Speech Recognition Chip
LD3320
Advanced Cheats

Update 2011.10.13

Speech Recognition Chip/Sound Control Chip
Single-chip/non-specific/dynamic edit identification list
Speech Recognition Solution Use voice to communicate VUI

(Voice User Interface)

translated by JimWoo (3dfilm@mail.ru)

ICRoute 用声音去沟通
VUI (Voice User Interface)

Web : www.icroute.com
Tel : 021-68546025
Mail: info@icroute.com

Introduction:

Based on the speech recognition chip LD3320 development products, refer to "LD3320 Development Manual." In order to improve the subjective experience of the end user for speech recognition, this article summarizes some of the high-end methods and tips, and assembles articles to share with everyone.

This document will be updated from time to time to gather experience in real time. Stay tuned.

1. Use "trigger recognition" mode in scenes with high recognition accuracy

About the LD3320's two modes of use, you can refer to the website: http://www.icroute.com/web_cn/LD332X_UserModel.html.
In scenes with high recognition accuracy, the "trigger recognition" mode should be used. Because:

1) When the user presses the hot key each time, the spirit is in the most concentrated state. At this time, the voice command spoken by the user will be more serious and clear. The recognition error caused by the overly casual pronunciation of the user is avoided.
2) Each time the hot key is pressed, the product should be given a clear start signal, such as a “due” sound or other prompt signal, which can give the user a clear start prompt, so that the user can grasp the time of speaking the voice command.
3) After the key is triggered, the user will be close to the microphone and speak the voice command to avoid misidentification caused by other environmental sounds being recorded into the LD3320 chip.

Another: This method is still a way to save power, in the absence of identification, completely do not let the chip work to save power.

2. Add "garbage keywords" - absorption error recognition

After setting the key words to be identified, in order to further reduce the false recognition rate, some additional vocabulary words may be added to the identification list to absorb the wrong recognition, thereby reducing the false recognition rate.

These keywords can be called "junk words."

For example, in an application scenario, there are four key words that need to be identified, "Forward," "Back," "Open," and "Close." After setting these 4 keywords into LD3320, you can set another 10~30 words into LD3320, such as "front door", "back door", "aaa", "呜呜" and so on.

Only when the recognition result is within 4 keywords, the recognition is considered valid. If the recognition result is “junk words”, it means that other voices caused misrecognition, and the product should start the recognition process again.

In this way, the false recognition rate can be reduced very very effectively. Greatly improve the subjective user experience.

The selection of “junk words and phrases” may be best performed by selecting words with the same number of words as the keywords to absorb possible false recognitions.

It should be noted that this method can be applied in "trigger recognition" mode or "circle recognition" mode.

The principle of this is as follows:

The non-specific human speech recognition technology ASR is a matching recognition technology based on keyword lists. The essence of the algorithm is to find the most similar word as the recognition result in the keyword list after extracting the characteristics of the input speech. (http://www.icroute.com/web_en/LD332X_principle.html)

Therefore, any sound input into the speech recognition chip will be matched against the words in the keyword list and will be scored in turn. In this way, other people chatting casually, or arbitrarily speaking a command that is not in the keyword list, or other unconnected speaking voices, may be matched to a certain keyword and output as a result. This leads to misidentification.

Although there are certain algorithms in the algorithm design to avoid such misidentification, they cannot be completely avoided. Product developers can target the outside of the chip to reduce the false recognition rate. The method provided in this section is a very effective method and has a very important position in practical applications.

3. Password trigger mode

In some applications, it is desirable to have high recognition accuracy, but it is impossible to require the user to “trigger” each time by pressing a key. At this point, "password trigger mode" can be used.

The product defines a phrase as a trigger password. For example, you can define "Open Sesame" as the trigger password.

When the product waits for a user trigger, it starts a “loop recognition” mode, sets the trigger password “Open sesame” and other dozens of words used to absorb errors into the LD3320. Only when it is detected that the identified result is the triggering password, it is considered that the terminal user has called this password. At this point, the tone is given, and a "trigger recognition mode" is activated, and the corresponding identification list is set into the LD3320, prompting the user to speak the operation to be performed within a few seconds after the tone.

When waiting for the user's process, if the recognized results are those that are used to absorb the errors, it is considered to be misrecognition, or other sound interference, without any processing, and directly enter the "circular identification" mode again.

This password triggering mode combines the advantages of the other two modes and combines the

The "garbage keywords" method can provide more convenient and practical voice operation features for products.

4. Clever use of key words ID

When setting keyword words into LD3320, the pinyin string of keyword words is passed to LD3320, and an ID is also passed in to represent the keyword.

The recognition result of LD3320 also outputs the ID of the identified keyword as a result.

In the LD3320 chip, different keyword words can correspond to the same ID. And the ID does not need to be continuous. This provides product developers with convenient programming methods.

For example, "Beijing" and "Capital" can be set to the same ID for follow-up processing.

For example, when using the "junk words" mentioned in the second section, the added IDs used to absorb the wrong keywords may be marked with a value, or they may be marked as a special ID value, such as greater than 200. In the program is relatively simple, it is easy to deal with misrecognition, to avoid adding a lot of keywords, write procedures need to add too many program branches for the processing of these keywords.

5. Working voltage

The LD3320 has three power inputs.
VDD Digital Circuit Power Input 3.0 V – 3.3 V
Power Input for VDDIO Digital I/O Circuit 1.65 V – VDD
VDDA Power Supply Input for Analog Circuitry 3.0 V – 4.0 V

However, in the actual design, a three-way power supply can be used with a uniform operating voltage of 3.3v. The minimum operating voltage is 3.0v. When the input voltage is lower than this value, the chip will not start working. This simplifies the circuit design. If conditions allow, you can isolate the analog power supply from the digital power supply to avoid interference and achieve the best power management.

6. Mark foreign languages or dialects in pinyin

Speech recognition identifies "voice". For non-specific human speech recognition, the keyword words to be identified are marked with phonetic symbols when the keyword words are described.

For the current Chinese recognition supported by the LD3320, Pinyin is used to describe keywords.

In other words, as long as Pinyin can spell pronunciation, it is possible to input chips and identify them.

Therefore, when you need to identify some simple foreign languages or pure dialects in certain situations, you can use pinyin annotation.

For example, some occasions need to identify some simple English words, you can use pinyin to mark: one - → wan

Two -→ tu
Three -→ si rui
For example, some occasions need to identify some pure dialect vocabulary words, you can also use pinyin to mark: Shanghai dialect "late" pronunciation is "ya", then "evening" word is marked in Mandarin "wan bao" if you want to label To pronounce in Shanghai dialect, it is "ya bao", so that the Shanghai Evening News can be identified.
It is worth noting that the LD3320 supports Mandarin Chinese, and some foreign languages or dialects cannot be described in pinyin, so the LD3320 may not be able to complete all required foreign language or dialect tasks.

7. Set multiple custom pronunciations for the same keyword ID

End users may have different pronunciation habits for the same vocabulary when speaking voice commands.

For example, "Turn on the light", the user may say "turn on the light", "turn on the light", "turn on the light", "turn on the light" and so on.

Taking full advantage of the LD3320's 50 dynamically editable key recognition entry features, developers can set these habitual utterances into the chip so that no matter how the user says it, it will be correctly identified, further increasing the end user's good experience.

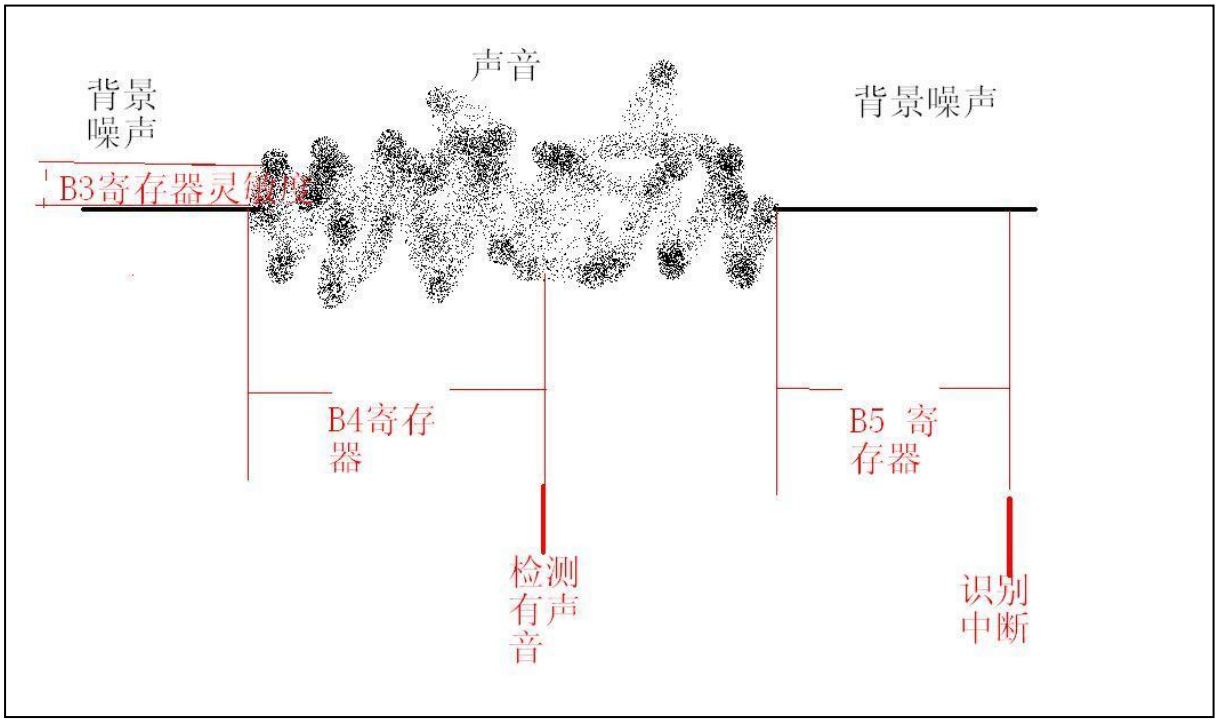
At the same time, it is possible to handle these multiple customary pronunciations conveniently in programming, in conjunction with the fourth article of Cheats, "Careful use of Keyword IDs."

It is worth noting that: if used to control the work, you need to add some spam keywords to absorb errors to reduce the misrecognition rate. See Section 2 "Adding Spam Keywords" - Absorbing Misidentification.

8. Adjust the response time after the end of the voice recognition results

The LD3320 chip internally uses the VAD (Endpoint Detection) mechanism to determine if the person has finished speaking and gives recognition results. For a detailed description of VAD and the mechanism for obtaining recognition results, please read the webpage introduction: http://www.icroute.com/web_en/LD332X_principle.html According to the VAD mechanism, after the speech recognition chip detects a continuous background noise, it is assumed that the user has finished speaking the speech recognition command, and then gives the recognition result. The default setting is to detect the recognition result only if there is continuous silence for 600 milliseconds after the start of the vocalization. There are three VAD related registers: B3, B4, and B5. According to the “LD3320 Development Manual.pdf”, the B3 register is equivalent to the sensitivity. The B3 and B4 registers together determine whether the sound starts. The B5 register is used to determine whether the sound is over. The recognition result will be given after the sound is over (interruption)

The working diagram of VAD is as follows:



That is to say, according to the default setting, from the end of human speech, to the voice recognition chip initiative to send the result to be interrupted, there must be an interval of at least 600 milliseconds. If the user wishes to adjust this reaction interval, the following aspects can be started:

1. Change the way of use

In a manner similar to that of a walkie-talkie, each time a person presses a key, he or she presses and begins to speak a command. After the command is completed, the key is released. Each time a key is released, the master control MCU sets the BC register. Get recognition results immediately. (The BC register is described in "LD3320 Development Manual.pdf")

2. Modify register B5

register judged by VAD

ASR: Vad Silence End After the voice data segment is detected, the background noise segment is detected again. The background noise segment that is detected continuously for a long time can be confirmed as the end of the actual voice segment. Every 1 unit, 10 milliseconds. Default: 60, equivalent to 600 milliseconds Value range: 20~200 (equivalent to 200~2000 milliseconds)

However, this modification will result in that, if the time is too short, causing the user to pause when speaking can also cause the VAD to detect that the end of the speech, thereby reducing the recognition rate of certain users.

3. Modify the microphone volume register

Modify the volume of the microphone, 35 registers, (adjust the adjustment range between 40H and 58H) to see which recording gain is suitable for the microphone and the environment.

4. Modify the B8 register

For example, to change to 2, then this means that, in any case, within two seconds after the start of each recognition, it must stop identifying and give a recognition result. (This setting does not affect VAD detection).

If the value of b8 is extremely small, for example setting 1,2,3, it needs to be given to the user before starting identification.

A very clear hint to start identifying. In case the user is not ready yet, the identification time passes.

However, this interval setting is too short, it will inevitably cause some of the possible misidentification, such as the voice command is relatively long, then this time is set too small, it will cause a relatively long voice command can not be completed in a specific time Misidentification.

Therefore, when this value is set to a relatively small value, it is generally recommended to use the "trigger recognition" user interface to avoid the use of "loop recognition" user interface.

5. Use environment

Changing the use environment may cause noise or echoes in certain environments to affect the judgment of the end of speech.

As well as the speaker's own volume, if the voice is low, it can also make it difficult to judge whether the person is speaking or not.

Change the content of the command words, compare the readings with good voices, etc., so that the user can easily and clearly pronounce the voice commands.

9. Microphone, related register setting and recognition effect and distance

The effect of speech recognition is the result of a subjective experience. It is related to the following factors:

1. The sound of the surrounding environment
2. Contents of the identification list: whether it is loud sounding sounds or closed sounds that are not easy to pronounce
3. Identify the degree of difference between the various words in the list
4. The speaker's articulation / size / speed / severity / accent
5. User Operation Flow Settings
6. Physical characteristics of an external microphone
7. Does the speaker release the volume and so on.

The quality of speech recognition and the effective distance between it and the microphone/microphone are very large. The quality of the microphone/microphone determines what kind of sound quality is sent to the recognition chip, so it also determines the distance of the recognition effect. In general, the distance of the microphone action is about 1 meter, and the carefully selected microphone action distance is about 2 to 3 meters.

Depending on the craftsmanship and quality of the production, the background/noise level of the microphone/mic head may cause the recording distance of different microphone/mic heads to be different. For example, the microphone head, which is also known as the sensitivity of 39db, is bought from different places in the market. The effect is completely different.

With poor quality, the electrical noise of the recording is very high, resulting in the annihilation of people’s voices and the need for people to increase the volume or to get closer.

The quality is good, the frequency response curve of the recording is relatively flat, the electrical noise is low, and the relatively distant vocal voice can be recorded relatively clearly.

There are many microphones on the market, which are produced for mobile phones. Their feature is close-range recording. This application on mobile phones can suppress distant noise. However, in speech recognition applications, such proximity microphones will result in near-identical distances.

There are also some microphone heads, the electrical noise generated by itself is very strong, which will cause the recognition effect to decline.

Developers should experiment with more microphones/microphones to choose the right one for their product. According to the application environment and positioning of the product, an appropriate microphone must be selected to fully utilize the recognition function of the LD3320 identification chip.

It is worth noting that for microphones or similar devices that have been augmented with amplifiers to increase the recognition distance, some amplifiers will destroy the waveform of the sound, causing “overshoot” of the sound input to cause serious distortion of the sound, which will greatly affect the recognition effect, resulting in The recognition rate has dropped dramatically.

It is worth noting 2: for some modules with noise reduction chips, some digital noise reduction chips will force some relatively small sounds to be zero, resulting in the loss of sounds (such as the consonant "si" that sounds lighter, "wu "etc." will also greatly affect the recognition effect and lead to a serious drop in the recognition rate.

When adjusting the register associated with the microphone AD input, there are some suggestions that can help increase the recognition distance and improve the recognition quality:

1. MIC gain register 0x35

The mic gain register (0x35) of the LD3320 chip is not set as good as possible, for 35 registers:

Normally in a normal room or quiet outside, the recommended range is 0x40 ~ 0x53. A higher value will result in excessive sampling of voice samples and a sharp drop in recognition rate. The reference program gives 0x43. If you need to increase the recording volume, it is recommended to 0x4c, it should be more appropriate. In this range, the distance from the user's mouth to the microphone should generally be more than 0.5 meters, and it is easy to produce excessive overshoot.

If it is in a very noisy environment, such as an exhibition, for example, nearby high-powered speakers are playing sounds, such as in the car or motorcycle, where there is a strong wind noise environment, in these environments, long-distance accurate identification is almost Impossible (unless the developer has added an active noise reduction chip at the front end). The 0x35 register should be set between 0x10~0x2f. At this point, the MIC of the LD will not be used for gain, which can effectively avoid over-recording. However, at this time, the distance between the user's mouth and the microphone is required to be within 0 to 50 centimeters, and the speaking volume is increased to obtain a better recognition effect.

It must be noted that the recording characteristics of different MICs and microphone heads are completely different. And the length of the MIC to LD chip connection will affect the MIC's recording effect and recording volume. Therefore, the developer must carefully experiment with each time after replacing the development board or changing the MIC to tune the register parameters to obtain the best results.

There is no universal register parameter that can be adapted to all application environments. The developer must set a most suitable parameter in conjunction with the occasion to use his own product. This parameter must be developed by the developer on his own board, using his own microphone microphones purchased, in the actual product application environment, to perform repeated experiments to debug.

2. B3 register

Register 0xB3, which can be roughly understood as sensitivity, is the sensitivity to the intensity of the surrounding sound. The default is 0x12H.

When it is necessary to increase the sensitivity and increase the recognition distance, adjust to 0x0FH or 0x0AH to achieve the effect. (The most sensitive is 0x1)

However, this adjustment is a double-edged sword. Sensitive to the sound, it will inevitably bring negative effects on the recognition rate. So the adjustment for the sound is very much in balance. Need to adjust slowly with the product.

This B3 register can be set to 0x0FH where 0x35 microphone volume is set.

3 VAD can be turned off in key trigger mode

If the product can use the key trigger mode, after the key is pressed, the recording is performed for a period of time (for example, 3 seconds or 5 seconds), and then the recognition result is obtained after the period of time is over.

Then there are two options, one is to keep the VAD turned on. This will still detect whether the person's speech starts. After the end of the person's speech is detected, the interruption and recognition results will be given. The other is to completely shut down the VAD. All the sound data of this time (for example, 3 seconds) are recognized. Only after this period of time will the interrupt and recognition result be given.

This key triggers and closes the VAD mode. The register is set as follows: 0xB3=0 0xB8=2 (length of time in seconds). When the 0xB3 register is set to 0, VAD detection is turned off. At this point, the chip will recognize all incoming sounds without checking and distinguishing between vocal and background noise. The value of B8 set determines the length of time for fixed recording, such as 2 or 3, etc., and is set according to the actual demand of the product.

4. B5 register

The B5 register, (see Section VIII), defaults to 60, which is equivalent to 600 milliseconds. If this is set too short, it will cause some of the consonant sounds that are lighter in the middle of a person's speech to be interpreted as saying that the speech has ended, resulting in the long words being separated, which leads to recognition errors.

5.

In summary, if you want to record a large amount of sound, the recognition function is farther away. In register adjustment, 0x35 registers can be set to 0x4c, 0xB3 registers can be set to 0xf, and B5 B8 defaults. At the same time, choose a suitable microphone with excellent quality, the combination of the two will achieve better recognition results.

Ten, voice recognition user usage patterns

Two different user usage patterns are introduced on the website page http://www.icroute.com/web_cn/LD332X_UserModel.html: trigger recognition mode and loop recognition mode.

The technical documents on the download page: "Voice Control Intelligent Product Voice Interface Design Guide.pdf." (http://www.icroute.com/web_cn/VUI_DesignManual.html) and the previous sections of this document, have also mentioned many times in the Different scenarios require different user modes.

In fact, if you look at the technical point of view of the chip, these different user modes, the workflow of the voice recognition chip LD3320 is the same:

Chip Initialization LD_Init_ASR()—>Add Keyword Identification List LD_AsrAddFixed()—> Turns on the microphone and starts speech recognition LD_AsrRun(). (This process, which is the content of the function RunASR())

The conditions for starting this identification process and ending the identification process are different and constitute different user modes.

The conditions for ending the identification process are:

1) When VAD is turned on, VAD detects the end of sound (there is a period of continuous silence), then the identification process ends and an interruption is given.
2) The time set in the B8 register is reached. If the identification operation is still performed in the identification flow, the identification flow ends and an interrupt is given.
3) Write 07H or 08H to the BC register to force the end of the identification process and give an interrupt.

These three conditions are independent of each other. Any one of the conditions is reached first, and the identification process will end. For a detailed description of the VAD registers, please read the "LD332X Development Handbook.pdf" http://www.icroute.com/web_en/Download.html#LD332X-Manual ; and the related discussion in Sections 8.9 of this document.

The conditions for starting the identification process are controlled by the master MCU.

Combining various conditions for starting and ending the identification process creates a variety of different user usage patterns. Developers can select the most appropriate one based on the characteristics of their products. These user usage patterns are as follows:

Cycle identification:

- Turn on VAD functionality (default).
- Setting the B8 register to a suitable value (default): It is generally recommended that this time is relatively long, so that during the loop recognition process, the user's speech is just split by the two recognition processes.
- Identification of process start conditions: Once the identification is interrupted, the recognition result is read immediately. After the processing is completed, the next identification process is started immediately.
- Identification process end condition: end by VAD or end by B8 condition.

Trigger recognition + single button + VAD:

- Turn on VAD functionality (default).
- Set the B8 register to an appropriate value: The time is set by the user according to the characteristics of the product. The meaning here is how long the key waits for a maximum time. After this time, the user's voice command will no longer be received.
- Identification process start condition: When the master MCU receives a button, it starts an identification process.
- Identification process end condition: end by VAD or end by B8 condition.

Trigger recognition + single button + VAD + avoid zero results:

This mode is an improvement over the previous mode (trigger recognition + single button + VAD). Since the sensitivity of different microphones is completely different and the ambient noise is not the same, even if people do not speak, there may be other reasons (other The sound entering the microphone causes the VAD to detect the sound to start) causing the VAD to start and end the recognition process. At this point, the recognition result is generally zero. At this time, the user often has not started to speak. To avoid this situation, you can make improvements.

- Turn on VAD functionality (default).
- Set the B8 register to an appropriate value: The time is set by the user according to the characteristics of the product. The meaning here is how long the key waits for a maximum time. After this time, the user's voice command will no longer be received.
- Identification process start conditions: When the master MCU receives a key press, it initiates an identification process. When the recognition result is zero, the recognition process is started again immediately. Until a certain identification process gives a recognition result.
- Identification process end condition: end by VAD or end by B8 condition.

Additional explanation: When the keyword list is set to spam words, the developer needs to decide at this time, as long as a non-zero recognition result is given, it ends (if it is a garbage word, the final recognition triggered by this button is recognized). The result is junk words); it is still necessary to identify the result of a non-spam word (it may happen that the user has not spoken or has not been seriously said, resulting in the identification of the valid identification results, but has been identified)

Trigger recognition + single button + VAD + avoid zero result + master MCU timing:

For the previous mode, we can make further improvements by introducing the timing function of the master MCU. The master MCU also starts a timer to clock. Once this time expires, if it is still in the process of identifying the flow, it will immediately set 07H or 08H to the BC register to force the end of recognition. At the same time, the master MCU does not restart the identification process.

- Turn on VAD functionality (default).
- Set the B8 register to an appropriate value: The time is set by the user according to the characteristics of the product. The meaning here is how long the key waits for a maximum time. After this time, the user's voice command will no longer be received.

- Identification process start conditions: When the master MCU receives a key press, it initiates an identification process. When the recognition result is zero, the recognition process is started again immediately. Until a certain identification process gives a recognition result.
- Identification process end condition: end by VAD or end by B8 condition or master MCU timing to active end.

Supplementary explanation: After the master MCU is timed out, I want the BC register to be set to 07H or 08H. It is up to the developer to decide whether or not he thinks that the recognition results obtained at the end of this timing will have an effect.

Trigger Recognition + Single Button + Off VAD:

For the single-button trigger mode, the VAD can also be turned off and the identification process can be ended by only relying on the B8 time setting. This method is the same as that of the voice king on the mobile phone. After the key is pressed, the recognition result must be given after the set time has passed. In this case, all the sounds in this period of time are sent to the recognition chip for calculation. If the user cooperates, the recognition rate should theoretically be better than the VAD, but if the user does not cooperate, it may be worse.

- Turn off VAD functionality.
- Set the B8 register to a suitable value: The time is set by the user according to the product features. It is generally recommended to set it to 3~5 seconds.
- Identification process start conditions: When the master MCU receives a key press, it initiates an identification process.
- Identification process end condition: Ended by condition B8. And each time must be after the completion of B8 set time, will give the recognition results, will not be given in advance. Even if the user finishes the command, the recognition result will not be given if it is less than the time set by B8.

Trigger recognition + double button + off VAD:

The dual-key mode is similar to the walkie-talkie mode. After a button is pressed, a voice command is started. When the voice command is spoken, the button is pressed again to obtain the recognition result. The user is required to press the button to notify the MCU after the command is spoken.

- Turn off VAD functionality.
- Set the B8 register to a suitable value: B8 It is generally recommended that the setting be longer, because this mode is controlled by the user to end the identification process. Therefore, the B8 should avoid excessively time and end the identification process prematurely.
- Identification process start conditions: When the master MCU receives a key press, it initiates an identification process.
- Identification process end condition: The master MCU receives the key again, sets 07H to the BC register, ends the identification process and obtains the recognition result.

Supplementary note: If the user is willing to cooperate, then this model should be the best. Because the user presses the key twice, strictly speaking, only the voice command is sent to the chip, and no background noise is sent to the chip. So the effect is best.

Of course, the developer can also modify the button to start talking after pressing the button, release the button and give the recognition result. This is the development work in the master MCU.

The above is a detailed description of several different modes of voice usage, all of which are implemented by the developer writing code in the main control MCU. The developers need to carefully select the most suitable method in combination with the actual characteristics of their products. At the same time, it is necessary to tell the user that if used correctly, the user is given a clear warning tone or prompting light in the product design to help the user speak the voice command in the right time and get the best recognition experience.