR$ to $Q_3 + 1.5IQR$ with $IQR = Q_3 - Q_1$. Horizontal bars inside the boxes denote the median.

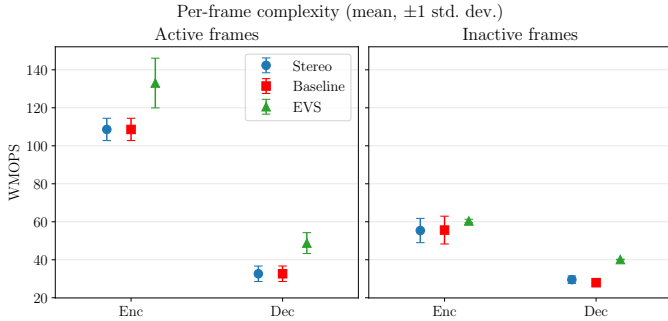


Fig. 4. Complexity numbers in Weighted Mega-Operations per Second (WMOPS) for the stereo system, the baseline system and dual-mono EVS.

dio. Microphone distances in smartphones will necessarily be rather narrow, whereas stationary conferencing systems may use large distances. The presented results show that the stereo system is versatile enough to address these different stereo input sources better than the baseline system. Together with the IVAS characterization testing [26], in which the stereo system achieved results close to the uncoded original signal on an absolute quality scale, the small coherence error for the narrow microphone distance items indicates that the system models and adapts to the background noise's inter-channel coherence well and produces spatially accurate comfort noise. For items with wide microphone distance, coherence modelling is less accurate, but this appears to have limited influence on perceptual quality relative to the baseline, which does not model inter-channel coherence at all.

Complexity-wise, stereo and baseline system show comparable numbers. For inactive frames, the stereo system's decoder shows a slightly higher mean value, which can be attributed to the added complexity of the stereo CNG algorithm; however, the difference is small and still within the spread bars. Overall, DTX operation achieves a significant reduction in complexity over active stereo signal coding, and the complexity of the stereo system for inactive frames is

comparable to dual-mono EVS, with a slight tendency towards being less complex. Note that the comparison to EVS is not entirely straightforward, as the per-channel signal pre-processing differs and EVS switches between two DTX/CNG mechanisms depending on the input signal [4]. These differences make it difficult to attribute the average complexity differences seen for inactive frames to the stereo-related contributions only. They do indicate, however, that the proposed system offers complexity comparable to an existing stereo communication codec while adding stereo-aware DTX functionality.

IV. CONCLUSION

A novel DTX system for stereo audio communication codecs is proposed. It addresses the need to model spatial characteristics of the background noise in CNG for spatial audio signals by introducing a stereo CNG method that adapts to the inter-channel coherence of the background noise. The system was evaluated in a listening test by comparing it to a baseline "dual-mono" DTX system. Results showed that the proposed system always performs at least as well as the baseline and outperforms it for background noise signals recorded with narrow microphone distance. These subjective evaluation results were complemented by measurements of the inter-channel coherence error with respect to the original background noise signals. The measurements revealed that the proposed system can model the inter-channel coherence of the input background noise better than the baseline system with a small median error for signals exhibiting high inter-channel coherence. In conclusion, its ability to adapt to varying degrees of inter-channel coherence in stereo background noise allows the proposed system to outperform the baseline system, especially for high inter-channel coherence signals, such as ones recorded with narrow microphone spacing. The system has been integrated into the high-bitrate stereo coding mode of the IVAS codec. Results from IVAS characterization tests indicate that it achieves good quality and generates perceptually accurate stereo comfort noise. Complexity measurements

showed that DTX operation allows for significant average complexity reduction and that complexity is comparable to dual-mono EVS.

ACKNOWLEDGMENT

We dedicate this work to the memory of our colleague Dr. Emmanuel Ravelli, who passed away before its completion. His ideas and contributions laid the groundwork for this research.

REFERENCES

- [1] ETSI, "Digital cellular telecommunications system; discontinuous transmission (DTX) for enhanced full rate (EFR) speech traffic channels," ETS 300 729 (GSM 06.81), version 5.2.1, Release 1996, 1997.
- [2] M. Polignano, D. Vinella, D. Laselva, J. Wigard, and T. B. Sorensens, "Power savings and QoS impact for VoIP application with DRX/DTX feature in LTE," in *Proc. IEEE 73rd Veh. Technol. Conf. (VTC Spring)*, 2011, pp. 1–5.
- [3] P. Frenger, P. Moberg, J. Malmolin, Y. Jading, and I. Gódor, "Reducing energy consumption in LTE with cell DTX," in *Proc. IEEE 73rd Veh. Technol. Conf. (VTC Spring)*, 2011, pp. 1–5.
- [4] 3GPP, "EVS codec detailed algorithmic description," Technical Specification (TS) 26.445, version 18.0.0, 2024. [Online]. Available: <http://www.3gpp.org/DynaReport/26445.htm>
- [5] 3GPP, "IP multimedia subsystem (IMS); multimedia telephony; media handling and interaction," Technical Specification (TS) 26.114, version 18.9.0, 2024. [Online]. Available: <http://www.3gpp.org/DynaReport/26114.htm>
- [6] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wide-band speech codec (AMR-WB)," *IEEE transactions on speech and audio processing*, vol. 10, no. 8, pp. 620–636, 2002.
- [7] J. Makinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, "AMR-WB+: A new audio coding standard for 3rd generation mobile audio services," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 2005, pp. ii–1109.
- [8] D. Virette, Y. Lang, L. Miao, W. Wu, B. Kövesi, C. Lamblin, and S. Ragot, "G.722 annex d and g.711.1 annex f - new ITU-T stereo codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 528–532.
- [9] M. Lutzky, M. Luis Valero, M. Schnell, and J. Hilpert, "AAC-ELD v2 - the new state of the art in high quality communication audio coding," in *Proc. 131st Audio Eng. Soc. Conv.*, 2011.
- [10] J.-M. Valin, K. Vos, and T. B. Terriberry, "Definition of the Opus audio codec," Internet Engineering Task Force (IETF), RFC 6716, Sep. 2012.
- [11] M. Multus, S. Bruhn, J. Torres, E. Fotopoulou, T. Toftgård, E. Norvell, S. Döhla, Y. Gao, H.-y. Su, L. Laaksonen *et al.*, "Immersive voice and audio services (IVAS) codec—the new 3GPP standard for immersive communication," in *Proc. 157th Audio Eng. Soc. Conv.*, 2024.
- [12] L. Turchet, M. Lagrange, C. Rottondi, G. Fazekas, N. Peters, J. Østergaard, F. Font, T. Bäckström, and C. Fischione, "The internet of sounds: Convergent trends, insights, and future directions," *IEEE Internet Things J.*, vol. 10, no. 13, pp. 11 264–11 292, 2023.
- [13] L. Turchet, C. Fischione, G. Essl, D. Keller, and M. Barthet, "Internet of musical things: Vision and challenges," *IEEE Access*, vol. 6, pp. 61 994–62 017, 2018.
- [14] E. Fotopoulou, K. Sagnowski, K. Prebeck, M. Chakraborty, S. Medicherla, and S. Döhla, "Use-cases of the new 3GPP immersive voice and audio services (IVAS) codec and a web demo implementation," in *Proc. IEEE 5th Int. Symp. Internet Sounds (IS2)*, 2024, pp. 1–6.
- [15] L. Vignati, G. Nardini, M. Centenaro, P. Casari, S. Lagén, B. Bojovic, S. Zambon, and L. Turchet, "Is music in the air? evaluating 4G and 5G support for the internet of musical things," *IEEE Access*, vol. 12, pp. 38 081–38 101, 2024.
- [16] L. Turchet, C. Rinaldi, C. Centofanti, L. Vignati, and C. Rottondi, "5G-enabled internet of musical things architectures for remote immersive musical practices," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 4691–4709, 2024.
- [17] L. Gabrielli and L. Turchet, "Towards a sustainable internet of sounds," in *Proc. 17th Int. Audio Mostly Conf.*, 2022, pp. 231–238.
- [18] K. Kurozumi and K. Ohgushi, "The relationship between the cross-correlation coefficient of two-channel acoustic signals and sound image quality," *J. Acoust. Soc. Amer.*, vol. 74, no. 6, pp. 1726–1733, 1983.
- [19] J. Herre and S. Dick, "Psychoacoustic models for perceptual audio coding—a tutorial review," *Appl. Sci.*, vol. 9, no. 14, p. 2854, 2019.
- [20] A. Lombard, S. Wilde, E. Ravelli, S. Döhla, G. Fuchs, and M. Dietz, "Frequency-domain comfort noise generation for discontinuous transmission in EVS," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2015, pp. 5893–5897.
- [21] G. Marković, E. Fotopoulou, J. F. Kiene, and C. R. Helmrich, "Efficient MDCT-based multi-channel coding with perceptual whitening and broadband ILD compensation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2025, pp. 1–5.
- [22] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *J. Acoust. Soc. Amer.*, vol. 33, no. 2, pp. 248–248, 1961.
- [23] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.
- [24] W. P. LeBlanc, B. Bhattacharya, S. A. Mahmoud, and V. Cuperman, "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 4, pp. 373–385, 1993.
- [25] 3GPP, "Codec for immersive voice and audio services (IVAS); detailed algorithmic description including RTP payload format and SDP parameter definitions," Technical Specification (TS) 26.253, version 19.3.0, 2026. [Online]. Available: <http://www.3gpp.org/DynaReport/26253.htm>
- [26] 3GPP, "Codec for immersive voice and audio services (IVAS); performance characterization," Technical Specification (TS) 26.997, version 19.1.0, 2026. [Online]. Available: <http://www.3gpp.org/DynaReport/26997.htm>
- [27] ITU-R, "General methods for the subjective assessment of sound quality," Recommendation BS.1284-2, 2003. [Online]. Available: <https://www.itu.int/rec/R-REC-BS.1284/en>
- [28] J. Thiemann, N. Ito, and E. Vincent, "Demand: a collection of multi-channel recordings of acoustic noise in diverse environments," Jun. 2013. [Online]. Available: <https://doi.org/10.5281/zenodo.1227121>
- [29] ITU-T, "Software tools for speech and audio coding standardization," Recommendation G.191, 1996.

APPENDIX

A. Listening test instructions

The following instructions were given to the participants of the listening test.

"This listening test assesses the DTX (discontinuous transmission) and CNG (comfort noise generation) algorithm for stereo signals of the IVAS codec. DTX is used to lower the bit rate of a codec during speech pauses by not transmitting the coded background noise, but instead generating comfort noise at the decoder. The generated comfort noise should ideally sound as similar as possible to the original background noise signal.

In this listening test, you will compare two stereo signals which were coded with two different versions of IVAS that differ in their stereo DTX functionality. As quality reference, the uncoded original signal is available for you to listen to. Your task is to rate if the first signal sounds better (that is, more like the reference signal) or worse (that is, less like the reference signal) than the second signal. The signals consist of speech recordings with different types of background noise. Pay special attention to the background noise parts (parts without speech) of the presented items as well as on the transition

sections between noise-only parts and speech parts. Focus on both timbral and spatial differences which might exist between the different signals, especially colouration and stereo width. The tests consists of two parts, a short training

session and the actual experiment. Complete the training session first before starting the second part. Adjust your headphone volume during the training to a comfortable level and don't change it during the actual listening test session."