Build and Optimize Cyberon DSpotter Model

Version: 1.3

Date of issue: 2021/1/13

Cyberon Corporation

Software solution provider for embedded system

http://www.cyberon.com.tw/

Contents

Change log:

Version	Date	Update
1.0	2020-08-04	1. First release.
1.1	2020-08-05	1. Modify the item 5 of chapter 2.5. 2. Interchange the item 5 and 6 of chapter 3.3.
1.2	2020-08-25	1. Modify chapter 1.1. 2. Modify the RAM and model size in the item 5 of chapter 1.2. 3. Add more description in the item 2 and 8 of chapter 3.3. 4. Add the description and example of sound junction in the item 2 of chapter 4.1. 5. Add the example of garbage command in the item 4 of chapter 4.2.
1.3	2021-01-13	For English version.

1. Cyberon DSpotter

1.1. Introduction

Cyberon DSpotter is a recognition engine for voice wake-up or predefined commands. It has the following features:

1. Using DNN technology, it has good noise immunity.
2. The demand for computing resources is not high. All fixed point operations are used, which is suitable for running on MCU. It has been optimized using its SIMD DSP instructions on ARM CM4 and RISC-V with P extension platforms, and adding an identification instruction will only use a little more resources. Please refer to 1.2 for detailed specifications.
3. Each language has two sizes of voice models to choose from. The voice model of level 0 is smaller than that of level 1, and uses less resources, but the false trigger is higher and the recognition rate is less at the noise environment (about 3% when SNR is 10 dB and 8% when SNR is 5 dB). The developers can choose flexibly based on system resources.
4. Provides the Windows tool DSMT, which can quickly create a recognition model by typing in text. It supports 30 languages at current time.
5. DSMT need 10 males and 10 females (at least 5 males and 5 females) command recording files, and each person reads each command 5-10 times, then you can use DSMT to optimize the model parameters and test the performance of recognition rate and false alarm rate.
6. If there are command recording files of 50 males and 50 females (at least 30 males and 30 females, the more the more the better), and each person records each command 10 times, we can use our own computing cloud to retrain the speech model. The adapt speech model will make the accent compatibility better, improve the recognition rate of various environments and reduce false trigger. In general, the adapt model is better than level 1 model, but uses resources similar to level 0 model.

1.2. Specification

The detailed specifications of DSpotter are as follows:

1. CPU (optimization option is -O3)
   About 45/60 MCPS for level 0/1 on ARM CM3.
   About 30/39 MCPS for level 0/1 on ARM CM4 (using SIMD DSP instruction).

About 26/34 MCPS for level 0/1 on RISC-V with P extension (using SIMD DSP instruction).

2. Code size
   About 40 KB on ARM CM4, different MCU platforms or compile options will be slightly different.
3. RAM
   The base requirement is about 37/40 KB for level 0/1.
4. Model size
   The base requirement is about 91/155 KB for level 0/1. Different languages will be slightly different.
5. Take the command word of 4 syllables as an example, adding a command will increase
   Level 0: 0.05 MCPS (on CM4), RAM 128 bytes, model 32 bytes.
   Level 1: 0.064 MCPS (on CM4), RAM 128 bytes, model 32 bytes.
   PS. A recognition command may company with some garbage commands.

2. Cyberon DSpotter Modeling Tool

2.1. Introduction

Cyberon DSpotter Modeling Tool is abbreviated as DSMT. It is a tool used to build, test, optimize and adjust voice models. It has the following functions:

1. An account and internet connection are required to use it. Only Windows platform is supported.
2. After inputting the command text, it will be sent to the cloud server. After the server has established the command model, it will be sent back to DSMT, which can be tested and verified immediately. Currently, 30 languages are supported, and the same voice model can only be used in one language, but it also supports Chinese-English, Korean-English, Japanese-English bilingual models.
3. The server has a built-in dictionary, which will find out the pronunciation and its multiple pronunciations from the input text. DSMT can be connected to the cloud TTS (Text To Speech) to play the pronunciation. If the pronunciation is incorrect, the user can modify it or add multiple pronunciations.
4. With online testing tools, you can record from the PC microphone. After the voice model is built, you can test it yourself. After testing, you can archive the recording and use it as the corpus for offline testing.
5. With offline test tools, read wave/script files for test, and allow users to mix various types of noise and set Signal Noise Ratio (SNR) to test. After test, the recognition rate and false accept (FA) will be counted. The score, energy, and time point recognized are provided as the basis for subsequent adjustment and optimization. It can be used for identification rate test and false trigger test.

DSMT include a detail user manual, please refer it to operate these functions.

2.2. Preparation for Noise Test

DSMT has built-in car, babble and office noise sound files, and you can also use external input sound files, for example:

1. Babble noise
   Please download Babble_X.wav from:
   https://github.com/microsoft/MS-SNSD/tree/master/noise_test
2. Pink/White/Brown noise
   You can use Audacity to generate 120 second noise wave files.

Audacity is a free and open source audio editing software (https://www.audacityteam.org/)。

3. Music noise
   Alexa Voice Service (AVS) is tested with music_xxx.wav in Acoustic_Noise_and_Calibration_Files.zip, which can be downloaded at: https://developer.amazon.com/alexa/console/avs/preview/resources/details/AVS%20acoustic%20audio%20files

2.3. Preparation for False Trigger Test

Each language needs to prepare its own false trigger test corpus, which can record the sound of news, variety shows, dramas and other programs on TV (or Youtube) for a total of 24 hours.

Amazon AVS provides English, German, French, Italian, Spanish, and Japanese false trigger test corpora:
https://developer.amazon.com/alexa/console/avs/preview/resources/details/AVS%20acoustic%20audio%20files

It is a 24-hour mp3, please convert it to 12 wav files first, and prepare the script(*.spt) file. The encoding of script file shall be Unicode 16 little endian (UTF16LE), and the format of each line is as follows:
full_path_file_name \r\n, for example D:\WAVE_OOV\CHT1.wav

2.4. Test Criteria

The recognition results have the following situations:

1. Recognition success: the number of successes divided by the number of tests can get the recognition rate.
2. False Reject (FR): There were instructions but no results were recognized.
3. False Accept (FA):
   A. The recognized command is different from what was said, such as "Turn on the light" => "Turn off the light".
   B. There is no instruction but the result is recognized by other sounds (False Trigger or False Alarm).

Customers can set the recognition rate and false trigger standard of test acceptance

according to their needs. The following provides various standards set by Amazon for Alexa as a reference:
https://developer.amazon.com/alexa/console/avs/preview/resources/details/AVS%20Acoustic%20Self-test%20Checklist

1. Recognition rate
   Quiet environment: 90%
   Stationary noise environment (Ex. Car or pink noise): 80%
   Music noise environment: 80%
   When the test device plays music (AVS is to play the song Happy-Pharrell Williams HD to test)：67%
2. False alarm
   No more than 3 times in 24 hours.

For devices used at short and medium distances (within 3 meters), such as earphones, smart watches, and bracelets, DSpotter should pass this standard.

For devices that may play music (such as smart speakers) and need to be used at a far distance, the SNR of the noise test will be less than 5 dB, and even the actual use environment may have severe reverberation, so a microphone array and signal pre-processing system are required to pass the test. For example, Amazon Echo and Google Home use 7 or 2 microphone arrays respectively. When used at a long distance, the Echo's measured performance is obviously better.

2.5. Optimal Adjustment of Parameters

In DSMT, the parameters that can be optimized and adjusted are described as follows:

1. After receiving the voice command, DSpotter will calculate the SG difference score and the confidence score. These two scores must meet the detection conditions before the command is recognized.
2. Confidence Score(CS)：
   It represents the similarity between the voice and the recognition command. The higher the score, the more similar and better.
3. SG Difference(SG)：
   It represents the degree of dissimilarity between the voice and silence/garbage. The higher the score, the larger the gap, the less resembling, and the better. Garbage stands for non-command speech and environmental noise.

4. The testing conditions are:
   (CS >= 10 && SG >= 0) or (10 > CS >= 0 && SG >= 10)
5. Adjusting the weighted score (reward) of these two indicators for each instruction will change the recognition rate and the number of false triggers. Decreasing the weighted score (reward is a negative value) will reduce the number of false triggers but the recognition rate will be worse. Increasing the weighted score (reward is a positive value) will make the recognition rate better but the number of false triggers will be higher. The purpose of optimization adjustment is to strike a proper balance between the recognition rate and the number of false triggers.

3. Preparation of Command, Test File and Device

3.1. The Design Requirement of Command

The command that have been recognized for 24 hours are called wake words. In order to improve the recognition rate of wake words and reduce the number of false triggers, the following precautions are required:

1. At least 3 syllables, preferably 4 syllables. The more syllables, the false triggering is inherently lower. It is more easier to adjust and optimize, because the recognition rate is less sacrificed in the adjustment process.
2. The ending is best pronounced as a vowel. For example, Alexa, Hey Siri, and Ok Google all end in a vowel, and Book and News end in a consonant.
3. Choose words with strong pronunciation.
   Take English wake-up words as an example. Hey Siri's Hey has strong pronunciation energy at the beginning, and it is not easy to be covered up in a noisy environment, so the recognition rate is better than Alexa and Ok Google.
4. Don't recognize too many wake-up words at the same time. The more wake-up words, the higher the probability of false triggering. For the requirements that need to recognize many commands at the same time, please design a wake-up word and many command words.
5. If multiple commands are recognized at the same time, the pronunciation difference between commands must be large enough (more than two syllables) so that it is not easy to recognize errors due to the influence of accent, environmental noise, and device recording quality.
6. Each syllable should preferably contain consonants and vowels, and especially do not have words with only vowels in succession, because when speaking faster, it is easy to connect with the preceding sounds, the pronunciation becomes inconspicuous, the interval becomes shorter, or even the sound changes. The recognition rate will get worse.
7. In order to reduce the chance of false triggering, please use words and sentences that are not easy to appear in daily speech.
8. You can prepare several sets of candidate words, and then choose the most effective instruction after adjustment and optimization.

3.2. The Audio File for Recognition Test

1. It is ideal to record with a target device. If there is no time to prepare in the early stage, you can use the iPhone to record first, and then you need to use the target device to test and confirm before mass production.
2. 10 men and 10 women, or more, at least 5 men and 5 women
   Ideally, it should be based on the target audience, including men and women, various accents, and ages.
3. Please record in a quiet and low-reverberation environment (such as a small meeting room). The recording format is mono, 16 bits, 16 KHz, and please save it as an uncompressed wav file.
4. Each person, each use distance (Ex. Close distance, 1M, 3M), each command record 5~10 times and save them as an audio file, each sentence is separated by 2 seconds, and when the last sentence is finished, wait 2 seconds before stopping recording.
5. If you are using iPhone to record, you must convert the .m4a format to a mono, 16 bits, 16 KHz wav file.
6. Use the audio editor to play, check, edit (end silence) all audio files, and write the script file (*.spt) for DSMT testing.
7. The encoding of script file shall be Unicode 16 little endian (UTF16LE), and the format of each line is as follows:
   full_path_file_name\tCmd1Text/Cmd2Text/... \r\n
   For example: D:\M1\Command1.wav Turn on/Turn on/Turn on/Turn on

3.3. Recording Quality of The Device

The recording quality has a great influence on the recognition rate. The following are the requirements for the recording quality of voice recognition:

1. Format:
   16 KHz, 16 bits(-32768~32767), mono channel, PCM
2. It probably need to disable the pre-process function Auto Gain Control (AGC) and Noise Reduction (NR).
   Generally speaking, unless AGC and NR are specially adjusted for speech recognition, they will reduce the recognition rate because of the destruction of the sound spectrum.
   PS. The AGC suitable for speech recognition will not frequently adjust gain. The NR suitable for speech recognition will not damage the voiceprint.
3. SNR:
   In a quiet environment, speaking at a volume of 70dB at the longest distance

specified by the product, the SNR can exceed 20 dB, at least 15dB. If it cannot be achieved, consider using a pre-processing system suitable for voice recognition, such as a microphone array.

PS. The speaking volume is the average energy measured by the decibel meter when a person speaks one meter away from the decibel meter.

4. Recording volume (adjustment of recording gain):

   At the farthest use distance specified by the product, with a speech volume of 64dB, the maximum peak recording amplitude must be above 1000, preferably 4000. If not, a pre-processing system should be used.

   At the nearest use distance specified by the product, use a speaking volume of 70dB, and there must be no clipping (the volume exceeds -32768~32767, and the sound is cut off). If so, please change to a microphone with a higher AOP.

5. Transient response or VAD cut sound:

   The cut length of VAD, or the data at the beginning of recording, if there is an unstable transient response, the length to be dropped should be less than 100 ms to avoid too much impact on the recognition rate.

6. Stability:

   Record a hundred times, 8 seconds each time, check whether there is the same quality, the length of the transient and cut sound should also be kept the same length to facilitate subsequent adjustment processing.

7. No noise band:

   The frequency spectrum recorded in a quiet environment does not have a specific noise band. If there is, it is usually caused by circuit interference.

8. No weak band:

   Use an artificial mouth (need to perform frequency response correction with an equalizer) or monitor speakers to play white noise (full frequency band equal intensity noise) for close-range recording (10 cm). It is best to operate in an anechoic box or an anechoic room. The recorded sound is subjected to spectrum analysis, and there is no particularly weak frequency band. If there is, it is usually a mechanism design problem.

   The requirements in mechanism design:

   A. The hole of the mechanism should be larger than the hole of the MEMS microphone, preferably above 1 mm.
   B. The microphone should be aligned and close to the opening hole of the mechanism to avoid sound distortion caused by the formation of a resonant cavity.
   C. Rubber pads around the microphone can be used to keep airtight to prevent

noise from coming in from behind.

D. When designing the product, it is necessary to make sure that there is no barrier between the opening of the mechanism and the user's sound source (for example, the design is on the back of the product), otherwise the sound waves are transmitted into the microphone through spatial reflection, which will cause distortion.

9. The 100~8000Hz frequency response curve is close to the horizontal curve, the better. The total harmonic distortion is less then 1%.

4. Optimize Speech Model

4.1. Solve Recognition Problems for Test Audio File

1. Use DSMT offline test tool to count the recognition rate of quiet environment with the default setting of CS reward and SG reward.
2. Review the wave file if it gets the false reject (FR) result:
   A. If there is a recognition result at the beginning of the next audio file, it means that the ending silence of this audio file is not long enough, please use the audio editor to lengthen it.
   B. If you are sure that the pronunciation is incorrect, skip it and remove it from script file.
   C. Please try to add multiple pronunciations command to solve accent problem or other pronunciation problem.
3. Review the wave file if it gets the false accept (FA) result:
   If there are many false accept results that are confused with each other, it means that the pronunciation difference is not big enough. Please re-design the command (3.1) for this case. For example, "Turn on the light" and "Turn off the light" are changed to "Turn on the light" and "Turn the light off".

4.2. Optimization Steps

1. Benchmarks test:
   Without adjusting CS reward and SG reward, test the recognition rate of quiet and noise environment (please determine the type and intensity of noise according to the use environment), and the number of false triggers in 24 hours.
2. Adjust SG reward:
   If the SG of the false trigger recognition result is relatively low, and the SG of the recognition rate test result is relatively high, you can lower the SG reward to reduce false triggers. After setting it, do the recognition rate test and false trigger test.
3. According to the recognition rate of the previous test and the result of the false trigger test, set the appropriate CS reward for each command, based on the principle that it can reduce false triggers and the recognition rate is not drop too much (compared to the benchmark test). This step may be repeated do it several times.
4. If the false trigger condition (usually a command of less than 4 syllables) cannot

be met, or the recognition rate drops too much, start adding garbage command:

A. Add partial command:
   For example, "Turn on the light" => "Turn on"、"the light"
B. Replace one or two words by a similar pronunciation word:
   For example, “Hey Siri” => “Hi Cirrus”
C. To listen to the audio file from the false trigger test result (the audio file and location information that occurred), the garbage command is determined by the sound that caused the false trigger. For example, Hey Siri adds “have very”, “in serial”...

After adding the garbage command, you must do a quiet and noisy recognition rate test, false trigger test, and leave a garbage that has less impact on the recognition rate and reduce more on false triggers. This step may need to be repeated many times.