SSP の音声認識と音声合成

SSP の音声認識＆合成には、SAPI (Speech Application Programming Interface) と呼ばれる仕組みを使用しています。これは、Windows 上で音声認識・合成を使うための「標準的な枠組み」とされており、現在のメジャーバージョンは 5 です。
この実装時に苦労したことを、さらっと書いてみようと思います。

◆音声合成を使うまで
音声合成で SAPI を使おうとした場合に、いきなり引っかかるのが、音声合成エンジンの選択方法です。使用できる音声合成エンジンを列挙しようとしてレジストリを覗くと、以下の 3 種類があることに気づきます。

① OS 標準の合成エンジン(Speech)
② Office と一緒に使うために開発された Microsoft Speech Platform の合成エンジン
③ Windows ストアアプリで使うための合成エンジン(Speech_OneCore)

Windows 8 以降では、②に相当するものが①の OS 標準合成エンジンとして添付されるようになったのですが、それより前では音質に難があったため、レジストリをいろいろ細工して②を①として無理やり使う方法などの先人の記事が残っています。
SSP では、①と③が自動的に合成エンジンの選択肢に現れるようになっています。また特に選択しない場合は①の OS 標準のものが使われます。
②については、現在利用者が少ないと思われるので、無視しています。

◆音声合成の SSP 実装
2017/9 のうかべん横浜#9 の時は、とにかく喋らせることを目標に、かなり雑な実装をしていました。喋らせるだけなら ISpVoice->Speak を呼べば良いだけですから、たいへん楽です。
問題は複数ゴーストを起動した時です。きれいに聞き取れるようにするためには、

- 「複数のトーク実行を一度に喋らないようにする（しゃべるトークは 1 つのみ)」
- 「1 回のトーク実行が終わるまでの一連のテキストを連続して喋らせる」
- 「1 回のトーク実行中に別のゴーストからの喋りが割り込まないようにする」

以上の条件を満たすように、各ゴーストからばらばらのタイミングでやってくるテキスト群をうまく制御してやる必要があります。この制御を矛盾なくこなすのに手間取り、何度かバグ修正版を出すハメになっています。
また、今のところリップシンク（バルーンシンク？）が実装できていないのは、このような制御をしながらスクリプト実行と同期させる方法を思いついていないからです。

◆音声認識は？

音声合成の時の問題は、音声認識にもあてはまりました。同じく 3 種類の実装が併存できるという難解な事態です。ここでさらに厄介なのは、Windows 7 またはそれより前のバージョンでは、そもそも標準の音声認識エンジンが存在しないということです。

後から追加で購入しないと、今ではスマホであたりまえにできる音声認識すらできませんでした。

幸い 2017 年頃の場合は特に気にしなくてもよくなりましたが、それより前には手軽に使えるものはなく、プラグイン「miccom」（＋デモゴースト「かしこ」）のような SAPI 非互換の大掛かりな仕組みを用意するしか選択肢はありませんでした。

◆なぜ決まった選択肢の認識にしか使われていないのか

当初、せっかく音声認識なんだから、自由な文章を喋ってもらい、ゴースト側ではコミュニケートとして扱われるようにしようと組んでいました。しかし、実際組んで動かしてみると、どうしようもないぐらいに認識精度が悪く、ハチャメチャな文章がゴーストに渡される結果となりました。

これを解決しようと、しばらく音声認識のトレーニング機能を使い、私の声の癖を覚えさせようとしてみましたが、あまり精度が改善されず実用的ではありませんでした。もしかして…関西弁だから…？

もう 1 つ、コミュニケート機能を実装してみるとわかると思いますが、プログラムする側はどんな言葉に反応したら良いかいまいち決めづらく、使う側は何を喋ればゴーストが反応してくれるかわかりづらいという致命的な問題があります。

使う側の立場としては、やみくもに喋って（またはテキストを打ち込んで）みても、言っていることがわからない、という反応ばかりでは使っている気がしなくなりますよね。

これらの問題を解決するために、ISpRecoGrammar を使って、認識すべき文字列（文法）をできるだけ制限し、誤った認識をできるだけ避けるようにしています。

また、喋る側が何を喋っていいか理解できるように、「誰に」「何を」やってもらうという極めて単純な形で、認識の仕組みを整理してあります。

◆これからの機能追加

音声認識については、とりあえずある程度の完成形に至ったので、喋れるコマンドの追加以外には今のところ考えていません。もっと認識精度が上がって、かつ形態素解析など自由な文章の解析手段が楽になったら考えます。

音声合成は、VTuber 用ツールとしての需要がありますので、早急にリップシンクさせたいですね。ただしその場合、かなり制約された仕様になると思います。

文責：ぽな@ばぐとら　挿絵：狼牙改