



# **ASR1600 Engine Parameters**

## **Application Note**

Version 1.2

バージョン	内容	日付	著者
1	標準レイアウト	04-March-1998	F. Vanpoucke
2	マイナー バージョン	05-March-1998	F. Vanpoucke
3	Accuracy デフォルト値の更新	09-March-1998	F. Vanpoucke
4	コンフィデンス値の追加	12-May-2000	F. Vancraeyveldt

**Lernout & Hauspie Speech Products NV**  
 Flanders Language Valley, 50  
 B-8900 Ieper, Belgium  
 Phone: +32 57 22 8888 Fax: +32 57 20 8489

#### **Copyright**

(C) Lernout & Hauspie Speech Products N.V.  
 Document No. D30011-21-01 ASR1600 Engine Parameters  
 Application Note  
 V1.0 © - March 1998

本書の 1 部または全部を、電子的、あるいは機械的、すなわち写真複写、録音、またはいかなる情報検索システムによる、いかなる形態、いかなる手段で、L&H の書面による許諾なく、複製または配布することを禁止します。

---

# 目次

はじめに.....	2
自動ゲイン制御.....	3
音声検出.....	4
開始モード.....	4
ストップモード.....	5
音声デコーダ.....	6
リジェクションとキーワードスポッティング.....	8
コンフィデンス値.....	10

---

## はじめに

The ASR1600 SPI (サービス プロバイダ インターフェイス) は、ASR1600 音声認識エンジンの下位レベルのインターフェイスです。このインターフェイスには、いくつかのエンジン パラメータが用意されており、さまざまな環境のもとで音声認識を行うアプリケーションで、これらのパラメータに最適な値を設定することが可能です。

このパラメータは機能別にグループ化されています。音声の取り込みから認識結果に至るまで、音声信号に影響を与えるパラメータ グループを以下に示します。

- マイクロホンの増幅度
- 音声認識の開始と終了の検出
- 音声認識エンジンの探索の精度
- リジェクションとキーワード スポットティング
- スペリング ポスト プロセッサの探索の精度

各グループについての詳細については、以下の節で解説しています。

---

## 自動ゲイン制御

パラメータ	レンジ	デフォルト値
AGCON	ブール型	false
FARTALK	ブール型	True

音声認識エンジンでは、音声信号のアナログ ゲイン（電氣的増幅）をソフトウェアで調整することができます<sup>1</sup>。ただし、自動ゲイン制御 (AGC) では、信号レベルの急激な変動が発生するため、特に必要がない場合はこの機能を無効にしておくことをお勧めします。音声認識の精度を維持するには、API 初期化時のマイク チェック処理で取得した最適なアナログ ゲイン値を固定する必要があります。

FARTALK フラグは、自動ゲイン制御の感度を決定します。オフィス環境では、話者とデスクトップ マイクの距離が特に固定されないため、FARTALK フラグを「有効」にしておくことで広範囲の信号レベルを捕捉できます。それ以外の環境では、話者とマイクの距離が固定されるため、AGC が弱い背景雑音で反応しないように FARTALK フラグを「無効」にしておくことをお勧めします。

FARTALK フラグの推奨値	
オフィスのデスクトップマイク	True
車内マイク	False

---

<sup>1</sup>ハードウェア側のオーディオ設定が可能であることが条件です。

---

---

## 音声検出

パラメータ	レンジ	デフォルト値
START_ENABLE	ブール型	True
SENSITIVITY	0 - 4000 (dB/100)	1200 (12 dB)
MINSPEECH	10 - 400 (ミリ秒)	60
TS_ENABLE	ブール型	True
TS	100 - 2000 (ミリ秒)	300
TIMEOUT	0 - 30 (秒)	0 (タイムアウト無し)

注：音声信号には、通常の音声を挟むように隣接する無音状態の音素が多数存在します。

ASR1600 エンジンには、音声開始と終了を検出するためのモードに、手動モードと自動モードの 2 つがあります。

## 開始モード

アプリケーションから音声認識エンジンを起動させるには、まず最初に `casrStart` 関数を呼び出します。START\_ENABLE フラグの設定により、音声検出の開始を有効にしたり無効にすることができます。有効で指定した場合、音声認識エンジンは SLEEP モードで開始します。SLEEP モードでは、音声認識エンジンは入力音声を検出するために CPU 負荷量の少ない状態で待機します。この場合、SLEEP モードから RUN モードに遷移する音声設定の基準として、最短の MINSPEECH ミリ秒の間で音量を (SENSITIVITY/100) dB 以上で定義します。入力音声を検出されると、音声認識エンジンでは、RUN モードに自動的に切り替わり、CPU パワーをフルに活用しながら、入力信号と最も一致する単語を絞り込むためにグラマーを探索します。START\_ENABLE を 0 で指定した場合は、音声認識エンジンは、最初から RUN モードで起動します。

注：未知量の無音状態が音声よりも先行する場合は、音声認識エンジンを SLEEP モードで起動することをお勧めします。

SENSITIVITY パラメータには、音声認識を行う環境の信号対雑音比に比例する値を設定するのが定石ですが、このパラメータに高い値を設定すると音声認識エンジンの性能が低下するため、音声認識時により大きい音量が必要になります。また極端に高い値の設定は音声認識エンジンがスリープ状態（認識不可状態）に陥る可能性があるため注意が必要です。逆に、このパラメータに極端に低い値を設定すると、「背景音声」のような弱い非定常雑音環境のもとで音声認識が発生する恐れがあります。

SENSITIVITY パラメータの推奨値	
オフィス環境	2000
車内環境	1200

MINSPEECH パラメータは、マウスのクリックやドアの閉まる音のような瞬時に発生する高いエネルギー事象に反応することを避けるために指定します。通常はデフォルト値 (60ミリ秒) の設定で問題ありませんが、非定常雑音の強い環境下では、これを変更する必要があります。

## ストップ モード

以下の 3 つのいずれかの条件を満たすと、音声認識は停止します。

1. casrStop 関数呼び出し時 (手動停止)
2. 単語の認識後、音声認識エンジンで TS ミリ秒以上の無音状態を検出した時。ただし、自動ストップ モードが有効になっているのが条件 (TS\_ENABLEが 1 であること)。
3. レコグナイザの SLEEP/RUN/RECORD<sup>2</sup> のいずれかのモードで、TIMEOUT 秒が経過しても音声入力が行われなかった場合、音声認識を停止します。

注 : TIMEOUTを 0 に設定すると音声認識の停止機能は無効になります。

TIMEOUT パラメータに 0 以外の値を設定することにより、音声認識エンジンのアイドル状態、すなわち音声入力が行われていないにもかかわらず、音声認識エンジンが稼動状態 (ビジー状態) であることを避けることができます。

自動ストップモードでは、TS パラメータは後続無音最小継続時間を決定します。離散単語文法では、音声認識エンジンのレスポンスタイムが早くなるため、このパラメータに低い値を設定します。連続音声の場合は、話者の発声途中で音声認識結果がアプリケーションに渡ってしまう恐れがあるため、2 単語間のポーズ持続時間よりも大きい値を TS パラメータに設定する必要があります。これは、音声認識エンジン側で、2 単語間のポーズ持続時間を誤って後続無音継続時間と捉え、連続音声の区切りと判断するためです。

TS パラメータの推奨値	
離散単語	300 ミリ秒
連続音声	800 ミリ秒

<sup>2</sup>RECORD モードは、ユーザー単語トレーニング時に使用します。

---

## 音声デコーダ

パラメータ	レンジ	デフォルト値
MAXNBEST	1-1000	10
ACCURACY	100-10000	300

音声認識エンジンの中核をなすデコーダ モジュールは、最も予測される MAXNBEST 個の単語文字列のグラマーを探索します。ACCURACY パラメータには、常時同時に探索可能な仮説の数の値を設定します。仮説<sup>3</sup>の数が ACCURACY で指定した値よりも大きい場合は、整合度の最も低い仮説が、その探索対象から除外されます。デコーダ モジュールには通常、同じ単語文字列の複数のコピーがあります。コピーは、各文字単位にタイムスタンプの記録された「開始点」「終了点」を両端に持ち、コピーによってそれぞれのタイムスタンプが異なります。

システムリソースの節減を考慮に入れた場合、音声認識性能のロスを抑える一方で、ACCURACY パラメータの値に極力小さい値を設定する必要がありますが、設定した値が極端に小さいと、音声信号との仮説照合が最適に行われているにもかかわらず、先頭から整合性の低い部分が発生し、その仮説が除外される恐れがあります。

ACCURACY パラメータに最適な値は、音声認識のボキャブラリ サイズやグラマーにより異なりがあり、予測するのが困難です。以下の表に、分岐数から最適なパラメータ値を算出する式を示します。分岐数とは、有限状態グラマー上のあるポイントの、単語、または無音状態から枝分かれするように続く単語の数をいいます。

たとえば、離散単語グラマーには、N 単語とそれに隣接する先行/後続無音状態があります。その先行無音状態の次に N 単語が続き、単語ごとに唯一の後続語である後続無音状態が続きます。その最大分岐数は N、平均分岐数は  $2N/(N+2)$  で算出されます。

誤った認識結果の高速計算処理に時間を費やさないよう、ACCURACY パラメータに極力高い値を設定することを強くお勧めします。ただし、現在のリソースの使用状況、適切な音声認識のレスポンス タイムを考慮に入れる必要があります。特定のグラマーがない場合は、ACCURACY パラメータに任意の値をセットして音声認識の性能をテストしながら ACCURACY パラメータの調整をする方法も考えられます。テストの方法としては、ACCURACY パラメータにある程度高い値をセットし、徐々にその値を下げながらテストを行います。音声認識の結果が許容限界を超える直前の値が最適値といえるでしょう。

---

<sup>3</sup> 仮説とは、単語文字列と時間配分のコンビネーションのことをいいます。



---

ACCURACY パラメータの値に設定する値が高ければ高いほど、CPU の負荷は、ある一定のレベル（オフセット）からほぼ直線的に増加します。メモリの消費量も同様に増加しますが、CPU の負荷増加率ほど急激な伸びは示しません。

ACCURACY パラメータの推奨値	
離散単語	Max (300,6 x 単語数)
連続数字	300
スペリング (連続文字)	1500
キーワード スポットティング	Max (300,6 x キーワード数)
連続音声	Max (300, 6 x 最大分岐数, 50 x 平均分岐数)

---

## リジェクションとキーワードスポッティング

パラメータ	レンジ	デフォルト値
REJECTION	0-100	50
GARBAGE	0-100	50

グラマーのすべての単語と音声信号のマッチングの結果に関係なく、音声認識エンジンでは、認識結果をアプリケーションに渡します。その認識された単語文字列をアプリケーション側で受理するか、棄却するかはアプリケーション側で判断します。この機能を「リジェクション」と呼びます。

音声認識エンジンは、各認識結果のスコアとコンフィデンス値をアプリケーションに渡します。スコアは単語文字列モデルと音声信号のマッチ度を表します。スコアの値が大きければ大きいほど音声信号とマッチングした単語文字列間のマッチ度が低くなります。スコアは信号の長さに比例するため、スコアの値が高い単語文字列は、リジェクションの対象(受理あるいは棄却を判断する)にはなりません。

スコア値とは異なり、コンフィデンス値は、信号の長さに影響されません。コンフィデンス値は、認識された単語文字列モデルの信号と「平均音声」単語で構成されるモデルの信号のマッチングを比較することによって算出されます。「平均音声」モデルとは、音声、ノイズを含めた幅広い意味の音のクラスを指します。

あるしきい値(通常は 50%) を境に単語文字列の受理/棄却を行うリジェクション機能を、アプリケーション側で制御できるように REJECTION パラメータが用意されています。REJECTION パラメータの値が大きければ大きいほど、「平均音声」モデルのマッチ度が高くなり、コンフィデンス値が低くなります。

音声認識エンジンでは、文や各単語ごとにコンフィデンス値を付加します。音声フラグメントがより長く、音素数が増えるほど、そのコンフィデンス値はより信頼性の高いものになります。したがって、'a' や 'the' のような非常に短い単語では、コンフィデンス値の評価対象にならないため、単語レベルのコンフィデンス値の見極めは慎重に行う必要があります。

REJECTION パラメータを調整するには、以下の2つケースを考慮する必要があります。

- 誤りリジェクション: 正しく認識された単語がリジェクトされた場合
- 誤り受理: 誤って認識された単語が受理された場合

デフォルトの REJECTION の値は、上記の 2 つの誤りの値の合計値が最小になるように選択されています。アプリケーションによっては、「誤り受理」の数が「誤りリジェクション」の数を上回った場合、認識の精度が落ちる可能性があるため、自分で録音した大量の音声データを収集/解析しながら REJECTION パラメータを調整する必要があります。

---

音声認識エンジンでは、無視すべき音声を吸収するすべての音声モデル<...><sup>4</sup>を用意しています。これは、BNF グラマーで、<...><KeyWord><...> のように指定し、ある特定のキーワードだけを認識する場合に使用します。

REJECTION パラメータと同様、GARBAGE パラメーターは、<...> モデルのマッチ度を調整するのに使用します。GARBAGE の値が大きければ大きいほど、<...> として認識される信号の部分が長くなります。キーワードのミスヒット率と誤ったキーワードの検出率を個別に調整する場合は、GARBAGE パラメータの値を一旦デフォルト値に戻して調整することをお奨めします。

認識されたキーワードとともにコンフィデンス値も戻ってきます。リジェクション調整の際は、このコンフィデンス値を参考にするとよいでしょう。

---

<sup>4</sup>BNF グラマーでは「平均音声」を <...> で定義します。

---

---

## コンフィデンス値

これは、エンジンのパラメータではなく、認識結果コールバックによって、返されるコンフィデンス値です。結果のコールバックは、単語コンフィデンス値（文中の各単語につき）と、文コンフィデンス値（各文につき）を返します。単語コンフィデンス値は、ある単語が発話されたという確信がどれほどあるかを示します。同様に、文コンフィデンス値は、ある文が発話されたという確信がどれほどあるかを示します。発話文は、このコンフィデンス値に基づいて、配列の枠の中に順番に並べられます。つまり、配列エレメント0の枠には、もっとも発話された可能性の高い文が表示されます。このコンフィデンス値は、値が充分出ない場合、認識を不適当であると否認するために使用することもできます。（認識の承認、否認のための境界を設けることによって）。

レンジ	最低値	最高値
コンフィデンス値	0	10000

この値は、確率ではないということにご注意下さい。

コンフィデンス値が10000ということは、音声認識エンジンがある文（もしくは単語）が発話されたと確信しているということを示しますが、エンジン自体が100%正確ということではありません。

また、この値は、（ある程度）コンテキストに依存するということにも留意下さい。すべてのコンテキストにおいて、共通な理想値はありません。あるコンテキストにおいては、他のコンテキストよりも、高い値が返されることはしばしばあります。コンテキストのサイズと、単語の近似度は、コンフィデンス値の違いに大きく関わってきます。アプリケーション開発者にとって、有効な方略は、新しく作成されたコンテキストをテストし、どのような値が返されるのか試してみることです。