

未来の没入型オーディオ・システム、コンピュータ・ビジョンが音声の再生を支える

著者：Santosh Singh、コンスーマ・システム・アプリケーション担当シニア・エンジニア
Aravind Navada、コンスーマ・システム・アプリケーション・アジア担当ディレクタ

概要

民生向けのエンターテインメントの分野では、没入型の体験（エクスペリエンス）がより強く求められるようになっていきます。つまり、ユーザが物理的な現実と区別できない形でコンテンツを楽しめるようにすることが重視されています。より優れた没入感を実現する上では、音声が必要な役割を果たすことになります。そのような背景を踏まえ、アナログ・デバイセズは近い将来、音声を再生するための新たな手法が活用されるようになると考えています。その手法は、ビジョン・インテリジェンスを活用したオーディオ・システムとして具現化されるはずです。そうしたシステムを開発するには、人間の脳が音声を処理して局所化する方法について深く理解する必要があります。技術的に言えば、最先端のToF（Time of Flight）イメージャとクラス最高のDSPが重要な役割を果たします。それらの適切な組み合わせにより、次世代の没入型オーディオ・システムを開発するための理想的なプラットフォームが実現されます。

あらゆる新世代の民生向けエンターテインメント機器について語る際には、例外なく没入感という言葉が使われます。それは具体的には、どういう意味なのでしょう。1999年に「マトリックス」という映画がヒットしました。その登場人物であるモーフィアスは、ネオに対し、「匂いを嗅ぎ、味を感じ、物体に触れる」という行為は現実のものなのか」と尋ねます。その上で、ネオがこれまで現実だと思っていたことは、人間の感覚を欺くためにコンピュータによって作られた虚構にすぎなかったという事実を示します。これが、まさに没入感という言葉が意味するところ。すべての人工的な没入体験は、そのような世界が目指すべきゴールだと捉えて開発されています。

体験の中に本当に入りこんでいるような感覚が得られるようにするためには、音声そのものと、それをどのように体験させるのかということが重要になります。それらは、高揚感を演出するため

の非常に重要な要素です。音声は、脳の中の原始的な反応として始まり、任意の状況に対する私たちの最初の反応へと発展していきます。脳は音声を活用することにより、周囲の環境や状況を表す明確なイメージを形成します。人工的に作られた没入体験を脳に信じ込ませることによって意図した没入感を与える上で、音声は非常に重要な役割を果たします。

音声を再生する技術は、長い年月の間に大きな進化を遂げました。当初使われていたのは基本的なモノラル・オーディオ・システムで、オーディオ用のチャンネルは1つしかありませんでした。それが現在では、非常に高度なサラウンド・サウンド・システムが使われるようになっていきます。ホーム・シアターでよく用いられるのは、5.1ch（6チャンネル）、7.1ch（8チャンネル）といった最小構成のシステムです。それに対し、映画館のスクリーン向けには、64以上のチャンネルを備える大規模なシステムが提供されています。ただ、そうしたシステムにおいて、音声に対する空間的な感覚と精度は、スピーカの数とそれらの位置によって制限されます。

次世代の没入型オーディオ・システムは、音声の再生方法に新たな手法を適用することによって実現されるはずです。その手法を実現するためには、人間の脳が音声を処理して局所化する方法に関する深い理解が必要になります。新たなシステムでは、多数のスピーカをリスナーの周りに配置する必要はありません。そのようなことを行わなくても、リスナーの周囲360度に対応した没入感あふれる音声体験がホーム・シアターにもたらされます。そのような新たなシステムを開発するためには、リスナーとそのリスニング環境に関する高度な知識を有していなければなりません。そうした知識が不足していた場合、没入型のオーディオ・システムに求められる要件を満たすことは難しいでしょう。では、次世代の没入型オーディオ・システムを実現するためには、具体的には何が必要なのでしょう。その答えは、音声の再生技術にビジョン・インテリジェンスを融合させるというものです。



現実の世界で音声を自然に耳にすると、人間の脳は、わずか2つのオーディオ信号に基づいて音源に関する空間的な手がかりを導き出します。2つのオーディオ信号の一方は左耳、もう一方は右耳に届きます。これは、人間の両眼による視覚系の仕組みによく似ています。その仕組みとは、左目と右目の視界を比較することによって、脳内で奥行き感を生成するというものです。それと同様に、人間の脳は左耳と右耳に届く音声を処理し、その振幅と遅延時間を比較することによって、音源の位置を大まかに割り出します。この能力は、人類の進化の過程で徐々に発達していききました。人類が自然界で生き延びるためには、そのような能力が不可欠だったからです。

自然な聴覚体験を再現することを目的とした手法に、バイノーラル・オーディオ再生（Binaural Audio Reproduction）というものがあります。これは、信号処理を利用することにより、現実の世界で音を聴いた場合と同じ状態が得られるように、左耳と右耳のそれぞれに向けて2つのオーディオ信号を生成するというものです（図1）。しかし、この手法を具現化するのは非常に難しく、様々な問題が発生します。

まず、バイノーラル・オーディオ再生に使用可能な音声を録音するにはどうすればよいのでしょうか。最も簡単な方法は、2つのマイクを実際に人の左右の外耳道に1つずつ配置し、その位置で音声信号を録音するというものです。この方法をバイノーラル録音と呼びます。音声の再生は、ヘッドフォンを介してリスナーの耳に向けて行われます。では、この手法は狙いどおりに機能するのでしょうか。その答えは、ある条件が満たされるならば「イエス」です。その条件とは、「同じ人を対象として音声の録音と再生が行われる」というものです。言い換えれば、録音時と再生時の対象人物が異なると、狙いどおりに機能しないということです。その原因は、私たちの脳が音声を局所化する仕組みにあります。私たちの頭／耳／体は、周波数領域において特定のシグネチャを放置することにより、脳による音声の局所化プロセスを補助するという形で影響を及ぼします。どのシグネチャが対象になるのかは人によって異なります。その違いは、頭部伝達関数（HRTF: Head-related Transfer Function）によって表されます。バイノーラルの手法を適切に機能させるには、音声を再生する際、リスナーの耳において、HRTFが音声に与える影響を正確に再現しなければなりません。

HRTFについては、リスナーごとに測定を行ってパーソナライズする必要があります。つまり、汎用的なソリューションによって対応できるものではありません。ある人が、別の人のHRTFを使用して生成された音声を聞かされたとします。その場合、人が音声を局所化する能力は著しく低下します。このことは、複数の研究によって確認されています^{1, 2, 3}。

ラウド・スピーカによってバイノーラル・オーディオを再生する際には、更に難易度の高い課題が浮上します。まず、複数のスピーカからの音声信号は互いに干渉を引き起こします。この現象は、クロストーク効果と呼ばれています（図2）。また、リスニング環境によっては、リスナーの耳に届く前に、音声に対して望ましくない影響が及ぶ可能性もあります。

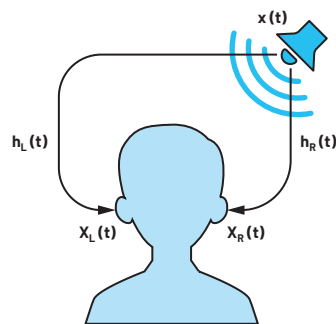


図1. 音源 $x(t)$ からの音声を聞く際の仕組み。 $X_L(t)$ は左耳に届くオーディオ信号、 $X_R(t)$ は右耳に届くオーディオ信号です。

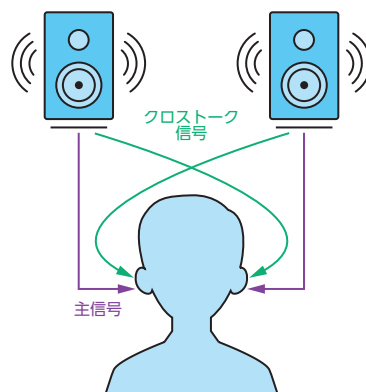


図2. ステレオ・スピーカにおけるクロストーク効果

目標とするのは、自然な聴覚体験を正確に再現することです。スピーカのクロストーク、HRTFのパーソナライズの必要性、リスニング環境の影響は、その実現を阻む主要な要因です。このようなバイノーラル・オーディオ再生が抱える課題を解決するために役立つものがあります。それは、リスナーとリスニング環境に関連して必要になるすべての詳細情報を捉えることが可能なビジョン・システムです。

例えば、コンピュータ・ビジョンのアルゴリズムにデータを供給するカメラを構築したとします。それを利用すれば、リスニング環境の3次元構造の詳細（音声を聞いている部屋の形状、各所表面の幾何学的な測定データ、物体の存在）を捉えることができます。そして、その情報（洞察）を利用すれば、リスニング環境が音声に及ぼす影響を計算することが可能になります。また、音声再生システムにおいて適切な係数を備えるフィルタを使用すれば、その望ましくない影響を排除することができます。実は、この種のシステムは、ホーム・シアターで利用するオーディオ向けのものとしては既に存在していました。従来は、キャリブレーションを実行する際、マイクを使用して音声に対する部屋の影響を把握するということが行われていました。ただ、この手法では、リスナーが音声を聞く位置が変わったり、部屋に構造的な変化が生じたりした場合、処理をやり直す必要がありました。

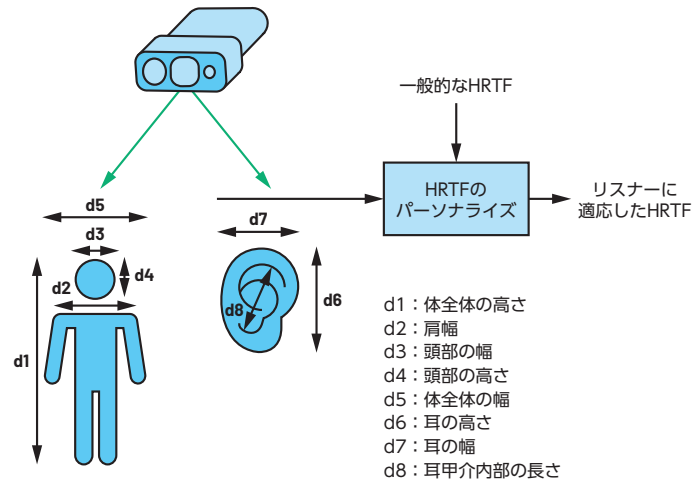


図3. 身体の測定値に基づく
HRTFのパーソナライズ

ビジョン・システムを利用すれば、身体的位置を測定したり、構造に関する詳細情報を把握したりすることができます⁴。それにより、空間的な手がかりを正確に把握し、それをレンダリングすることで、HRTFをパーソナライズするための計算を行うことができます（図3）。例えば、スピーカに対するリスナーの頭の位置と頭部のサイズに関する情報を使用し、クロストークを除去するアルゴリズムを適用するといった処理を行います。それにより、ラウド・スピーカからのバイノーラル・オーディオをリアルタイムにレンダリングすることができます。結果として、リスナーが部屋の中を動き回ったとしても、理想的な音声体験が得られるようになります（図4）。

ビジョン・システムを利用する場合には、ある一般的な問題が浮上します。それは、ユーザのプライバシーの侵害をどのようにして防ぐのかというものです。例えば、エッジの専用プロセッサを使用して、ビジョン・データの分析を行うのであれば、ユーザのプライバシーが侵害されることはありません。その場合、ビジョン・システムのカメラでキャプチャされた情報はリアルタイムに処理されます。それらのデータを保存したり、別のリモート・マシンに転送したりする必要はありません。そのため、プライバシーの問題は生じないということです。

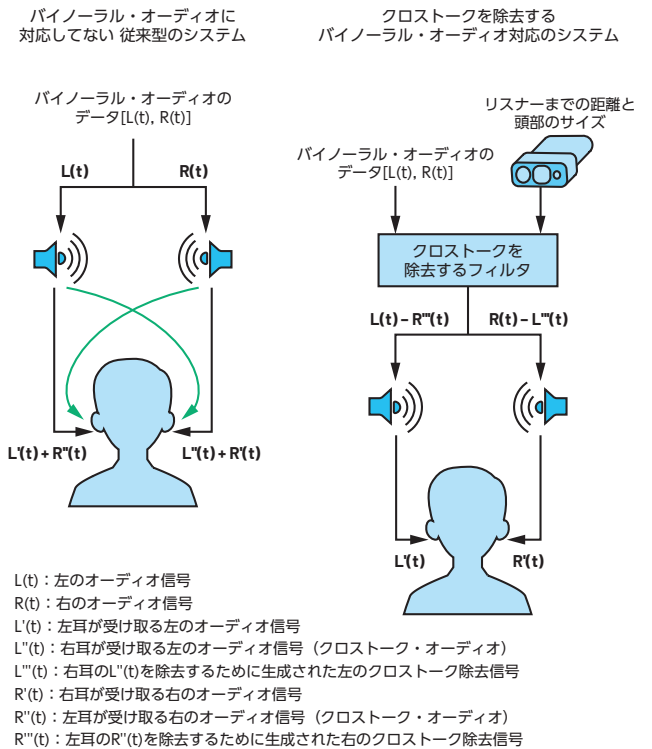


図4. クロストークを除去するための実装方法。
フリー・フィールド・スピーカ・システムによる
バイノーラル・オーディオの再生を実現します。

次世代の没入型オーディオ・システム

▶ アナログ・デバイセズのHDMI、DSP、オーディオ・アンプ、ToFセンサーが基盤

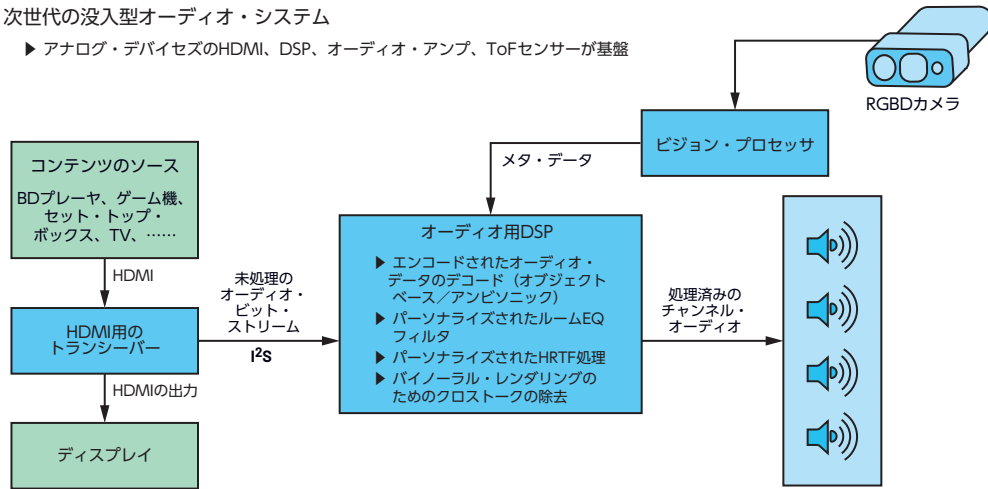


図5. 次世代の没入型オーディオ・システム

ここまで説明したように、次世代の没入型オーディオ・システムを構築するには、ビジョンとオーディオの融合が必要です。つまり、ハードウェア・プラットフォームをベースとして、図5のようなシステムを構築しなければなりません。アナログ・デバイスの場合、最新のマルチコアSHARC[®] DSPと最先端のToFイメージャを主要な構成要素としてプラットフォームを構成しています。

アナログ・デバイセズの「ADSP-SC598」は、2個のSHARCコアと1個のArm[®] Cortex[®]-A55コアを搭載したSoC（System on Chip）です。大容量のオンチップ・メモリと、外付けメモリ用のDDRインターフェースを備えています。また、遅延を小さく抑えつつ、メモリに関連する負荷の大きい演算に対応することが可能です。同製品は、真の没入型オーディオ・システム向けのものとして理想的なプラットフォームだと言えます（図6）。

ADSP-SC598であれば、SHARC DSPが備える演算リソースを使用して負荷のパーティショニングを実現できます。例えば、オーディオ・データのデコードに関連する処理を1つ目のSHARCコアで実行し、オーディオを再生するための後処理とパーソナライズを2つ目のSHARCコアで実行するといった具合です。また、Arm Cortex-A55は、様々な制御の処理に使用できます⁶。図5に示したビジョン・システムは、RGBカメラと深度カメラを組み合わせるか、またはスタンドアロンの深度カメラによって実現できます。アナログ・デバイセズは解像度が1MPのToFイメージャ「ADSD3100」を提供しています。この製品は、mmのレベルの分解能で深度マップをキャプチャし、様々な照明の条件下で機能するように設計されています。これを使用すれば、パーソナライズ用のアルゴリズム（クロストークの除去、ルーム・イコライゼーション [EQ]、HRTFのパーソナライズなど）の適用対象となる非常に精度の高い幾何学的な測定データが得られます。

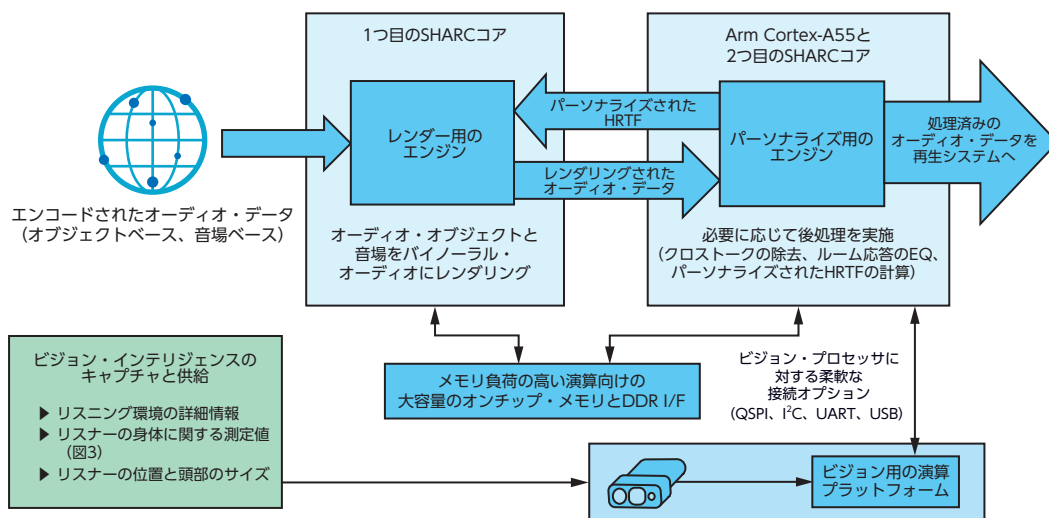


図6. 次世代の没入型オーディオ・システムにおけるパーティショニング

アナログ・デバイセズは、深度の検出を担うToFイメージャとしてADSD3100を採用したToFモジュール「[ADTF3175](#)」も提供しています。その解像度は1MPで、視野角（FOV：Field of View）は75°×75°です。ADTF3175には、ADSD3100用のレンズと光学的バンドパス・フィルタに加え、光学系／レーザー・ダイオードとドライバ／フォトディテクタを備える赤外照射光源、フラッシュ・メモリ、ローカル用の電源電圧を生成するレギュレータが統合されています。このモジュールは、複数の範囲と分解能のモードに対して完全にキャリブレーションされています。

完全な深度検出システムでは、ADTF3175によって得られる未処理のイメージ・データを、ホスト側のシステム・プロセッサまたは深度ISP（Image Signal Processor）によって外部で処理することになります。ADTF3175から出力されるイメージ・データは、MIPI CSI-2（Mobile Industry Processor Interface Camera Serial Interface 2）に対応する4レーンのトランスミッタ・インターフェースを介してホスト・システムに引き渡されます。モジュールのプログラミングと制御には、4線式のSPI（Serial Peripheral Interface）とI²Cのシリアル・インターフェースを使用します。

アナログ・デバイセズは、開発プラットフォーム「[EVAL-MELODY-8/9](#)」、評価用ボード「[EV-2159X/SC59x-EZKIT](#)」、Eclipseベースのエディタ・ツール「[CrossCore® Embedded Studio](#)」も提供しています。これらを利用すれば、ADSP-SC598を使用するアプリケーションのリアルタイムの配備とデバッグを直ちに開始することができます⁷。

[Melodyプラットフォーム](#)は、AVR（AV Receiver）やサウンドバーのアプリケーションに適したシグナル・チェーン・ソリューションです。この統合型のソリューションには、ビデオ、DSP、オーディオ、電力、ソフトウェアの面でアナログ・デバイセズの中でもクラス最高レベルのコンポーネントが盛り込まれています。これを利用すれば、例えば年に1度のアップグレードのスケジュールに合わせて、最新技術を採用した製品を素早く市場に投入することができます⁸。

図7に示したのは、次世代の没入型オーディオ・システムを実現するためのハードウェア・プラットフォームです。ご覧のように、ToFモジュールであるADTF3175をビジョン用の演算プラットフォームに接続し、そのプラットフォームをMelodyのボードに接続しています。なお、RGBカメラとADTF3175を組み合わせれば、ビジョンに関する高度な分析を行うためのRGBD（色＋深度）カメラを構成することが可能です。

まとめ

アナログ・デバイセズは、DSP、HDMI用のトランシーバー、D級アンプ、ToFイメージャなどを含むポートフォリオを有しています。現実世界の音声と区別がつかない音を再生可能な真の没入型オーディオ・システムの実現に向けて、全力で取り組みを行っています。

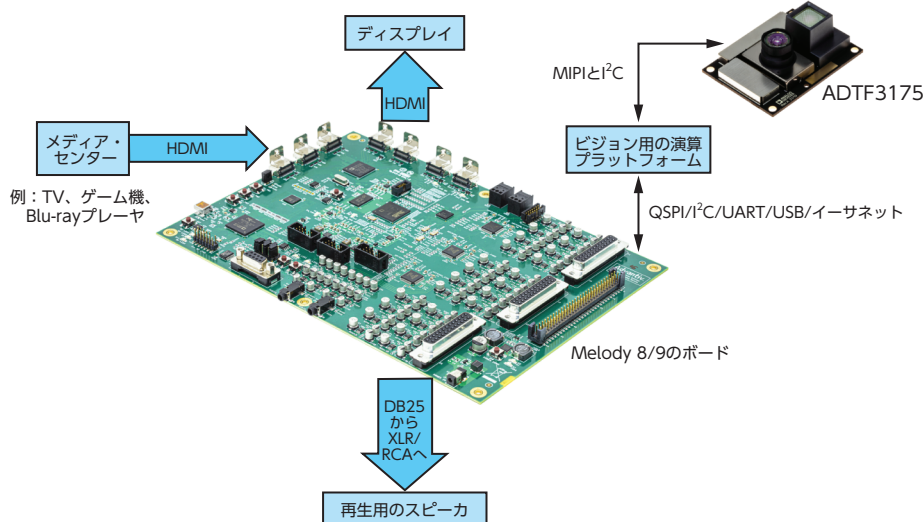


図7. アナログ・デバイセズのプラットフォームによって構成した没入型オーディオ・システム

参考資料

¹ Philipp Paukner, Martin Rothbucher, Klaus Diepold「[Sound Localization Performance Comparison of Different HRTF-Individualization Methods](#) (異なるHRTFのパーソナライズ手法による音声の局所化性能の比較)」 Technische Universität München、2014年4月

² Parham Mokhtari, Ryouichi Nishimura, Hironori Takemoto. 「[Toward HRTF Personalization: An Auditory-Perceptual Evaluation of Simulated and Measured HRTFs](#) (HRTFのパーソナライズに向けて : HRTFのシミュレーション結果と実測値の聴覚知覚評価)」 International Conference on Auditory Display (ICAD)、2008年7月

³ Jenny Claudia, Christoph Reuter 「[Usability of Individualized Head-Related Transfer Functions in Virtual Reality: Empirical Study With Perceptual Attributes in Sagittal Plane Sound Localization](#) (仮想現実におけるパーソナライズされたHRTFの有用性 : 矢状面の音声局所化における知覚属性による実証研究)」 JMIR Serious Games、2020年9月

⁴ Geon Woo Lee, Hong Kook Kim 「[Personalized HRTF Modeling Based on Deep Neural Network Using Anthropometric Measurements and Images of the Ear](#) (身体測定と耳の画像を用いたディープ・ニューラル・ネットワークに基づくパーソナライズされたHRTFのモデル)」 Applied Sciences、2018年11月

⁵ Sanket Nayak, Mitesh Moonat 「[EE-436: Using ADSP-SC59x/2159x High Performance FIR/IIR Accelerators](#) (ADSP-SC59x/2159xによる高性能のFIR/IIRアクセラレータ)」 Analog Devices、2022年6月

⁶ 「[ADSP-SC59x/2159x SHARCシリーズがポートフォリオの性能を2倍に向上し、複雑なオーディオ・アプリケーション向けプラットフォームを実現](#)」 Analog Devices

⁷ 「[CrossCore® Embedded Studio](#)」 Analog Devices

⁸ 「[ホーム・シアターとゲーム](#)」 Analog Devices

著者について

Santosh Singhは、2016年にインドのビルラ工科大学で電子工学と通信工学の学士号を取得しました。インターンを経て、アナログ・デバイセズに入社。デジタル設計者としてキャリアをスタートし、最先端の技術を活用した民生市場向けの様々な製品を担当しました。現在は、シニア・システム・アプリケーション・エンジニアとして、コンシューマ・ビジネス・ユニットに所属。アプリケーションと技術のギャップを埋めるべく、システムの主要な課題を解決することに主眼を置いて業務に取り組んでいます。

Aravind K. Navadalは、アナログ・デバイセズで、アジア全域のお客様を対象とするコンシューマ・システム・アプリケーション・チームを統括しています。20年以上にわたり、ICの設計に従事。様々なミックスド・シグナルIoT製品や民生向けのSoC製品を市場に投入してきました。現在は、製品のターゲットを最大化することを目指し、民生市場向けのプラットフォームとソリューションの開発を担当しています。インドのバンガロールにあるヴィスエソラヤ国立工科大学 (UVCE) で電子工学と通信工学の学士号を取得。米国テキサス大学ダラス校でマイクロエレクトロニクスに関する修士号も取得しました。

EngineerZone®

オンライン・サポート・コミュニティ

アナログ・デバイセズのオンライン・サポート・コミュニティに参加すれば、各種の分野を専門とする技術者との連携を図ることができます。難易度の高い設計上の問題について問い合わせを行ったり、FAQを参照したり、ディスカッションに参加したりすることが可能です。



Visit ez.analog.com

*英語版ソート・リーダーシップ記事は[こちら](#)よりご覧いただけます。