音声認識格闘録

もしくはUWPに敗北した記録

自己紹介

- 「ふぃるみすと」といいます。
- たまに「ふぃるちゃんねる」という名前で昼間に配信をしています。
- プログラムを作った経験はあんまりありません。
- C#で作った経験はもっとありません。
- UWPアプリを作った経験は皆無です。

動機

かわいい声になりたくて

最近Vの者が流行ってますね。かわいくなりたいですね。かわいい声で配信したいですね。

動機

かわいい声になりたくて

最近Vの者が流行ってますね。かわいくなりたいですね。かわいい声で配信したいですね。

方法は数あれど

大きくわけて2つの方法があります。

動機

かわいい声になりたくて

最近Vの者が流行ってますね。かわいくなりたいですね。かわいい声で配信したいですね。

方法は数あれど

大きくわけて2つの方法があります。

- ボイスチェンジャー的なもので声を変換する(例:まぐろなさん)

動機

かわいい声になりたくて

最近Vの者が流行ってますね。かわいくなりたいですね。かわいい声で配信したいですね。

方法は数あれど

大きくわけて2つの方法があります。

- ボイスチェンジャー的なもので声を変換する(例:まぐろなさん)
- 音声を文字に変換してから文字を読み上げる(例:のらきゃっとさん)

2つ目の方法は **ゆかりネット** の登場で普及しました。

音声認識からの読み上げ

ゆかりネット

ゆかりネットはChromeを経由して音声を文字に変換します。

音声認識からの読み上げ

ゆかりネット

ゆかりネットはChromeを経由して音声を文字に変換します。

つまり**Chromeがないと音声認識できない**ということです。

音声認識からの読み上げ

ゆかりネット

ゆかりネットはChromeを経由して音声を文字に変換します。

つまり**Chromeがないと音声認識できない**ということです。

宗教上の理由でChromeを使いたくない人(私)は 使えません。

音声認識からの読み上げ

ゆかりネット

ゆかりネットはChromeを経由して音声を文字に変換します。

つまり**Chromeがないと音声認識できない**ということです。

宗教上の理由でChromeを使いたくない人(私)は 使えません。

他の方法

音声認識からの読み上げ

ゆかりネット

ゆかりネットはChromeを経由して音声を文字に変換します。

つまり**Chromeがないと音声認識できない**ということです。

宗教上の理由でChromeを使いたくない人(私)は 使えません。

他の方法

Windows 10の機能のひとつになっている音声認識が使えそうです。

どうやって配信にのせようか

どうやって配信にのせようか

OBSに文字をのせる方法

1. テキストソース(ファイルから読み込み)
2. ブラウザソース(nodejsか何かでhtmlを表示させる)
3. OBSのWebSocketプラグインでソースを操作

どうやって配信にのせようか

OBSに文字をのせる方法

1. テキストソース(ファイルから読み込み)
2. ブラウザソース(nodejsか何かでhtmlを表示させる)
3. OBSのWebSocketプラグインでソースを操作

テキストソースを使うのが実装する上では一番楽です。

どうやって配信にのせようか

音声認識をするアプリを作る

Windows 10で音声認識をする場合、 標準の音声認識機能を使うにはUWPアプリを作らないといけません。

具体的に言うとUWPアプリからWinRT APIというものを使うことになります。

どうやって配信にのせようか

音声認識をするアプリを作る

Windows 10で音声認識をする場合、 標準の音声認識機能を使うにはUWPアプリを作らないといけません。

具体的に言うとUWPアプリからWinRT APIというものを使うことになります。

UWPアプリは作ったことがないので作り方を調べるところから始めました。

どうやって配信にのせようか

UWPアプリの作り方

流れとしては:

1. 画面をXAMLで作る
2. XAMLに対応したコードを作る

XAMLはXMLで書かれた画面記述用の言語です。

Visual Studioで作る場合、実際に表示される画面を見ながら作成できます。

UWPアプリを作る

UWPアプリを作る

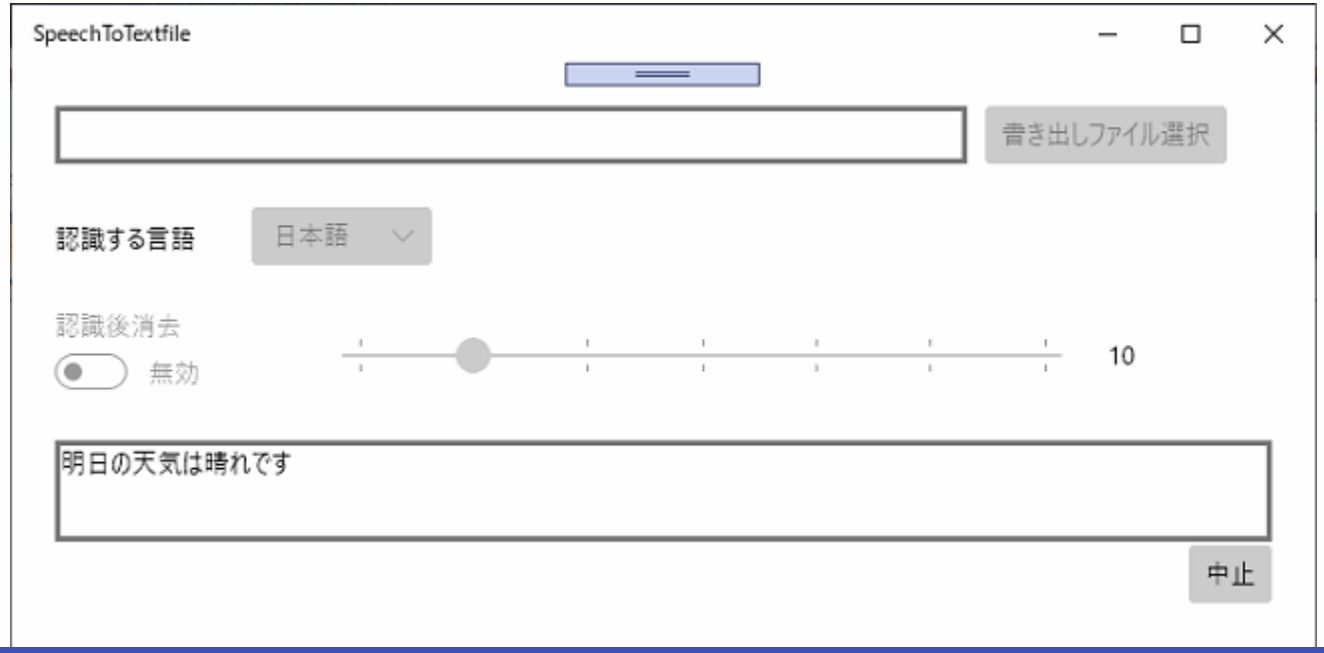
襲いかかるC#とUWPの仕様

- eventって何？
- async voidとasync Taskの違いって？
- 直接画面に値を書きこめないの？　UIスレッドっておいしい？
- バックグラウンドでタスクまわしっぱなしってできるの？

UWPアプリを作る

なんやかんやで

ながくくるしいたたかいだった。



UWPアプリを作る

問題点

UWPアプリを作る

問題点

アプリがバックグラウンドになると音声認識を停止してしまう。

- UWPアプリはバックグラウンドになると停止状態になる。

参考: https://docs.microsoft.com/en-us/windows/uwp/launch-resume/app-lifecycle

どうして(猫略)

UWPアプリを作る

UWPアプリ以外でもWindows 10の音声認識を使いたい

- Windows 10の音声認識は通常のアプリでも使えます。
- アプリをMSIXパッケージにして、固有のidentityをつけるとWindows 10の音声認識が呼びだせるようになります。

UWPアプリを作る

UWPアプリ以外でもWindows 10の音声認識を使いたい

- Windows 10の音声認識は通常のアプリでも使えます。
- アプリをMSIXパッケージにして、固有のidentityをつけるとWindows 10の音声認識が呼びだせるようになります。

Windows 10の音声認識をWPFアプリから使った

結果:

- 音声認識は通常のアプリでもバックグラウンドになると停止する。
- APIの使用に制限がかかっているのかも。

最終的にどうしたか

最終的にどうしたか

Microsoft AzureのSpeech To Textを使う

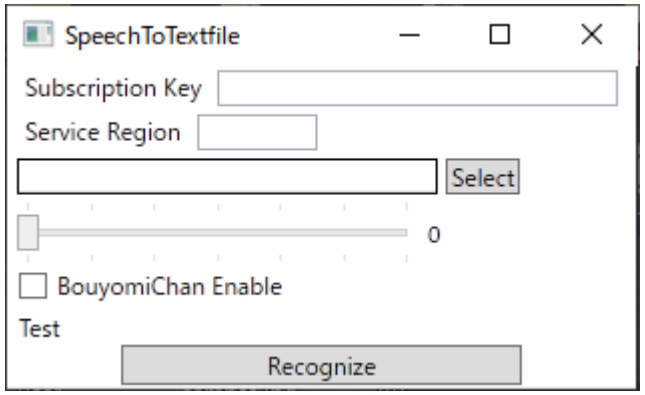
- Azureの音声認識サービスを使えばとりあえず認識できるでしょう。
- 無料枠が少しだけあるので試すのは簡単です。

最終的にどうしたか

一週間でできました。

最終的にどうしたか

一週間でできました。



かなり雑に作ったし、コードを流用できたから……

- ファイルの更新検知では読み上げが上手くいかなかったので、棒読みちゃん付属のソースを使いました。

最終的にどうしたか

そして

かわいくなりました。

最終的にどうしたか

問題点

- 自分の声以外がマイクに入ると認識率が格段に落ちます。
- 認識完了してから表示させるので、どうしてもラグが存在します。
- 声を出していない時間も含めて課金対象になります。
- 標準の状態では思っていたより認識精度が出ません。

単一指向性のマイクを使うなどして、自分の声以外の音が入らない環境を作るのが大前提になります。

その上でさまざまな余裕がある場合、 自分で学習させてカスタムされたモデルを使うことが認識精度の向上につながります。

最終的にどうしたか

価格

最終的にどうしたか

価格

Azure Speech Services (Microsoft)

- 無料プラン: 月に5時間まで (一応カスタムも可能)
- 標準プラン: 1時間あたり1.00USD、無料枠なし
- 標準カスタム: 1時間あたり1.40USD + エンドポイント1つあたり0.0538USD毎時(1か月39USDくらい)

ドキュメントを読む限り、クラウド主要3社の中では一番Windowsにおいて使いやすいと思います。

追加で語句を認識させたいとかノイズの多い環境に合わせたいときは、自分でデータを用意して学習させる必要があります。

最終的にどうしたか

価格

Google Cloud Speech To Text (Google)

- 標準モデル: 月に60分までは無料、それ以上は15秒あたり0.006USD(1時間あたり1.44USD)
- Googleにデータの利用を許可すると15秒あたり0.004USD(1時間あたり0.96USD)に割引
- 料金はリクエストごとに15秒単位へ切り上げ

言葉が豊富でノイズにも対応できるという売り文句です。

自分で単語を追加するのは無理そうです。

最終的にどうしたか

価格

Amazon Transcribe (Amazon)

- 1秒あたり0.0004USD(1時間あたり1.44USD)
- 最初に文字起こしをした時から12か月間は、月あたり60分まで無料
- 1リクエストの秒数が15秒未満の場合は15秒に切り上げ

AWSで学習のためにデータを利用するみたいですが、サポートに問い合わせて削除することも可能なようです。

最終的にどうしたか

価格

補足

有料でいいなら他にも音声認識サービスはあります。

- IBM Watson
- AmiVoice

いろいろ！

オフラインで認識したいならJulius(https://github.com/julius-speech/julius)が使えるかもしれません。

かわいくなるための補足

かわいくなるための補足

Vの者といえば絵が必要ですよね

絵を動かすソフトはいくつかあります。

- facerig

  > Steamで買えるとっても有名なソフト。Live2Dな絵を動かす人が多い。

- VRMを動かせるソフト

  > 3tene、VDRAW、VMagicMirror、バーチャルモーションキャプチャー

かわいくなるための補足

VRMって何ですか？

https://vrm.dev/

> 「VRM」はVRアプリケーション向けの人型3Dアバター（3Dモデル）データを扱うためのファイルフォーマットです。glTF2.0をベースとしており、誰でも自由に利用することができます。

- 主にニコニ立体(ニワンゴ)とかVRoid Hub(Pixiv)、booth(Pixiv)で配布されている。
- 人型モデルをUnityで使う形式に変換できればそこから作れる(UniVRM)。

VRM FANBOOK(m2wasabiさん著)に詳しく載っています。

https://booth.pm/ja/items/1037223

かわいくなるための補足

リップシンク

文字表示と同時に口を動かすのは難しいですが、合成した音声に合わせて口を動かすのはなんとかなります。

1. VB-CABLE(https://www.vb-audio.com/Cable/)を入れて、仮想オーディオデバイスを作ります。
2. VOICEROID2の出力先をVB-CABLE Inputに設定します。
3. VB-CABLE Inputをデバイスの設定から「スピーカーで聞く」ように設定します。
4. リップシンクできるソフトの入力をVB-CABLE Outputに設定します。

同じところが出しているVoicemeeter Bananaでもなんとかなると思います。

まとめ

- 実質100行に満たないコードを書けば済むくらいに今の音声認識は楽でした。
- GUIアプリを作るのもとても楽になりました。
- Googleに屈さなくても方法はあるよ。
- でも素直にGoogleのを使うほうが精度いいんですよね。

今回使ったプログラムのソースコードはそのうち公開するかもしれません。

謝辞

- 自分の声を出さずに配信できると知ったのはすまいるさんの配信を見たからです。
- MicrosoftのUWP音声認識サンプル(https://github.com/Microsoft/Windows-universal-samples/tree/master/Samples/SpeechRecognitionAndSynthesis)で最初の一歩を踏み出しました。
- 棒読みちゃんがあったからこそ声を出すことができました。
- VOICEROID2:琴葉葵で合成音声を出力しています。
- 3Dモデルは金子卵黄さんの「まんぞく葵ちゃんVRM(袖なし)」(https://3d.nicovideo.jp/works/td35206)を使っています。
- フェイシャルキャプチャソフトは3tene(株式会社プラスプラス)を使っています。
- 配信のノウハウを築いてきた先人に感謝します。
- YPv6(http://ypv6.pecastation.org/)いいよね。みんなもIPv6にしよう？