Excitement Through Sound

Voice Interaction comes of age
Controlling consumer electronics with voice commands

October ’13

Uses of Voice

CONEXANT

Voice is increasingly a part of the consumer's everyday electronics experience

Voice Communication: Clear voice quality expected in phones and videoconferencing

Voice Input: As a replacement to typing, into a web browser, for example

Voice Control: Using natural language speech instead of buttons or touch

Smart TV

- Voice Control
- Voice Search
- Voice Chat - Skype

Gaming

- Voice Chat
- Voice Control
- Voice Interactive Games

Smartphone/Tablet

- Voice Communication
- Virtual Assistant (e.g. S-Voice, Siri)

Appliance

- Voice Control (air conditioning on/off, Timer)

Auto

- Voice Input (enter address..etc)
- Voice Communication

STB

- Voice Chat
- Voice Search
- Voice Interactive Games

Voice

Since the dawn of civilization, humans have used voice to communicate with each other…

Human Machine Interface

… yet since the dawn of personal computing, we still interact with machines by point-and-click

1990

2009

Evolution of Voice Input and Control

Voice input and control requires high performance speech recognition

Pre 2008:

Local speech recognition engine running on a PC supported dictation (input) and limited control

Post 2008:

Cloud-based speech recognition engine + Large bandwidth =

Unlimited Natural Language Voice Input and Control

Local speech recognition still used

Market needs support of both cloud & local engines

Examples of Voice Input and Control

CONEXANT

Basic Commands

Only requires limited speech recognition performance, but offers limited user input and control functionality

"On"
"Up"

Natural Language Voice Input and Control

Requires very high performance speech recognition, but offers true natural language speech interface

"I want to watch Minority Report on Netflix"
"Can you recommend a Mexican restaurant in Irvine?"

Voice Pre-Processing enables seamless performance

Natural Language Voice Input and Control

Natural language speech recognition engines are more powerful

Siri from Apple and *Google Now* from Google are examples of applications that use natural language voice input and control

"Near Field"

Natural language voice input and control performance is impressive when speech is captured in a well controlled environment

Natural Language Challenges

Most natural language voice input and control systems tend to fail or become unreliable when used in real-world circumstances

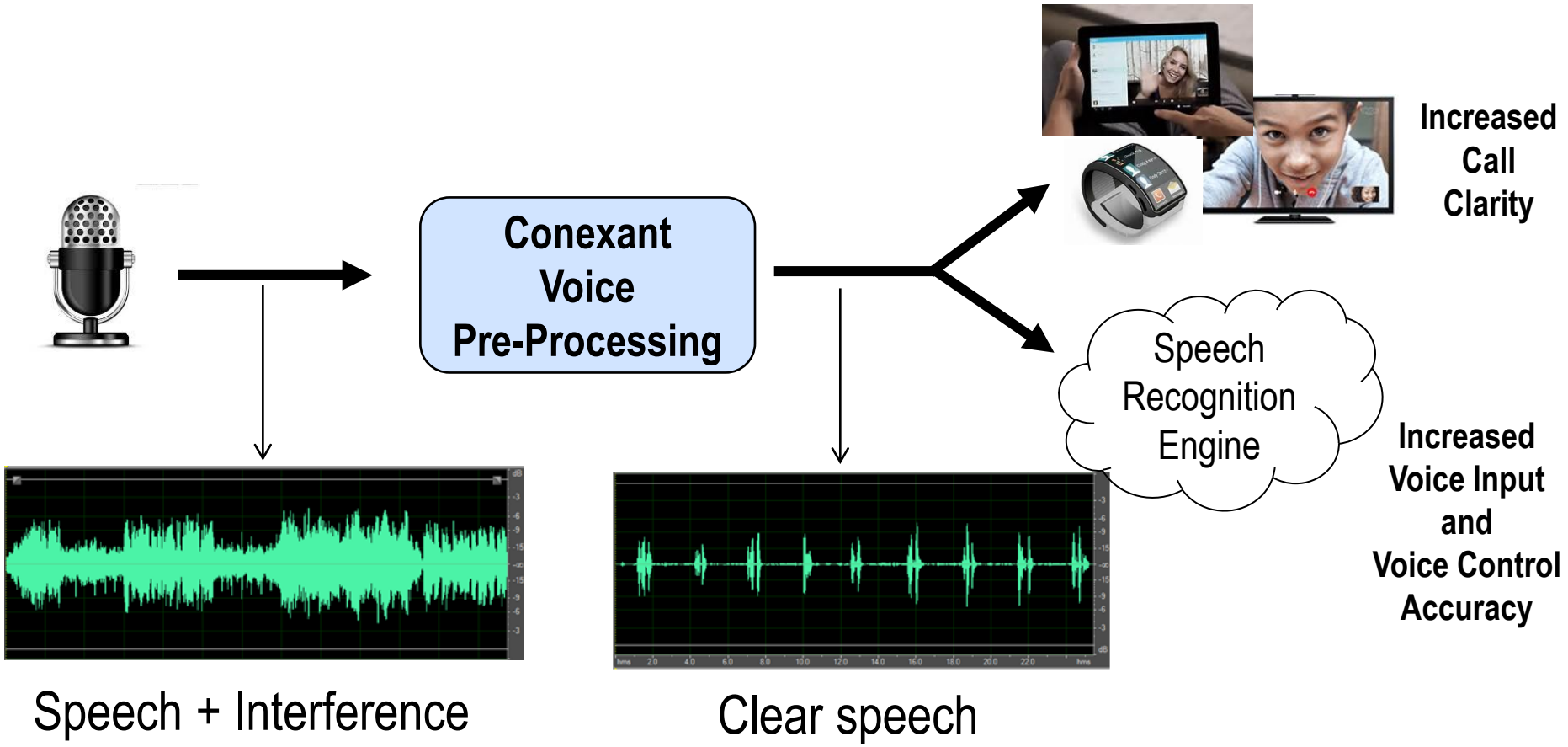
"Far Field"

Interference from the environment and reverberations severely degrade the ability of natural language speech recognition

Overcoming these limitations is necessary to truly realize the promise of natural language voice interaction

Enabling Voice Input & Control in the Real World

Mitigating environmental noise and interference is the key to natural language voice input and control and high quality voice communication in "far field" (real world) conditions



SmartAudio from Conexant

Conexant enables turnkey integration of high quality voice pre-processing for all consumer electronic devices

We offer a portfolio of solutions specifically tailored for clear voice communication and voice input and control

VOICE PRE-PROCESSING TO ENABLE CLEAR VOICE COMMUNICATION AND NATURAL LANGUAGE INPUT AND CONTROL

Genesis of Conexant's Technology

Originated from research in NL Pre-Processing for far field communication and speech recognition

- Far field: User assumed to be 0.5m to 5m from microphone
- Originally targeted control of Smart TVs, Appliances, smart homes and other gadgets as well as enabling clear, hands-free, truly full-duplex communication

Far field scenarios impose many challenges

- Interferences can be at same level or even higher than the desired signal
- Reverb from the room distorts the signal which can have a severe effect on speech clarity and speech recognition performance
- Dynamic range of desired signal and interference can be very large

Conexant's Voice Pre-Processing solution is designed for these challenging conditions

Methods of Voice Pre-Processing

CONEXANT

Blind Source Separation (BSS)

Uses statistical independence to decompose the acoustic scene into their atomic sound components

Computational Auditory Scene Analysis (CASA)

Attempts to model human auditory system and brain

Humans and Computers are fundamentally different

	Humans		Computers
1.	Can integrate a lot of multimodal information (audio, video, …)	1.	Can often access only single modalities
2.	Are continuously learning, memorizing and self-organizing data collected since birth	2.	Cannot access a lot of memory (for HW limitation and cost reasons)
3.	Can do a lot of simple operations but with huge parallelism	3.	Can do quick structured complex mathematical operation (but with limited parallelism)
4.	Describe the world with exemplars	4.	Describe the world with mathematical models

Conexant implements Blind Source Separation modified for real-world robustness

Optimizes for computers' intrinsic attributes to enable super-human capabilities!

Conexant's Approach

- **We use statistical independence to decompose the acoustic scene into their atomic sound components (known as Blind Source Separation)**

- **Our algorithm framework is based on the state-of-art multichannel BSS theory**

- **We overcome robustness issues through a Semi-BSS approach that is conditioned with dynamic models of the acoustic scenes**
  - Traditional BSS approaches lack of robustness in real-world.

- **We do not attempt modeling the human ear we dynamically model the acoustic environment.**

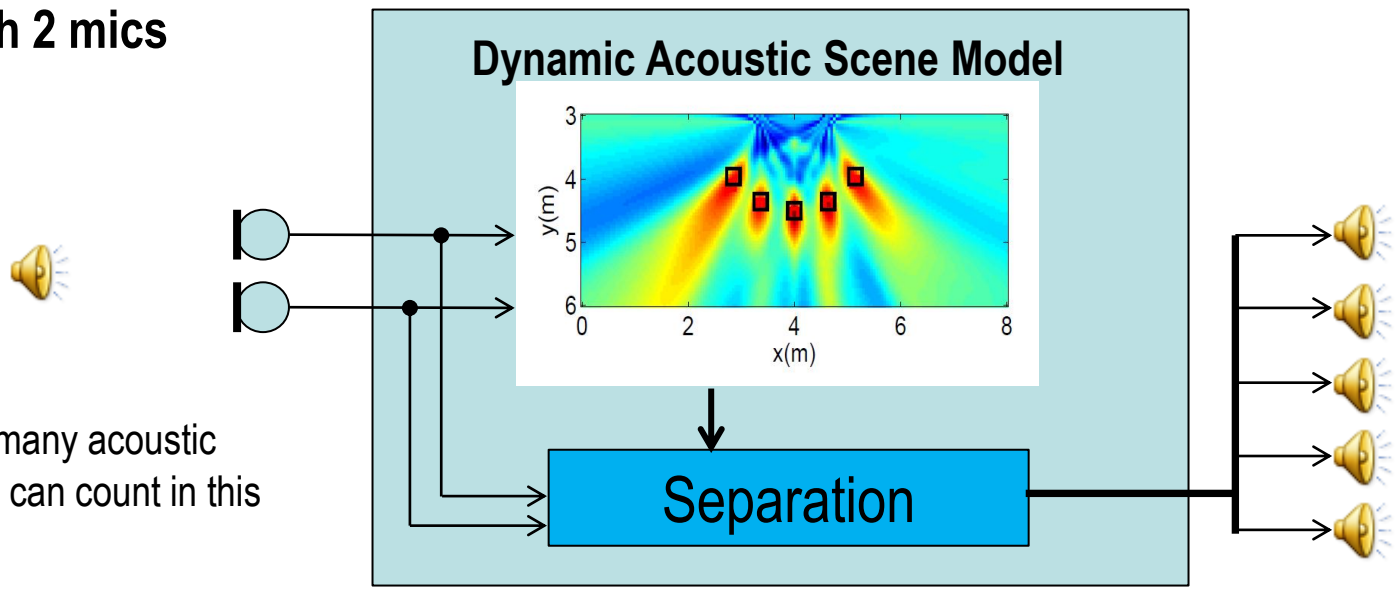
Example of Super-Human Capability

Example: Mixture of 5 voices in a real room

With no visual cues or prior language knowledge, average humans cannot determine the number of voices let alone isolate what each voice is saying

BSS solutions can simultaneously detect and separate multiple sources using only a few microphones.

E.g. with 2 mics



Test: How many acoustic sources you can count in this audio file?

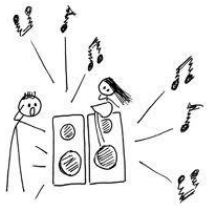
Other Advantages of Conexant Solution

In addition to enabling better call/videoconferencing quality and natural language voice input/control, Conexant's approach offers:

Low Cost, Small Footprint
Flexible implementations using as few as two microphones

Orientation Independence
Clear voice quality regardless of device orientation

Distance and Location Independence
Excellent voice clarity when user is far from device and/or moving relative to the device

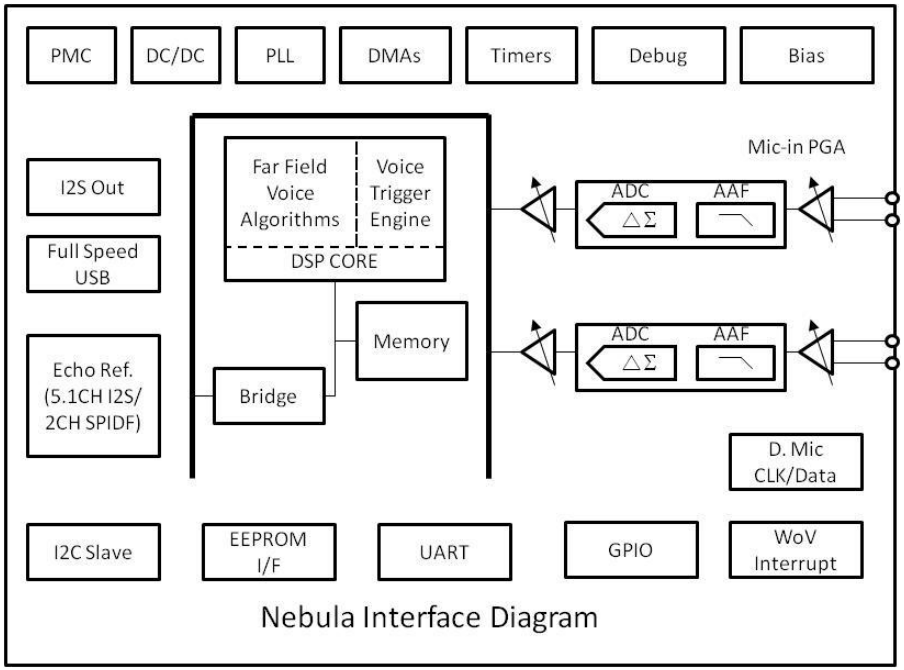


Noise Environment Robustness
Reduces both ambient stationary and non-stationary noise regardless of user position relative to device

HIGH FIDELITY ADC & LOW POWER DSP

Nebula CX2092x Overview

Standalone Subsystem - Far Field Voice Input Processor with integrated ASR Engine



Nebula Interface Diagram

7x7mm 60QFN 0.40mm pitch

- **Lowest System Power Wake on Voice**
- **Command Buffering**
- **Voice Trigger Assisted SSP**
- **Expanded 747kB internal RAM**
- Pre-Tuned Hidden Microphone Module Reference Design
- Most Advanced Far Field Voice Processing
- SWB Watch Live and Talk® experience
- Single Chip, 7x7mm QFN Package, ROHS

Multiple innovative technologies integrated into a standalone device

Nebula 2 Die Solution

Reno:

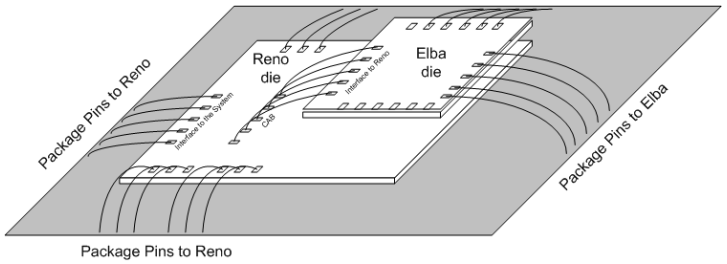
Digital Die
40NM
Supplies:

- 1V Nominal: 200Mhz
- 0.9V Nominal:80~100Mhz
- 3.3V for the oscillator

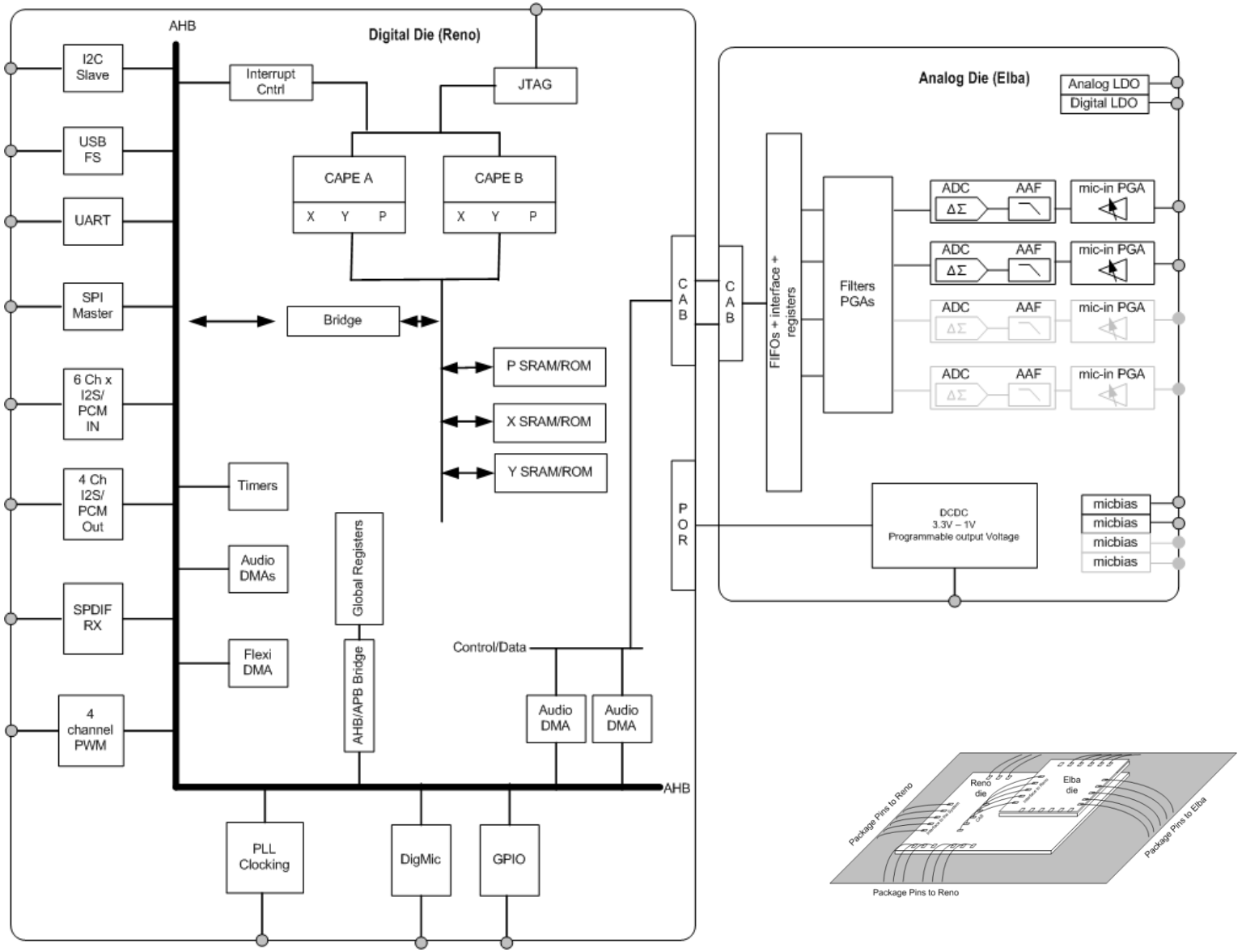
Elba:

Mixed Signal Die
0.18um

- Supplies:
  - Digital: 1.2V
  - Analog: 3.3V



Nebula Block Diagram

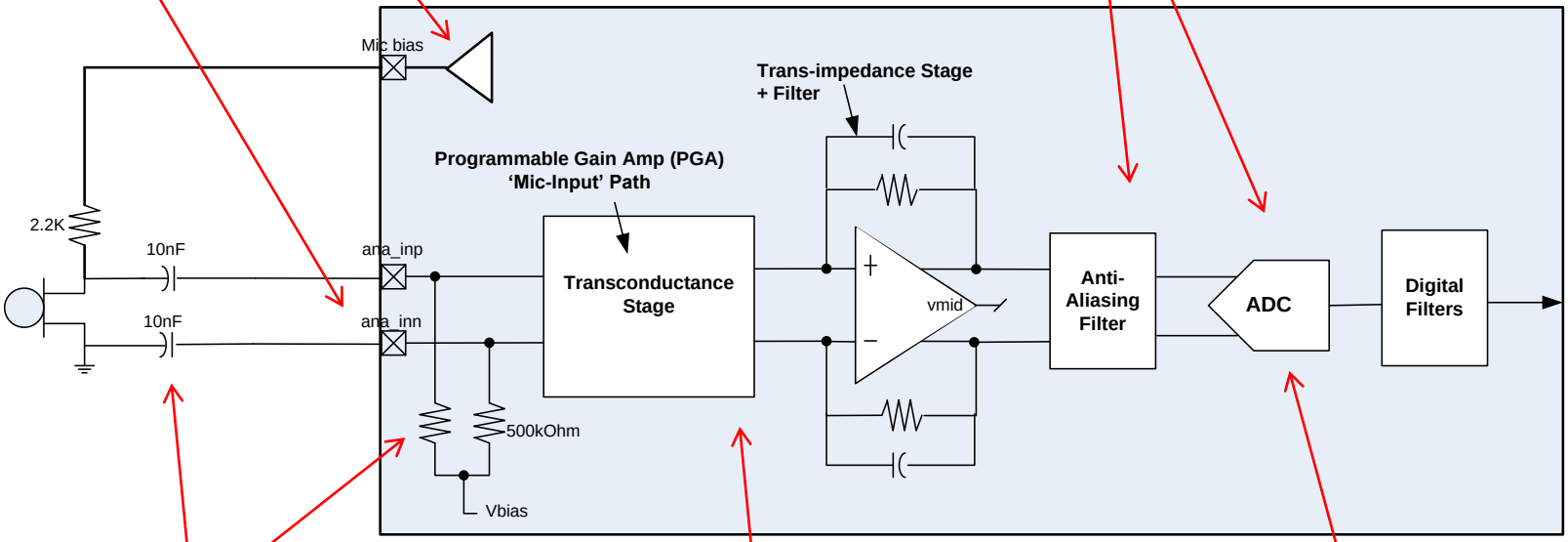


Voice Input ADC - architecture

low-noise microphone bias

signal-adaptive biasing minimizes power consumption

differential inputs for best performance



high input impedance → small external components

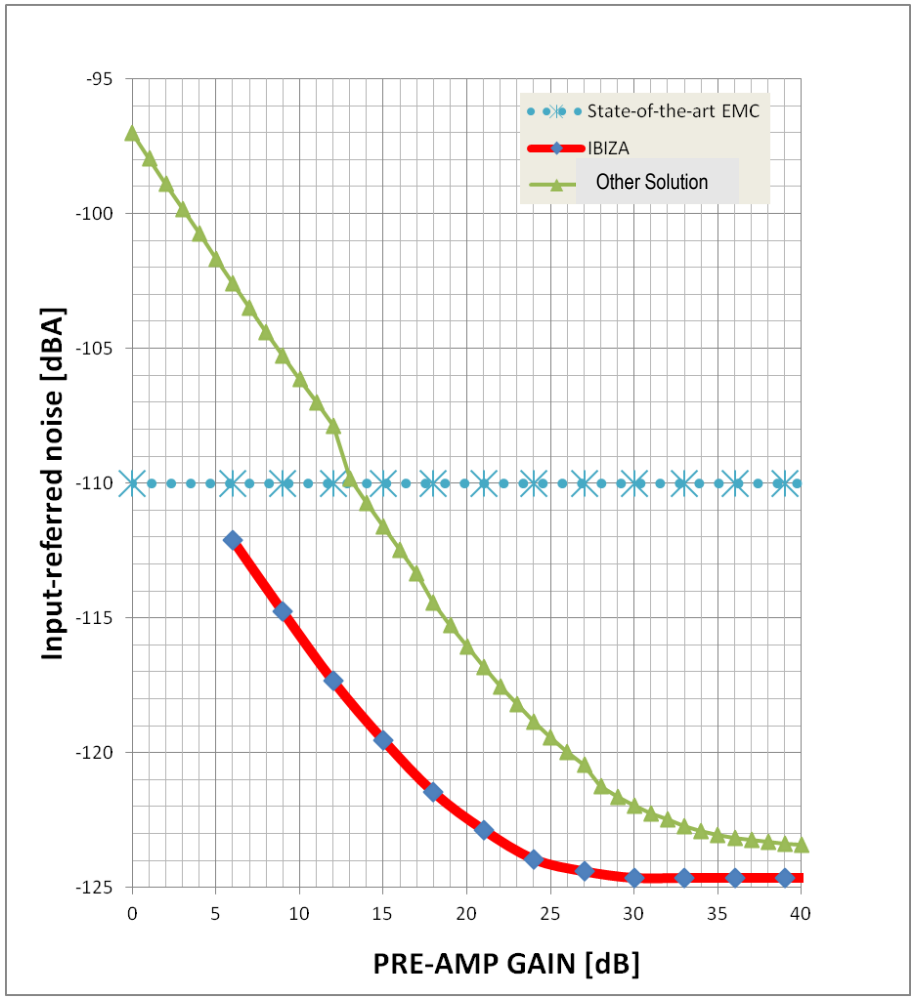
transconductor / opamp architecture achieves lowest noise floor with high input Z

high-performance switched-capacitor sigma-delta ADC

ADC and pre-amp input noise

- Input-referred noise level for ADC is lower than intrinsic noise of most microphones, even at low preamp gain
- High gain in preamp unnecessary
- Use low gain →prevent saturation!

ADC input-referred noise



CONCLUSION

Broad Applications

- Our solution was born as a far field solution
- Personal and Near field are much simpler problems to solve
- For personal and near field we provide superior performance with high SNR gain without speech distortion

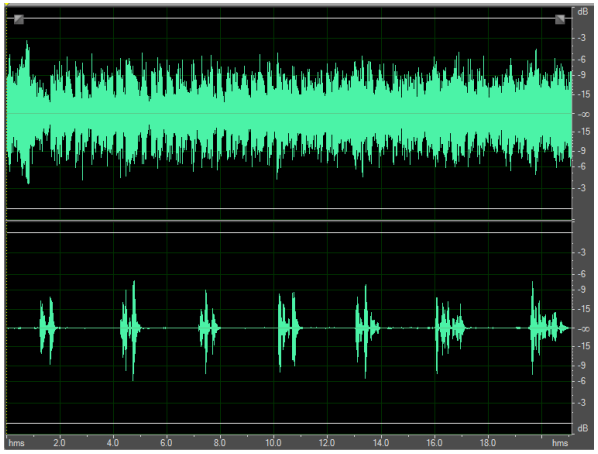
Current Applications

Conexant's technology provides exceptional call clarity and natural language voice input and control in millions of products

- Smart TV
- Mobile Computing
- Appliances
- Communications

TV playing back media content

No Processing



WER = 100%

With Processing

WER = 8.5%

Future Applications

Exceptional call clarity and natural language voice input and control enables new applications

Wearables

Contextual awareness enables a new level of intelligence for "smart" electronics

Future: Conexant is working on it!

?

Thank you

http://www.conexant.com