MediaTek Natural User Interface

MediaTek White Paper

October 2014

Table of Contents

1 Introduction

The extraordinary pace in which computing devices have penetrated into people's everyday lives could not have been imagined by the pioneers of this industry. Computing devices first entered the business workplace. Then, such devices entered the home. After that, people carried mobile computing device around, in their pockets. Finally, we're entering the era in which daily usage gadgets are becoming computing devices. We see the rise of objects that connect to the internet—the internet of things (IoT). They're all computing devices; however, they have different user interfaces.

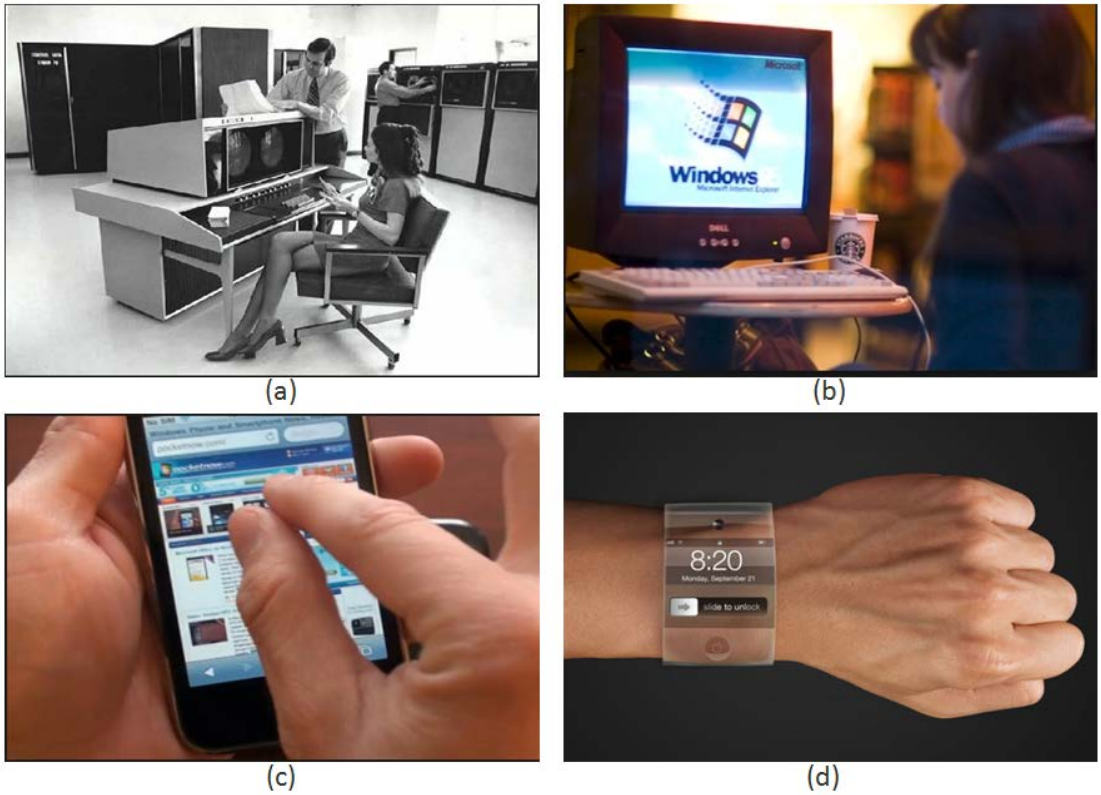

Figure 1. Natural User Interface Progression

The term "natural user interface" for the computing devices has changed as technology has advanced its scope, as is shown in Figure 1 above. Decades ago, natural user interface referred to the keyboard and command-line interface on the screen (Figure 1a). Then, the graphical user interface with mouse (Figure 1b) emerged in the market in the 1980s as the natural interface.

Each new interface mode has brought more versatile computer applications for customer use. Keyboard and command-line interface boosted the productivity of business computing. Graphical user interface and mouse have enabled an extension of computer usage from business into the consumer space at home. The multi-touch screen for mobile devices liberated the way people acquire information. The goal of the touchless user interface, shown in Figure 2 below, is to employ human language to control machines, thereby freeing people from controlling the machine with machine language. In 2007 Apple popularized multi-touch screen for mobile device (Figure 1c). Today touchless user interface is the new natural user interface. (Figure 1d)

Then, what benefits will touchless user interface bring?

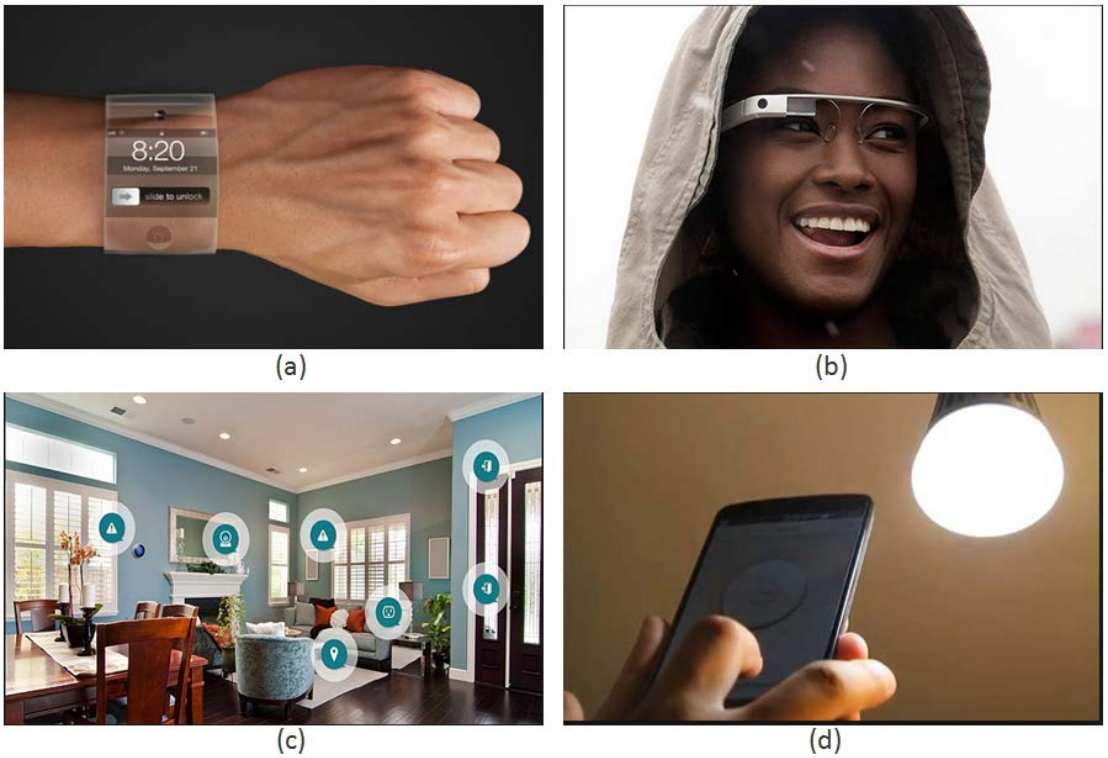

Figure 2. Touchless User Interface Modes

Arguably within years all devices surrounding humans will be more or less smart devices. Multiple versatile devices without a unified way to control them will pose a problem for the typical user. Some devices may have a touch screen (Figure 2a); some may not (Figure 2b). Some

may require complicated controls and settings (Figure 2c); some might only need simple on-and-off (Figure 2d). It is tiresome for humans to control them all one-by-one and to remember each setting. A ***smart*** device should know how to interact with its owner, naturally, with human language -- gesture and speech.

Enabling a unified way to control a multitude of computing devices is the goal of Mediatek's Natural User Interface project—a machine that can communicate with humans using human language, whether a cell phone, TV, tablet, watch, or any computing device. The smart device should at least recognize the user's gesture and voice (Figure 3a). More, it can verify the user by face or voice (Figure 3b). Finally, the device can gradually adapt to user's preference and learn their habits, as shown in the picture below (Figure 3c). Of course, there are other capabilities that machine can learn and acquire to support the user's daily life (Figure 3d).

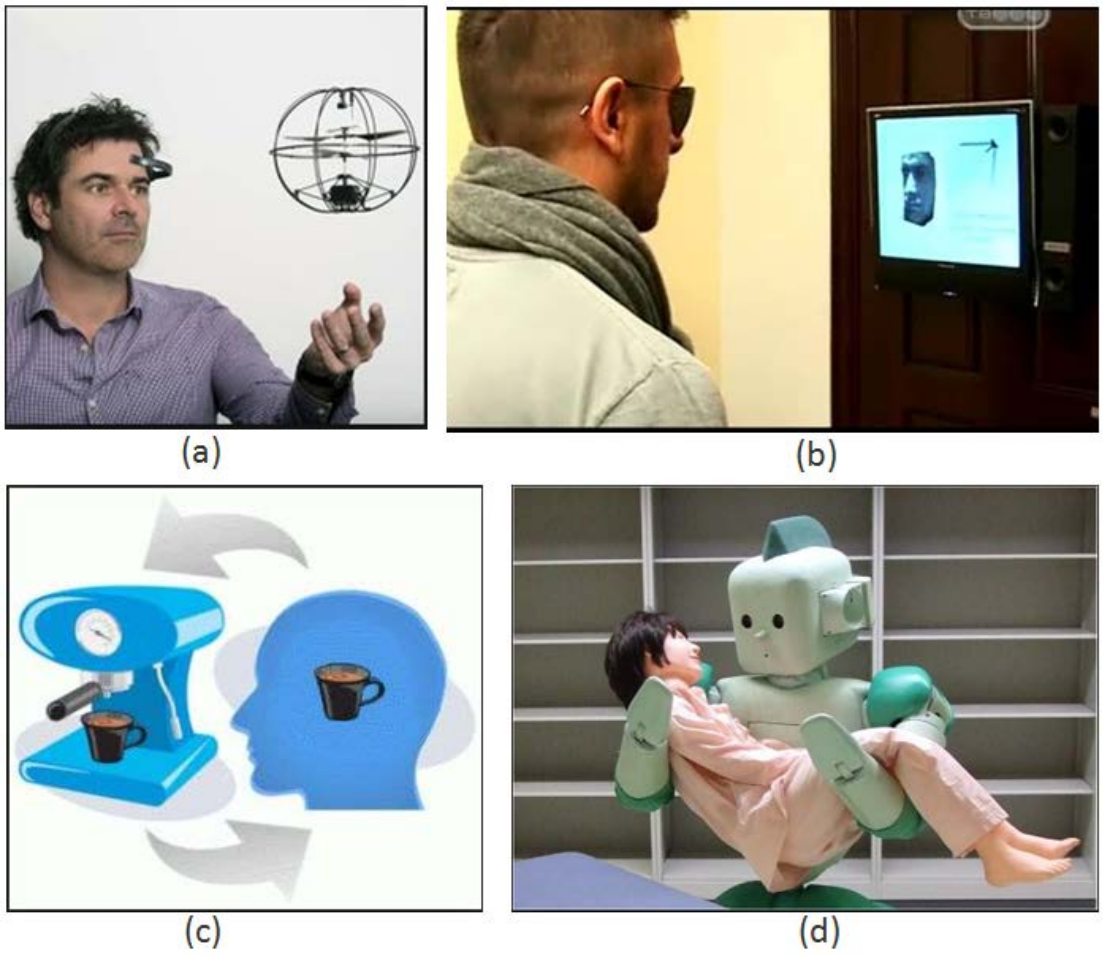

Figure 3. Communication Modes

To realize this target, Mediatek has developed the Natural User Interface software framework shown here in Figure 4.

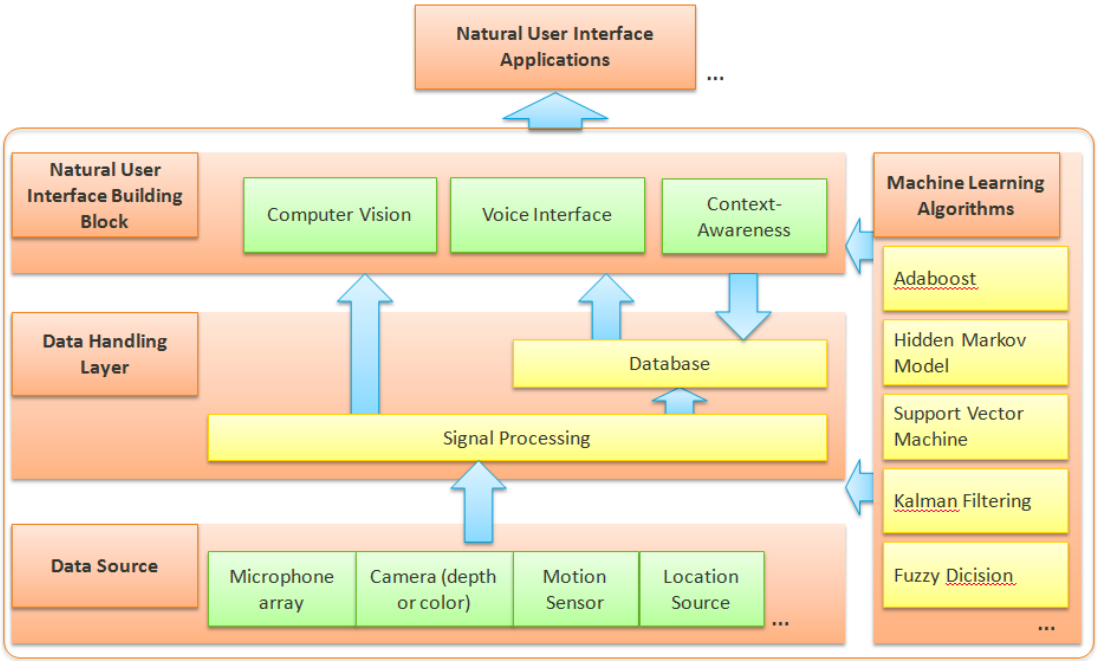

Figure 4. MediaTek Natural User Interface Software Framework

This framework is divided into four sub-layers—Data Source, Data Handling, Natural User Interface (NUI) Building Block, and Machine Learning Algorithms. The **Data Source** layer accepts the NUI required input, such as camera (depth and color), microphone array, motion/light sensor, and location information.

The Data Source sends the data into the **Data Handling** layer where data are time-aligned for coherent audio-visual processing. Within the Data Handling layer is a **Database** for acquiring user statistics, like voice and visual information that can used to learn user habits.

The data is sent to the **NUI Building Block** to perform the pattern recognition and machine learning tasks, such as computer vision, voice interface, and context-aware computing. With the NUI Building Block information the NUI application could be expanded.

Another important block is the **Machine Learning Algorithms** layer where the required pattern recognition algorithms are packaged to be utilities for each layer to use. This design greatly enhances the productivity of new NUI building blocks.

The next section provides an in-depth discussion of the NUI building blocks, and how the combination of this technology can realize the project goals.

2 Computer Vision Technology

MediaTek has developed computer vision technology for both television and smart devices for over a decade. A software/hardware integrated vision-based recognition core has been developed, as shown in Figure 5 below. This vision-based recognition core aggregates a variety of processing units to obtain different recognition functions.

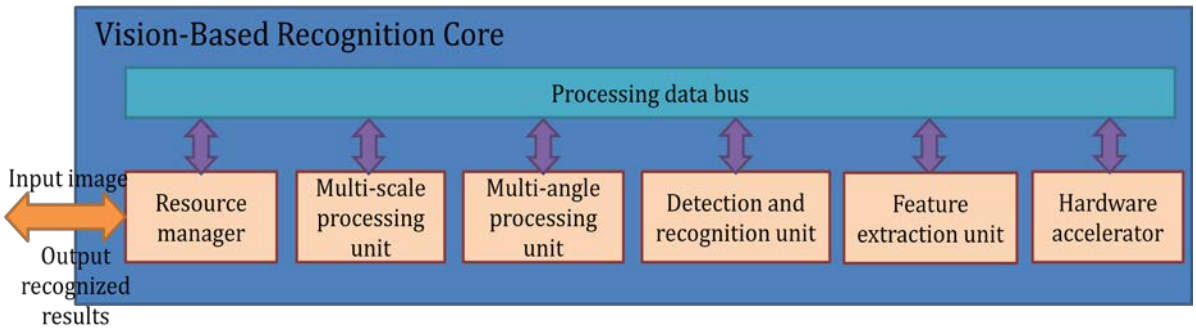

Figure 5. Vision-Based Recognition Core Architecture

The vision-based recognition core enables the vision functions of MediaTek SoC s with low power consumption and high flexibility. The core comprises the following six units: **Feature Extraction**, **Detection and Recognition**, **Multi-Scale Processing**, **Multi-Angle Processing**, **Resource Manager**, and the **Hardware Accelerator**.

- The **Feature Extraction** unit extracts several different kinds of features from captured input.
- The **Detection and Recognition** unit consists of three components: the General Model, the Online Adaptation Algorithm, and the Scene Modeling Algorithm. The General Model receives the object features from the Feature Extraction unit, and provides the system with basic classification and recognition functionality through common algorithms such as AdaBoost, SVM, etc. The Online Adaptation Algorithm then updates the general model with newly learned human characteristics. The Scene Modeling Algorithm gives the General Model more environmental knowledge by scene

understanding and background modeling. These three components are utilized to produce machine learning results. Due to different viewing angles and distances of the camera that captures the image for recognition, the viewing angles and sizes of objects in the natural scenes vary a lot and thus increase the complexity of the algorithms.

- In such cases, the **Multi-Scale** and **Multi-Angle** processing units help to normalize the sizes and capturing angles of objects. These two units reduce the burden of the Detection and Recognition unit to cover a large variety object transformations.
- The **Resource Manager** monitors the availability of the computational cores, hardware accelerators, and the power budget The RM determines whether to use the hardware accelerator with a core adaptation algorithm, as shown in Fig. 6. It provides the recognition core with more flexible computation and low power consumption. Some algorithms create a bottleneck in the overall system are implemented in the Hardware Accelerator on this recognition core. The Hardware Accelerator can be enabled by the Resource Manager to reduce power consumption by largely parallelizing the learning and detection processes.

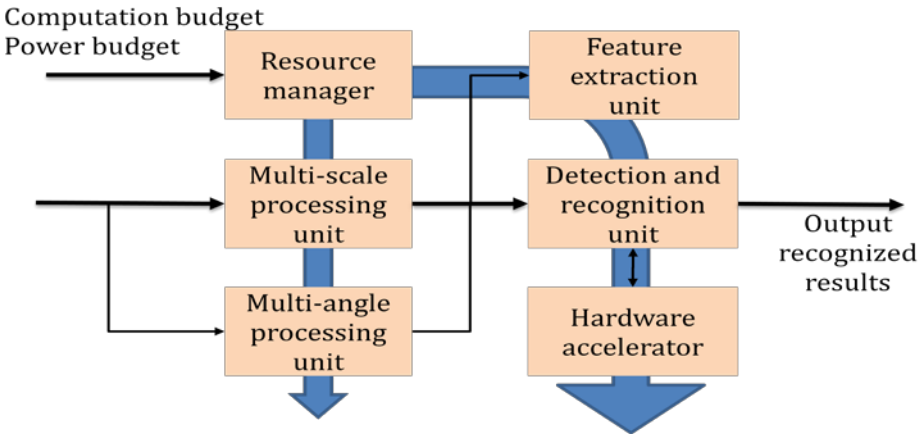

Figure 6. Computation and Power Budget Monitored and Dispatch by Resource Manager

Utilizing these units, the Vision-Based Recognition core is capable of recognizing hand gestures and objects. The purpose of hand gesture recognition is to help users to communicate with the devices. Hand gesture recognition can be used as a remote control of a television, with touchless panel devices, and several other natural user interfaces. With **Multi-Scale** and **Multi-Angle** processing units, the recognition core is able to support multiple sizes and capturing angles of hand gestures and objects. This also increases the detection rate of these recognition algorithms.

Currently, this Vision-Based recognition core is used in MediaTek products: the gesture user interface on TV and the gesture user interface on mobile devices. The gesture user interface lets the users control the devices in a more natural way without pre-training of user behavior. The power consumption is relatively low compared to the original processor based architectures.

3 Voice Interface Technology

3.1 Overview

Voice is the most efficient means of human communication. Acoustic, semantic, and emotion information can be conveyed in a succinct segment of human voice. As a result, using voice as an interface for human-to-machine communication is always an ultimate goal for the development of computing devices (Figure 7). We demonstrate in the sections below status and features that Mediatek has developed to facilitate the human-to-machine communication.

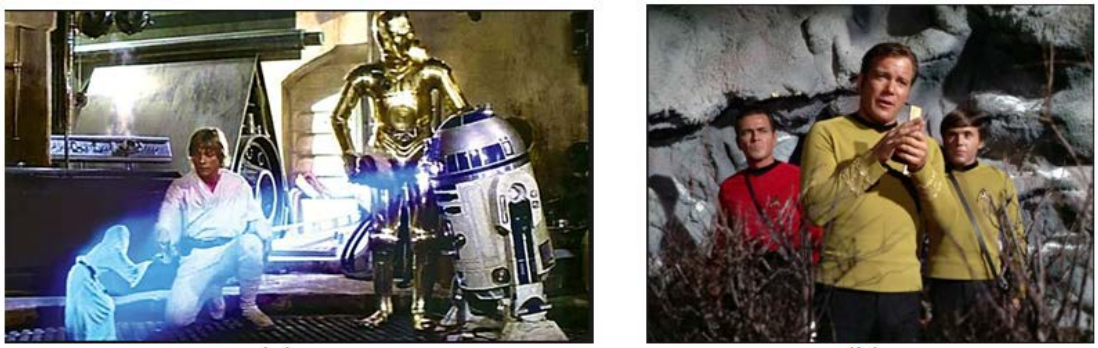

Figure 7. Human-to-Machine Communication

3.2 Voice Keyword Control

Voice commanding is one of the oldest research topics in voice recognition research and has achieved very high accuracy in most usage scenarios. Traditionally, voice commanding requires the user to press a button to prompt the device to listen to the command. If user's hand is available to press a button, he can use hand instead of voice to control a device.

The Voice Keyword Control feature is always listening to the voice keywords. Voice Keyword Control eliminates the necessity of the finger to control a device. And the device will perform the designated action when it recognizes a voice prompt. This is especially useful when finger use is not convenient. For example, in Figure 8(a), when using large screen cell phone, it is

difficult to hold the phone and press the shutter: the user can take self-shot by simply saying “cheese”. Another example in Figure 8(b), people who like to reading during eating can use keyword control to flip the page without dirtying the touch screen.

Figure 8. Voice Control instead of the Finger

3.3 Voice Verification and Wakeup

In some situations, voice verification becomes the only suitable control method, as seen in Figure 9.

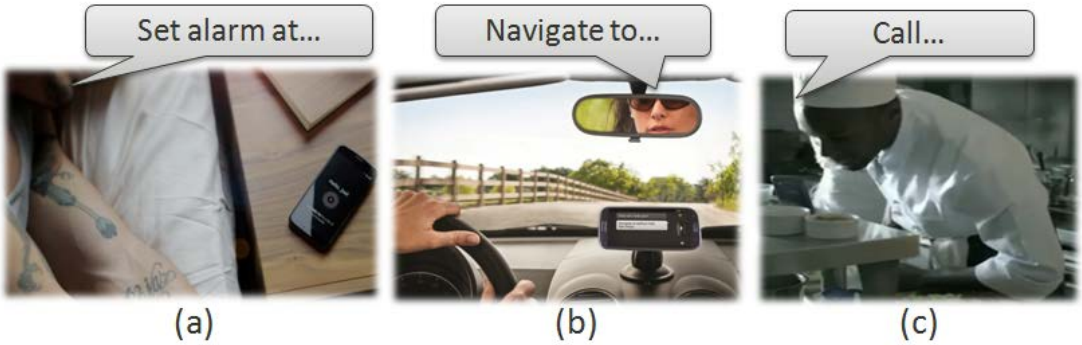

Figure 9. Voice Verification Solution

Two designs are necessary to ensure a satisfying user experience: **voice verification** and **voice wakeup**. In the first situation, we’d like our device only responds to its owner, instead of other users, so voice verification is necessary . In the second type, because hand control of the device is inconvenient—the user doesn’t have the chance to press the button to wake-up the device, the device has to be woken up by voice.

3.4 Voice Search for Contact Information

Voice search is yet another very useful feature. In addition to voice search in a cloud service with an internet connection, Mediatek has developed the **off-line voice search** of device-based information. This is an excellent feature to search contact information, music, or APPs, thereby saving time the user would have to spend retrieving such information via touch screen as in Figure 10.

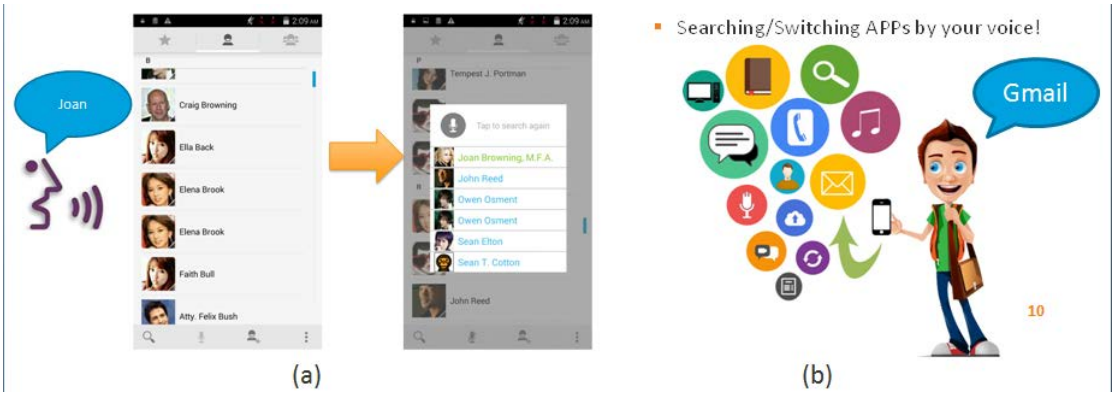

Figure 10. Off-Line Voice Search

This design offers several advantages. First, without the Internet connection requirement, the search is never out-of-service; off-line search can always respond to user in predictive time latency. Second, the search target is a limited set, even if the set has thousands of entries; we can have very effective algorithms to boost the accuracy. Third, the search is single-domain, which makes it very suitable to integrate into the original APP interface, already familiar to the user.

3.5 Robustness Design to Ensure High Accuracy

Users always desire high accuracy voice interface. Many have frustrating experiences when using voice interface, which impedes the popularity of voice interface.

However, because of the advancement of hardware and algorithm capabilities, the accuracy of voice interface is gradually raised to an acceptable level. Currently fewer people complain about the accuracy; indeed, more users are experiencing the benefit of a robust voice interface.

Figure 11. Robust Voice Interface

One key advancement of mobile phone voice interface accuracy is the popularity of the dual microphone. With the help of the dual microphone, we can create a narrow acoustic beam which is directed to the target user, like Figure 11(a), and can exclude the interference of ambient noise or voices. This is achieved by very effective dual microphone interference rejection algorithm developed in Mediatek, like Figure 11(b), which properly use the signal picked up by main microphone and reference microphone to reject the interference signal. When we can successfully reject interference signal and leave only target user speech, the accuracy of voice recognition boosts considerably.

3.6 Hardware Deep Integration and Acceleration

Mediatek's Natural User Interface project is deeply integrated into Mediatek's chip for best user experience; it's a vertical integration from algorithm to hardware, a special advantage of this project.

In Figure 12 below, we've shown that with deep integration to Mediatek's "true octa-core" processor, MT6592, our algorithms response time is comparable to that of the competitor's flag-ship chip. As a result, we can always deliver state-of-the-art performance on algorithm capability and user experience.

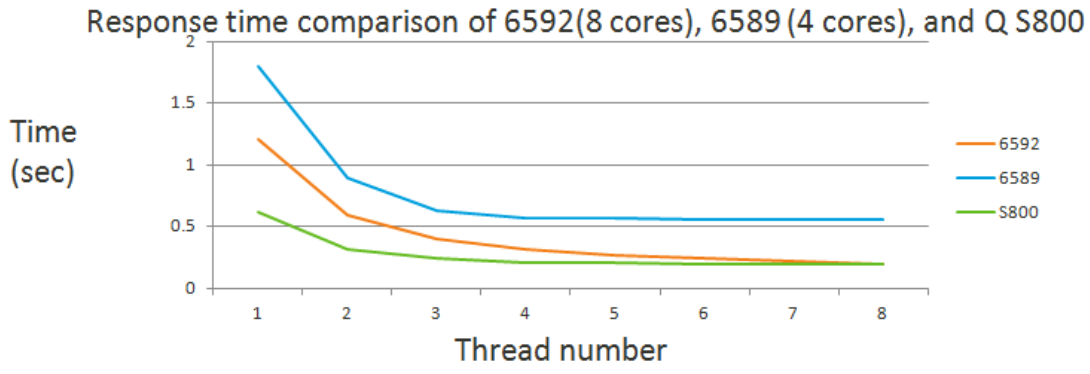

Figure 12. Response time Comparison

3.7 Voice Interface Extension SDK

Finally, the packages described above are released to our customer packed into a SDK called the **Voice Interface Extension SDK**. With the Voice Interface Extension SDK, phone makers and APP designers can create their own innovative voice interface applications without worrying about developing a voice interface engine on their own, which requires both considerable professional expertise and R&D investment.

Meanwhile, for our customers who would like to support multi-language for internationalization purposes, Mediatek's Voice Interface Extension SDK allows model-plug-in. Our customers can plug a multi-language voice model into the framework to support mult-language instantly, without worrying if any changes should be made in the APP's API layer. This design flexibility greatly enhances the productivity to develop voice interface-enabled application.

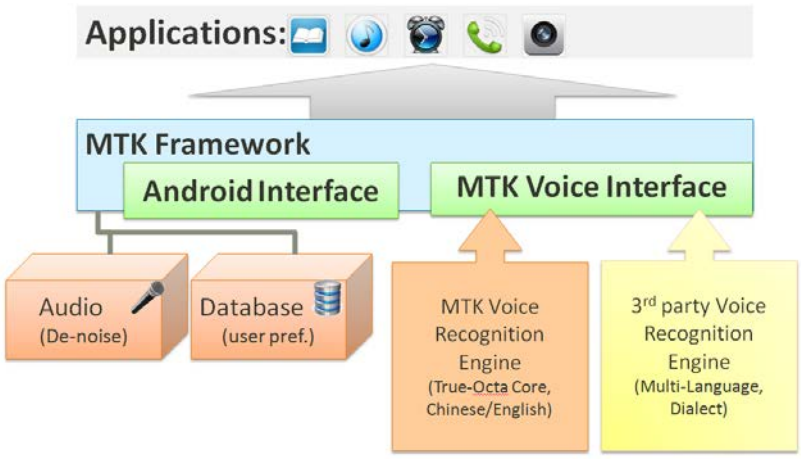

Figure 13. Voice Interface Extension SDK

4 The Fusion of Eye, Ear, and Context

4.1 Overview

Humans are used to communicating with humans. Therefore, the most natural user interface is human interface, which comprises seeing, hearing, speaking, touching, tasting and smelling. Computer science fields today include computer vision, voice recognition, text-to-speech, and motion sensor recognition.

Human communication is of higher quality if using multiple human interfaces. For example, face-to-face communication, in which non-verbal information can be transmitted, always has much better communication quality than a tele-conference. Eye contact, head or hand shaking, and gestures all add quality to the verbal communication.

Human-to-machine communication is also the same. A computing device that can hear, see, and sense the motion will understand more of the human's intent. We are thereby able to deliver higher accuracy or more delicately control the machine.

4.2 The Fusion of Eye and Ear

Computer vision enables the computer the ability to see, voice recognition gives computer the ability to hear. Then, it is intriguing to consider what if a computer can see and hear simultaneously.

Let's take biometrical user verification as the first example. The security level of the commonly used 4-digit password verification system is 99.99%, which means that there is only 0.01% chance of being hacked given your 4-digit is not the commonly used one like "1234". However, to remember an uncommon 4-digit is difficult, researchers consider biometrical verification as an alternative, which verification is done through the biometrical characteristic ex like voice or face. The state-of-the-art of face recognition and voice verification algorithm both gives 1~5% equal-error-rate(EER), when the possibility of false-accept and false-reject are tuned to equal, which is far below the 4-digit verification system. It's obvious that either voice or face itself cannot build a user verification system that is secure enough. However, it is intuitively to combine face and voice recognition together to achieve the security level that is comparable to the 4-digit verification system, and develop a much more natural user interface to the user.

The second example is device control. Voice control is a convenient feature for the user. However, this technology is error-prone due to the environment noise. As a result, even though users wish to utilize voice control features, the technology has not matured enough to achieve real popularity. Computer vision gives great support to voice control by identifying lib-moving or even performing lip-reading. Accuracy could boost significantly with the combining of voice recognition and computer vision technology.

In today's market, microphones and cameras are both cost-effective enough to be equipped in every entry-level computing device. Computing power and storage are both so cheap that is available everywhere. Therefore, without any hardware limitation, the above usage scenario would not be restricted to high-end computing device, but rather will be quite common in various kind of computing device, even to your door lock.

4.3 The Inclusion of Context-Awareness

The same control language has different meanings within different contexts. Take “check out” as an example. It is the process you need to do before leaving a hotel; it means to examine if anything wrong when you driving a car; it asks the phone to collect the interesting stories when you surfing the web. The same condition also happens with human body gestures. An eye-blinking may transmit different meaning under different circumstances.

Context, location, and motion information enrich the meaning of a control. Therefore, context-awareness is the key to give a truly satisfying user experience. In our architecture (see Figure 4 above), we can use location information to identify the user’s intent more clearly, like the example in previous paragraph. Furthermore, we can use motion-sensor information to improve the performance of computer vision or voice recognition performance.

5 Conclusion

MTK's natural user interface combines technologies ranging from computer vision, voice recognition, and context-aware computing to deliver an user experience that approaches human-to-human communication, which is also multi-modal. These achievements can be applied to any kind of computing devices, ranging from handheld, wearable, or devices in the so-called internet-of-things.

In the future there many kinds of computing devices will come to the market to assist human life; each of them may have its most natural user interface to fit its application. With MTK's achievements in natural user interface, there will be plenty of choices and algorithms to satisfy this need.

In the future, we'd like to see humans communicate with machines as naturally as we're communicating with people or our best friend. (Figure 14)

Figure 14. Communicating Future