(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2005-115386
(P2005-115386A)
(43)公開日　平成17年4月28日(2005.4.28)

(51)Int.Cl.[7]	F I			テーマコード（参考）
G10L 15/28	G10L	3/00	561A	5D015
G10L 15/10	G10L	3/00	531F	
G10L 15/14	G10L	3/00	535A	
G10L 15/18	G10L	3/00	537H	

審査請求 未請求　請求項の数 14　OL　外国語出願　（全 20 頁）

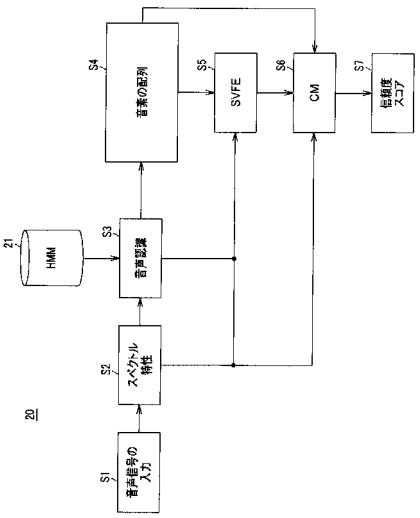
(21)出願番号　特願2004-294232 (P2004-294232)
(22)出願日　平成16年10月6日 (2004.10.6)
(31)優先権主張番号　03022646.8
(32)優先日　平成15年10月6日 (2003.10.6)
(33)優先権主張国　欧州特許庁 (EP)

(71)出願人　598094506
ソニー　インターナショナル　（ヨーロッパ）　ゲゼルシャフト　ミット　ベシュレンクテル　ハフツング
ドイツ連邦共和国　１０７８５　ベルリン　ケンパーブラッツ　１
(74)代理人　100067736
弁理士　小池　晃
(74)代理人　100086335
弁理士　田村　榮一
(74)代理人　100096677
弁理士　伊賀　誠司

最終頁に続く

(54)【発明の名称】誤認識予測方法

(57)【要約】
【課題】　入力音声信号（２）における変動は信号（２）が音声信号であるか非音声イベントであるかによって異なるという洞察に基づいた誤認識予測方法を提案する。
【解決手段】　この誤認識予測方法は、音声信号を入力するステップ（Ｓ１）と、入力音声信号（２）の少なくとも１つの信号変動特性を抽出するステップ（Ｓ２）と、入力音声信号に信号変動メータを適用して信号変動尺度を導出するステップ（Ｓ５）とを有する。
【選択図】　図２



【特許請求の範囲】
【請求項１】
音声認識システムにおいて誤認識を予測する誤認識予測方法において、
音声信号を入力するステップと、
上記入力音声信号の少なくとも１つの信号変動特性を抽出するステップと、
上記入力音声信号に対して信号変動メータを適用して信号変動尺度を導出するステップとを有する誤認識予測方法。
【請求項２】
上記信号変動メータは、上記入力音声信号の部分語に適用されることを特徴とする請求項１に記載の誤認識予測方法。
【請求項３】
上記入力音声信号の２つ以上の部分語から導出された信号変動尺度を組み合わせて、信頼度尺度を形成することを特徴とする請求項２に記載の誤認識予測方法。
【請求項４】
上記入力音声信号の２以上の部分語から導出された信号変動尺度の組合せは、相加平均、相乗平均、最大値、最小値、あるいは分散尺度に基づくことを特徴とする請求項３に記載の誤認識予測方法。
【請求項５】
上記分散尺度は、範囲、標準偏差又は相対分散であることを特徴とする請求項４に記載の誤認識予測方法。
【請求項６】
上記信号変動メータは、ユニークステート比の評価に基づくことを特徴とする請求項１乃至５のいずれか１項に記載の誤認識予測方法。
【請求項７】
上記信号変動メータは、同一ステート比の評価に基づくことを特徴とする請求項１乃至６のいずれか１項に記載の誤認識予測方法。
【請求項８】
上記信号変動メータは、ユニークステートエントロピーの評価に基づくことを特徴とする請求項１乃至７のいずれか１項に記載の誤認識予測方法。
【請求項９】
上記信号変動メータは、平均スペクトル変動の評価に基づくことを特徴とする請求項１乃至８のいずれか１項に記載の誤認識予測方法。
【請求項１０】
上記信号変動メータは、スペクトル不一致距離の評価に基づくことを特徴とする請求項１乃至９のいずれか１項に記載の誤認識予測方法。
【請求項１１】
上記信号変動メータは、ステート長単一比の評価に基づくことを特徴とする請求項１乃至１０のいずれか１項に記載の誤認識予測方法。
【請求項１２】
上記信号変動特性の抽出は、上記入力音声信号から取り出されたスペクトル特性に基づくことを特徴とする請求項１乃至１１のいずれか１項に記載の誤認識予測方法。
【請求項１３】
上記信号変動特性の抽出は、さらに、音声認識仮説ベースの隠れマルコフモデルによる音素配列に基づくことを特徴とする請求項１２に記載の誤認識予測方法。
【請求項１４】
音声認識システムにおける入力音声信号に関する認識仮説のための信頼度尺度を定めるコンピュータソフトウェア製品において、
請求項１乃至１３のいずれか１項に記載の誤認識予測方法を実行するようなデータ処理手段により処理される一連の状態要素により構成されることを特徴とするコンピュータソフトウェア製品。

【発明の詳細な説明】
【技術分野】
【０００１】
音声認識システムにおいて、一音一音はっきりと発音された（articulated）音声又は発話は、それぞれの音声信号を解釈することによって、それぞれ書き言葉（written language）に変換される。最新技術の音声認識システムの場合でも、雑音の多い環境で使用すると、通常、認識エラーと呼ばれる誤解釈が起こることが多い。入力音声信号に重畳された周囲の雑音は、入力信号の特性を変化させ、あるいは音声認識装置によって誤って音素として解釈されてしまう。
【背景技術】
【０００２】
誤認識が起こったことを検出するためには、所謂信頼度尺度（confidence measure）が用いられる。信頼度尺度とは、ある単語又は部分語（sub-word）が音声信号の特定部分に対応するときの信頼度を判定するものである。そして、単語又は部分語は、算出された信頼度尺度に基づき、認識処理において容認され又は拒否される。
【０００３】
様々な種類の言葉による表現（expression）が非常に類似した音声を有するので、ある発話の解釈として幾つかの選択肢がある場合が多い。いずれか１つを特定するには、例えば、信頼度尺度を、ある表現がそれぞれの発話に対応する尤度として計算する。これは通常、何らかの特別な統計的仮説検定により行われる。これらの処理は、通常、非常に複雑であり、特に、ある音素が隣接した音素の影響によって音響的変動を生じる可能性がある場合、すなわち同時調音（coarticulation）として知られる効果がある場合は複雑である。
【０００４】
また、上述の音声信号に重畳した周囲の雑音のような非音声イベント（non-speech event）も、音声信号に音響的変動を生じる。したがって、音声信号を書き言葉と同等なものに変換するための単語又は部分語の正しい認識処理は、煩雑な作業であり、満足な解決策は未だ得られていない。
【発明の開示】
【発明が解決しようとする課題】
【０００５】
したがって、本発明の目的は、音声認識システムにおける認識エラーの検出を改善する方法を提案することである。
【課題を解決するための手段】
【０００６】
上述の目的は、独立請求項に定める発明により達成される。
【０００７】
音声信号の特性は、例えば、周囲の雑音やノイズバースト等の非音声信号の特性とは非常に異なる。非音声信号と異なり、音声信号は、準周期的に、比較的小さく変動するという性質を有する。このため、入力音声信号の通常安定している部分に急激な変化や大幅な変動があれば、非音声成分が存在する可能性が高い。
【０００８】
この洞察に基づいて上述の目的を達成するために、本発明に係る誤認識予測方法は、音声認識システムにおいて、音声信号を入力するステップと、入力音声信号の少なくとも１つの信号変動特性を抽出するステップと、入力音声信号に対して信号変動メータを適用して信号変動尺度を導出するステップとを有する。
【０００９】
さらに、上述の目的を達成するために、本発明は、音声認識システムにおける入力音声信号に関する認識仮説のための信頼度尺度を定めるコンピュータソフトウェア製品を提供する。このコンピュータソフトウェア製品は、本発明に係る誤認識予測方法を実行して信

頼度尺度評価システムを形成するようにデータ処理手段により処理される一連の状態要素により構成される。

【００１０】

本発明の更なる有利な特徴については、各従属請求項に記載している。

【００１１】

信号変動メータは、以下に説明するように、入力音声信号の部分語、特に、フレーム、ステート又は音素に適用することが有利である。有利な展開によれば、入力音声信号の２以上の部分語から導出された信号変動尺度を組み合わせて、信頼度尺度を形成する。ここで、入力音声信号の２以上の部分語から導出された信号変動尺度の組合せは、好ましくは、相加平均、相乗平均、最大値、最小値、あるいは分散尺度に基づくものであり、分散尺度は、範囲、標準偏差又は相対分散により形成される。

【００１２】

信号変動メータは、ユニークステート比（Unique State Ratio）、同一ステート比（Same State Ratio）、ユニークステートエントロピー（Unique State Entropy）、平均スペクトル変動（Average Spectral Variation）、スペクトル不一致距離（Spectral Mismatch Distance）又はステート長単一比（State Length One Ratio）を評価し、あるいはこれらの２つ以上を組み合わせると、効果的である。

【００１３】

本発明の更なる好ましい実施例では、信号変動特性の抽出は、入力音声信号から導出したスペクトル特性に基づくものとすることにより、信号変動特性の抽出は、隠れマルコフモデルベースの音声認識仮説の音素配列に、十分基づくことができる。

【発明を実施するための最良の形態】

【００１４】

以下、図面を参照して、本発明の実施例を詳細に説明する。

【００１５】

図１は、時間の経過に対する入力音声信号２振幅の一部を示す波形図である。入力音声信号２は、音素の境界４、４’により互いに分離された幾つかの音素３、３’、３”に分割される。音素は、例えば「ｂ」を「ｐ」と区別するような最小の単語を識別できる信号として特徴付けられる音声である。これは、例えば音節等の部分語（sub-word）を既に表している場合もあるが、通常はもっと小さい単語成分である。この明細書においては、部分語という用語は、限定的なものではなく、音声として又は書き言葉として表現した場合の単語の一部を表すものである。

【００１６】

音声認識処理では、各音素が、所謂ステート（states）５_{1}、５_{2}、５_{3}及び$５'_{1}$、$５'_{2}$、$５'_{3}$等に細分化され、各ステートが更にフレーム６又は６’と呼ばれる小さい解析単位に細分化される。認識処理では、記録している信号の波形（以下、記録信号波形という。）からステート５_{2}内の入力信号に一致しているものを探索する。通常、完全に一致した波形は存在せず、すなわち、記録信号波形は、各ステート５_{2}の１つ以上のフレーム６内の入力音声信号２からずれている。したがって、様々な記録信号波形をステート５_{2}内の入力音声信号２と比較して、そのステート５_{2}内に存在するフレーム６に基づいて一致するものを検出する。したがって、様々な記録信号波形がステート５_{2}の各フレーム６内においてそれぞれ最も一致する可能性がある。これを図１に示すが、ここでは各記録信号波形が指示符号７で示されている。記録信号波形番号「１７」は、ステート５_{2}の１番目、４番目及び最後フレーム６においてよく一致し、記録信号波形番号「５」は、２番目、３番目及び５番目のフレームにおいて最も一致する。最後から３番目のフレームは、記録信号波形番号「２０」と一致し、最後から２番目のフレームは番号「２１」と一致する。

【００１７】

ステート５_{2}のレベル内の記録している信号からの上述の入力音声信号２のずれ（deviation）は、例えば背景雑音の侵入や話者の気分の変化等、非音声イベントにより生じる

ことが多い。このため、完全に一致することは殆どなく、ある程度の変動（variation）や不確実性を伴う。最も一致した場合であっても、間違っている可能性はある程度残る。

【００１８】

信頼度尺度（confidence measure system）方式の精度を向上させるために、本発明では、検出された各候補（match）の信頼度の特徴付け（characterisation）を行う。この特徴付けには、検出された各候補の信頼度を評価するための仮説を導き出す。仮説では、ある候補がおそらく正しいか誤っているかについてのステートメント（statement）を作成する。ステートメントは、通常、音声認識システムの後続の処理を簡単にするために、数値という形態をとる。仮説の値は、例えば、非常に確実な（very certain）、確実な（certain）、確実であろう（possibly certain）、確実かもしれない（perhaps certain）、おそらく有り得ない（more or less unlikely）、完全に有り得ない（completely unlikely）等、口語的な用語で表される判定に対応している。

【００１９】

本発明によれば、フレーム６のレベルまでの音声信号の変動を考慮した信号変動メータ（signal variation meter）を用いて仮説を生成する。本明細書の内容において理解される信号変動メータは、ステート５_{2}内の各フレーム６に対する一致結果を用いて、入力音声信号２から信号変動特性を抽出し、音声イベントのみから得られる信号成分に関する最高の候補を特徴付ける値を探し出す。各信号変動メータは、好ましくは、音声認識システムの信頼度尺度評価部内で用いられる。

【００２０】

以下に詳細に説明するが、上述した方法により仮説を生成するために、以下の信号変動メータを提案する。すなわち、ユニークステート比（Unique State Ratio：ＵＳＲ）、同一ステート比（Same State Ratio：ＳＳＲ）、ユニークステートエントロピー（Unique State Entropy：ＵＳＥ）、平均スペクトル変動（Average Spectral Variation：ＡＳＶ）、スペクトル不一致距離（Spectral Mismatch Distance：ＳＭＤ）、ステート長単一比（State Length One Ratio：ＳＬＯＲ）である。

【００２１】

ユニークステート比（Unique State Ratio：ＵＳＲ）：ステート５_{2}内の各フレーム６において、入力音声信号２に最も一致している記録信号波形、すなわち最高のフレーム候補（the best frame match）の識別子７を識別する。次に、ステート５_{2}内において、異なる最高のフレーム候補の数をカウントし、この数をステート５_{2}内にあるフレーム６の数で除算する。図１を参照して具体例を説明する。第１の音素３のステート５_{2}は、８個のフレーム６から構成されている。最高のフレーム候補は、フレーム毎に異なり、合計４個の異なる識別子、すなわち「１７」、「５」、「２０」、「２１」がある。ステート５_{2}内には全部で８個のフレーム６が存在するので、この場合のＵＳＲは０．５となる。

【００２２】

同一ステート比（Same State Ratio：ＳＳＲ）：上述と同様に、まず、ステート５_{2}内のフレーム６毎に最高のフレーム候補（最適フレーム一致）を識別する。次に、同じ最高のフレーム候補（最適フレーム一致）を有するフレームの数を判定する。そして、最も大きなカウント値を、各ステート５_{2}内に存在するフレーム６の数で除算する。図１に示す具体例では、最高のフレーム候補の識別子（最適フレーム一致）「１７」及び「５」が、それぞれ３つの異なるフレームに対して識別されている。他の２つ、すなわち「２０」及び「２１」は、それぞれ１つのフレーム６のみに対して識別されている。したがって、最も大きなカウント値は３であり、その結果、ＳＳＲ＝０．３７５となる。

【００２３】

ユニークステートエントロピー（Unique State Entropy：ＵＳＥ）は、以下のように定義される。

【００２４】

【数１】

$$USE = \frac{-\sum_{s}^{Ns}\left(\frac{c(s)}{N}\cdot\log\left(\frac{c(s)}{N}\right)\right)}{\log(Ns)}$$

【００２５】

ここで、Ｎｓは、例えばステート５$_2$等のステート内における様々な記録信号波形の総数であり、Ｎは、ステート（例えば５$_2$）内のフレーム６、６’の数であり、ｃ（ｓ）は、ステート（例えば５$_2$）内の各記録信号波形についてのフレームカウント数であり、ｓは、記録信号波形の識別番号である。図１の具体例では、ステート５$_2$内における一致を検出するために、４つの異なる記録信号波形、すなわち「１７」、「５」、「２０」、「２１」という識別番号７の記録信号波形を用いている。したがって、Ｎｓ＝４であり、ステート５$_2$を構成するフレームの総数（Ｎ）は８であり、ｃ（１７）＝３、ｃ（５）＝３、ｃ（２０）＝１、ｃ（２１）＝１となる。１つのステートのみにおける１つの記録信号波形に基づいて計算を行う場合、すなわちＮｓ＝１の場合、ゼロによる除算の問題が生じる。これについては、ＵＳＥの値を０に設定することにより対処する。

【００２６】

ステート５$_2$内における隣接した２つのフレーム６間の平均スペクトル変動（Average Spectral Variation）は、スペクトルフラックス判定（spectral flux determination）に基づく一種のオーディオコンテンツ解析を表す。

【００２７】

【数２】

$$ASV = \frac{1}{N_{coef}\cdot(W_{sw}-1)}\sum_{n=bf_{sw}}^{ef_{sw}}\sum_{k=1}^{N_{coef}}\left[\log\left(\left|F_n(k)\right|+1\right)-\log\left(\left|F_{n-1}(k)\right|+1\right)\right]^2$$

【００２８】

ここで、ｎは、ステート５$_2$のユニットｓｗにおけるフレームインデックスを表す。下限の値は、各ステートユニットｓｗの開始フレームｂｆ$_{sw}$であり、上限の値は、終了フレームｅｄ$_{sw}$である。Ｗ$_{sw}$は、ステート５$_2$内のフレーム数を表す。Ｎ$_{coef}$は、スペクトル係数の総数であり、｜Ｆ$_n$（ｋ）｜は、ｋ番目のスペクトル係数に対応するｎ番目のフレームの振幅スペクトルである。

【００２９】

スペクトルを表すのに、この具体例のような振幅スペクトルではなく、メル周波数ケプストラム係数（Mel Frequency Cepstrum Coefficient：ＭＦＣＣ）等の別のスペクトルベクトルも用いることができる。

【００３０】

スペクトル不一致距離（Spectral Mismatch Distance：ＳＭＤ）：記録信号による最高の候補の仮説と、各ステート５$_2$内における入力音声信号２との間の不一致量を、好ましくは、距離メータ（distance meter）によって判定する。例えば、平均ユークリッド距離を用いると、仮説の最高ガウスフレーム候補（最適ガウスフレーム一致）μ（ｋ）とステート５$_2$のユニットｓｗ内のスペクトルベクトルの間のスペクトル不一致距離ＳＭＤは、以下のようになる。

【００３１】

【数３】

$$SMD = \frac{1}{W_{sw}} \sum_{n=bf_{sw}}^{ef_{sw}} \left(\sum_{k=1}^{N_{coef}} (F_n(k) - \mu(k))^2 \right)^{1/2}$$

【００３２】

平均マハラノビス距離（average Mahalanobis Distance）を用いると、仮説の最高ガウスフレーム候補（仮説の最適ガウスフレーム一致）とステート５$_2$のユニットｓｗ内のスペクトルベクトルとの間のＳＭＤは、以下のようになる。

【００３３】

【数４】

$$SMD = \frac{1}{W_{sw}} \sum_{n=bf_{sw}}^{ef_{sw}} \left(\sum_{k=1}^{N_{coef}} \frac{(F_n(k) - \mu(k))^2}{\sigma(k)} \right)^{1/2}$$

【００３４】

これは、重み付けユークリッド距離分散（weighted Euclidian distance variance）に対応する。

【００３５】

平均スペクトル変動メータについて既に述べたように、ＭＦＣＣ等、他のスペクトルベクトルをスペクトル表現を用いることもできる。

【００３６】

ステート長単一比（State Length One Ratio：ＳＬＯＲ）は、ステート５$_2$内のフレーム数Ｎでステート５$_2$を分割することによって得られる単一のフレーム６内の最高フレーム候補の数によってられる。図１の具体例では、ステート５$_2$の６番目（識別番号７が「２０」）と７番目（識別番号７が「２１」）のフレームのみが一度だけ起こっている。したがって、この２つのフレームが条件を満たし、ステート５$_2$内のフレーム総数Ｎが８であるので、ステート長単一比は０．２５となる。

【００３７】

上述の信号変動メータの全てを組み合わせて、入力音声信号２を認識するための信頼度尺度を導き出すことができる。信頼度尺度は、ステート５$_2$のレベル又はそれより高レベルのいずれかに基づいて得ることができる。特に、発話が２つ以上のステート５$_2$からなる場合、後続のステート５$_2$に対して得られる信頼度尺度を組み合わせて、高レベルの信頼度尺度を形成することが有利である。高レベルは、一音素、一連の音素、一単語、あるいは発話全体とすることができる。組合せは、ステート５$_2$のレベルで算出された一連の信頼度尺度の相加平均又は相乗平均に基づくものでもよいが、最大値又は最小値の判定、あるいは範囲、標準偏差又は相対分散等の分散尺度に基づくものでもよい。分散尺度は、１つの単語及び／又は発話仮説におけるステート５$_2$のユニット信号変動尺度の統計的分布を抽出するのに用いられる。

【００３８】

信頼度スコアは、上述した信号変動メータの１つ又はそれらの組合せから直接導出することができ、あるいは多層パーセプトロン（multilayer perceptron）等の最新の分類による１つ以上の信号変動メータを組み合わせることによって、導き出すことができる。

【００３９】

本発明に係る信号変動メータの適用例について、図２を参照して説明する。フローチャート２０は、本発明に基づく信頼度スコア又は尺度を生成する音声認識処理の各ステップを示す。ステップＳ１において、入力音声信号２が供給される。ステップＳ２において、入力音声信号２からスペクトル特性を導出する。これは、振幅スペクトルでもよいが、メ

ル周波数ケプストラム係数等、他のスペクトルベクトルでもよい。ステップＳ３において、ステップＳ２で取り出したスペクトル特性から、認識仮説を生成する基礎をつくる。仮説を評価する音声認識装置は、隠れマルコフモデル（ＨＭＭ）２１に基づいている。ステップＳ４において、音素配列により、ステート5_2に対する少なくともフレーム単位で一致する記録信号波形の識別情報に基づいて、最高候補の情報を得る。ステップＳ５において、上述の信号変動メータのうちの１つ以上により、ステップＳ２～Ｓ４のスペクトル特性解析、ステップＳ３の認識の仮説、ステップＳ４の音素配列の各結果を統合して、信号変動特性の抽出（signal variation feature extraction：ＳＶＦＥ）を行う。ステップＳ６において、ステップＳ２～Ｓ４で得られた結果を用いて、信頼度尺度器ＣＭはステップＳ５の結果を信頼度尺度に変換し、ステップＳ７において、この信頼性尺度を信頼度スコアとして出力する。なお、ステップＳ６において、ステップＳ２、Ｓ３、Ｓ４の結果を必ずしも利用しなくてもよいが、これらの結果を利用することにより、音声認識を大幅に改善することができる。

【図面の簡単な説明】

【００４０】

【図１】例えば信号変動尺度を導出するのに用いる、幾つかに細分化された部分を有する入力音声信号を示す波形図である。

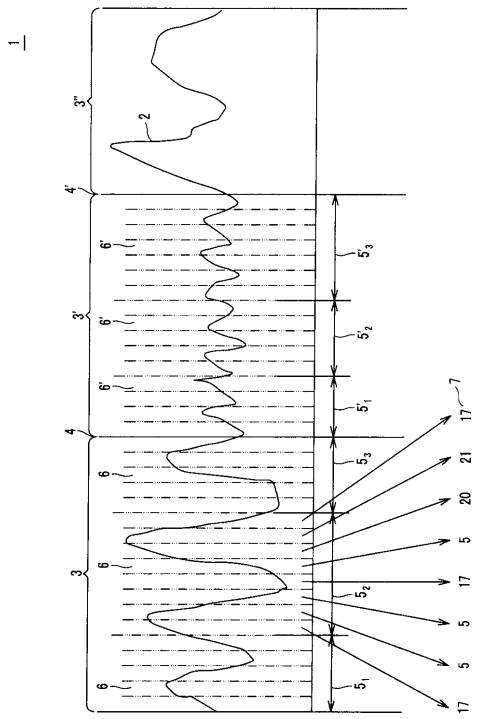
【図２】本発明に係る信号変動測定を用いて、信頼度尺度を決定する処理を示すフローチャートである。

【符号の説明】

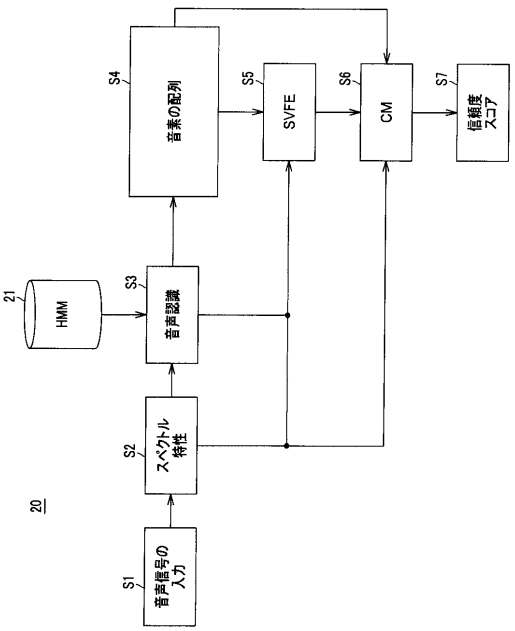
【００４１】

２　入力音声信号、３，３’，３”　音素、４，４’　音素の境界、５，5_1，5_2，5_3　ステート、６，６’　フレーム、７　記録信号波形の識別番号、Ｓ１　音声信号の入力、Ｓ２　スペクトル特性の導出、Ｓ３　音声認識仮説の生成、Ｓ４　最高候補の情報の取得、Ｓ５　結果の統合、Ｓ６　結果を信頼度尺度に変換、Ｓ７　信頼度スコアの出力

【図１】



【図２】



フロントページの続き

(72)発明者　イン　ハイ　ラム
ドイツ連邦共和国、７０３２７　シュトゥットゥガルト　ヘデルフィンガ　シュトラーセ　６１　ソニー　インターナショナル　（ヨーロッパ）　ゲゼルシャフト　ミット　ベシュレンクテル　ハフツング　シュトゥットゥガルト　テクノロジーセンター内
(72)発明者　トーマス　ケンプ
ドイツ連邦共和国、７０３２７　シュトゥットゥガルト　ヘデルフィンガ　シュトラーセ　６１　ソニー　インターナショナル　（ヨーロッパ）　ゲゼルシャフト　ミット　ベシュレンクテル　ハフツング　シュトゥットゥガルト　テクノロジーセンター内
(72)発明者　キルジストフ　マラセク
ドイツ連邦共和国、７０３２７　シュトゥットゥガルト　ヘデルフィンガ　シュトラーセ　６１　ソニー　インターナショナル　（ヨーロッパ）　ゲゼルシャフト　ミット　ベシュレンクテル　ハフツング　シュトゥットゥガルト　テクノロジーセンター内

Ｆターム(参考) 5D015 HH23 LL03

【外国語明細書】

Signal Variation Feature Based Confidence Measure

In speech recognition systems articulated sounds or utterances, respectively, are converted in written language by interpreting a respective speech signal. Misinterpretations which are usually referred to as recognition errors frequently occur with state-of-the-art speech recognition systems when used in a noisy environment. Ambient noise superimposing an input speech signal either modifies the characteristic of the input signal or may mistakenly be interpreted as a phoneme by a speech recogniser.

In order to detect if misrecognitions occur, so called confidence measures are used. A confidence measure judges the reliability with which a word or sub-word corresponds to a particular part of a signal. The word or sub-word is then accepted or rejected by the recognition process on the base of the confidence measure calculated for it.

As many different expressions sound very similar, there are often several alternatives possible for interpreting a certain utterance. To decide for one in particular, a confidence measure is e.g. calculated as the likelihood with which a certain expression corresponds to a respective utterance. This is usually accomplished by some form of special statistical hypothesis testing. These processes are usually very complicated, particularly as a phoneme can undergo certain acoustic variations under the influence of neighbouring phonemes, an effect which is known as coarticulation.

But also non-speech events, like the above mentioned ambient noise superimposing a speech signal result in an acoustic variation of the speech signal. A correct identification of the word or sub-word being the speech signal's written equivalent is therefore an elaborate task which is yet not been brought to a satisfactory solution.

It is therefore an object of the present invention to propose a system for improving the detection of recognition errors in a speech recognition system.

The above object is achieved by the invention as defined in the independent claims.

The characteristic features of a speech signal are quite different to those of a non-speech signal like e.g. ambient noise or noise bursts. In contrast to a non-speech signal, the

quasi-periodic behaviour of a speech signal results in comparatively small signal variation. A rapid change or considerable variation in normally steady parts of a speech input signal therefore most likely indicates the presence of a non-speech component.

Based on this insight, the above defined object is achieved by a method for predicting a misrecognition in a speech recognition system with steps for receiving a speech input signal, extracting at least one signal variation feature of the speech input signal, and applying a signal variation meter to the speech input signal for deriving a signal variation measure.

The above object is further achieved by a Computer-software-product for defining a confidence measure for a recognition hypothesis concerning a speech input signal in a speech recognition system. The computer-software-product comprises hereto a series of state elements which are adapted to be processed by a data processing means such, that a method according the present invention may be executed thereon to form a confidence measure evaluation system.

Additional advantageous features of the present invention are claimed in the respective sub-claims.

The signal variation meter is advantageously applied to a sub-word of the speech input signal, particularly to a frame, a state or a phoneme as described below. According to an advantageous development, the signal variation measures derived for two or more sub-words of the speech input signal are combined to form a confidence measure. Hereby, the combination of the signal variation measures derived for two or more sub-words of a speech input signal may preferably be based on an arithmetic mean, geometric mean , maximum value, minimum value or on a dispersion measure, whereby the dispersion measure is formed by a range, standard deviation or relative dispersion.

The signal variation meter is effectively based on a Unique State Ratio, Same State Ratio, Unique State Entropy, Average Spectral Variation, Spectral Mismatch Distance or State Length One Ratio evaluation or a combination of two or more of these.

According to a further advantageous embodiment of the present invention, the extraction of the signal variation is based on a spectral feature derived from the speech input signal, whereby the extraction of the signal variation may suitably be based on a phoneme alignment of a Hidden Markov Model based speech recognition hypothesis.

In the following, the present invention will be explained in detail and by way of example only, with reference to the attached Figures, wherein

Figure 1 shows a speech input signal with an indication of several subdivisions as for example used to derive a signal variation measure, and

Figure 2 shows a system for determining a confidence measure with a signal variation metering according to the present invention.

The diagram 1 of Figure 1 shows part of a speech input signal 2 in an amplitude versus time representation. The speech input signal 2 is divided into several phonemes 3, 3' and 3'' which are separated from each other by phoneme boundaries 4 and 4'. A phoneme is a sound characterised as the minimal word distinguishing signal like e.g. a *b* in contrast to a *p*. It may already represent a sub-word like for instance a syllable but usually it is a smaller word component. The term sub-word is used within this specification without any restriction and to indicate that only a part of a word in its acoustical or written representation is referred to.

For the speech recognition process, each phoneme is subdivided into so called states 5_1, 5_2, 5_3, and $5'_1$, $5'_2$, $5'_3$ etc., and each state is further subdivided into the smallest analysing unit called frame 6 or 6', respectively. The recognition process looks for matches of a recorded signal form to the input signal within a state 5_2. Usually, there will be no exact match which means that the recorded signal form will deviate from the speech input signal 2 in one or more frames 6 of a respective state 5_2. Different recorded signals forms are therefore compared to the input signal 2 within a state 5_2 of interest, and the matching is controlled on the basis of the frames 6 present in that state 5_2. Different recorded signal forms are therefore likely to form the respective best match in different frames 6 of a state 5_2. This is indicated in Figure 1 where each recorded signal form is referenced by an identification number 7. While the recorded signal form number '17' produces a good match in the first, fourth and last frame 6 of state 5_2, the recorded signal form number '5' fits best in the second, third and fifth frame. The third last frame is found to match with the recorded signal form number '20', and the last but one frame with number '21'.

The described deviation of the speech input signal 2 from a recorded signal within the level of a state unit 5_2 is mostly caused by non-speech events like e.g. background noise insertions or changes in the mood of the person talking. A match will therefore seldom be perfect but rather tainted with a certain variation or uncertainty. Even for the best match there remains a certain probability for it being wrong.

To improve the accuracy of a confidence measure system the present invention introduces a characterisation of the reliability of each detected match. For the characterisation a hypothesis is derived for assessing the reliability of each match found. The hypothesis forms a statement about a certain match being probably right or wrong. It usually takes on the form of a numerical value for an easy further processing by a speech recognition system. The value of the hypothesis corresponds to a judgement which may be expressed in colloquial terms as e.g. very certain, certain, possibly certain, perhaps certain, more or less unlikely, completely unlikely or something like that.

According to the present invention a hypothesis is generated by utilising signal variation meters which consider the speech signal variations down to a frame 6 level. A signal variation meter as it is understood in the context of this specification uses the matching results for each frame 6 in a state 5_2 to extracts signal variation features from the speech input signal 2 so as to ascertain a value characterising the best match regarding the signal component originating from a speech event only. A respective signal variation meter is preferably used within a confidence measure evaluation unit of a speech recognition system.

The following signal variation meters, which are explained in detail below are proposed to generate a hypothesis according to the previously explained: Unique State Ratio (USR), Same State Ratio (SSR), Unique State Entropy (USE), Average Spectral Variation (ASV), Spectral Mismatch Distance (SMD), and State Length One Ratio (SLOR).

Unique State Ratio: For each frame 6 in a state 5_2, the identity 7 of the recorded signal form which matches the speech input signal 2 therein best, i.e. the best frame match, is identified. Next, the number of different best frame matches for a state 5_2 is counted and divided by the number of frames 6 present within said state 5_2. An example can be given with reference to Figure 1. The state 5_2 of the first phoneme 3 is composed of eight frames 6. The best frame matches differ from frame to frame with a total of four different identities, namely '17', '5', '20', and '20', As there are altogether eight frames 6 present in state 5_2, its USR is computed to 0.5.

Same State Ratio: Like before, first the best frame matches are identified for each frame 6 within a state 5_2. Next, the number of frames having the same best frame match are determined. The highest count is then divided by the number of frames 6 present in the respective state 5_2. In the example illustrated in Figure 1, the best frame matches

'17' and '5' are each identified for three different frames respectively.. The other two, '20' and '21' each only for one frame 6. The highest count therefore amounts to three, resulting in an SSR = 0.375.

The Unique State Entropy is defined as:

$$USE = \frac{-\sum_{s}^{Ns}\left(\frac{c(s)}{N}\cdot\log(\frac{c(s)}{N})\right)}{\log(Ns)}, \qquad (1)$$

wherein N_s denotes the total number of different recorded signal forms in a state as e.g. 5_2, N the number of frames 6, 6' within the state (e.g. 5_2), c(s) the count of frames for a respective recorded signal form in the state (e.g. 5_2), and 's' is the identification number of a recorded signal form. In the example of Figure 1, four different recorded signal forms were used to match state 5_2, namely the signal forms with the identification numbers 7 of '17', '5', '20', and '21'. Hence $N_s=4$, the total number of frames (N) constituting state 5_2 is 8, and $c(17)=3$, $c(5)=3$, $c(20)=1$, $c(21)=1$. If the calculation is performed on the basis of one recorded signal form only in one state, eg., i.e. $N_s=1$, a division-by-zero problem arises, which is handled by setting the value for USE to 0 for this case.

The Average Spectral Variation between two adjacent frames 6 in a state 5_2 represents a sort of audio content analysis based on a spectral flux determination. It is defined by:

$$ASV = \frac{1}{N_{coef}\cdot(W_{sw}-1)}\sum_{n=bf_{sw}}^{ef_{sw}}\sum_{k=1}^{N_{coef}}[\log(|F_n(k)|+1)-\log(|F_{n-1}(k)|+1)]^2; \qquad (2)$$

Herein n signifies the frame index in the state 5_2 unit sw; Its lower value is the begin frame bf_{sw} and its upper value is the end frame ef_{sw} of the respective state unit sw. W_{sw} denotes the number of frames within said state 5_2, N_{coef} the total number of the spectral coefficients, and $|F_n(k)|$ the amplitude spectrum of the n^{th} frame corresponding to the k^{th} spectral coefficient.

Instead of an amplitude spectrum like in the example given, other spectral vectors such as a Mel Frequency Cepstrum Coefficient (MFCC) may be used for the spectral representation.

Spectral Mismatch Distance: The amount of mismatch between a hypothesis for the best match formed by a recorded signal form and the speech input signal 2 in a respective

state 5_2, is preferably determined by a distance meter. By e.g. using the average Euclidean distance, the Spectral Mismatch Distance between the best Gaussian frame match μ(k) of the hypothesis and the spectral vectors in the state 5_2 unit sw is

$$SMD = \frac{1}{W_{sw}} \sum_{n=bf_{sw}}^{ef_{sw}} \left(\sum_{k=1}^{N_{coef}} (F_n(k) - \mu(k))^2 \right)^{1/2} . \quad (3)$$

Using the average Mahalanobis Distance, the SMD between the best Gaussian frame match of the hypothesis and the spectral vectors in the state 5_2 unit sw will become:

$$SMD = \frac{1}{W_{sw}} \sum_{n=bf_{sw}}^{ef_{sw}} \left(\sum_{k=1}^{N_{coef}} \frac{(F_n(k) - \mu(k))^2}{\sigma(k)} \right)^{1/2} , \quad (4)$$

which corresponds to a weighted Euclidian distance variance.

Like mentioned before with reference to the Average Spectral Variation meter, other spectral vectors such as an MFCC can be used for the spectral representation.

The State Length One Ratio is given by the number of best frame matches in a state 5_2 which last for only one frame 6 divided by the number of frames N within said state 5_2. In the example of Figure 1 only the sixth (with an identification number 7 of '20') and the seventh frame (with an identification number 7 of '21') of state 5_2 occur just once. Therefore two frames fulfil the condition, and with the total number of frames N within state 5_2 being 8, the State Length One Ratio amounts to 0.25.

All signal variation meters described up to now may be combined to derive a confidence measure for the speech input signal 2 to be recognised. The confidence metering may either be based on a state 5_2-level or on a higher level. Particularly, when an utterance consist of more than one state 5_2, the confidence measures obtained for subsequent states 5_2 may be advantageously combined to form a higher level confidence measure. The higher level may be a phoneme, a series of phonemes, a word or a complete utterance. The combination may be based on an arithmetic or geometric mean of a series of confidence measures calculated on a state 5_2-level, but also on a maximum or minimum determination or a dispersion measure as e.g. a range, standard deviation, relative dispersion or the like. The dispersion measures are used to extract the statistical distribution of a state 5_2 unit signal variation measure in a word and/or utterance hypothesis.

A confidence score may be derived directly from one of the above described signal variation meters or a combination thereof, or by combining one or more signal variation meters with a state-of-the art classifier like a multilayer perceptron.

The application of a signal variation meter according to the present invention is described with reference to Figure 2. The flowchart 20 illustrates the individual steps of a speech recognition process yielding a confidence score or measure, respectively, according to the present invention. Beginning with step S1, a speech input signal 2 is received. Step S2 retrieves the spectral features from the speech input signal 2. This may be an amplitude spectrum or any other spectral vector such as for instance a Mel Frequency Cepstrum Coefficient. The spectral features derived in step S2 form the base for producing a recogniser hypothesis in step S3. The speech recogniser evaluating the hypothesis is based on a Hidden markov Model (HMM) 21. A phoneme alignment applied in step S4 provides the best match information based on an identification of at least framewise matching recorded signal forms for a state 5_2. In step S5, the results from the spectral feature analysis of step S2 to S4, the recogniser hypothesis of step S3, and the phone alignment of step S4 are merged by one or more of the above described signal variation meters to perform a signal variation feature extraction. Using results obtained in step S2 to S4, the confidence Meter CM transforms in step S6 the result of step S5 into a confidence measure delivered as a confidence score in step S7. Although the usage of the results from S2, S3, and S4 in step S6 is not mandatory, it is preferred, since it improves the speech recognition significantly.

List of Reference Signs

1	diagram
2	speech input signal
3, 3', 3''	phoneme
4, 4'	phoneme boundary
5, 5_1, 5_2, 5_3	state
6, 6'	frame
7	identification number of recorded signal form
20	flowchart
21	HMM
S1	receiving speech input
S2	retrieve spectral features
S3	produce a recogniser hypothesis
S4	provide best match information
S5	merging results
S6	transform results in confidence measure
S7	deliver confidence score

Claims

1. A method for predicting a misrecognition in a speech recognition system, the method comprising the following steps:
   - receiving a speech input signal,
   - extracting at least one signal variation feature of the speech input signal, and
   - applying a signal variation meter to the speech input signal for deriving a signal variation measure.

2. A method according to claim 1,
characterised in
that the signal variation meter is applied to a sub-word of the speech input signal.

3. A method according to claim 2,
characterised in
that the signal variation measures derived for two or more sub-words of the speech input signal are combined to form a confidence measure.

4. A method according to claim 3,
characterised in
that said combination of the signal variation measures derived for two or more sub-words of a speech input signal is based on an arithmetic mean, geometric mean , maximum value, minimum value or on a dispersion measure.

5. A method according to claim 4,
characterised in
that said dispersion measure is the range, standard deviation or relative dispersion.

6. A method according to one of the claims 1 to 5,
characterised in
that the signal variation meter is based on a Unique State Ratio evaluation.

7. A method according to one of the claims 1 to 6,

characterised in
that the signal variation meter is based on a Same State Ratio evaluation.

8. A method according to one of the claims 1 to 7,
characterised in
that the signal variation meter is based on a Unique State Entropy evaluation.

9. A method according to one of the claims 1 to 8,
characterised in
that the signal variation meter is based on an Average Spectral Variation evaluation.

10. A method according to one of the claims 1 to 9,
characterised in
that the signal variation meter is based on a Spectral Mismatch Distance evaluation.

11. A method according to one of the claims 1 to 10,
characterised in
that the signal variation meter is based on a State Length One Ratio evaluation.

12. A method according to one of the claims 1 to 11,
characterised in
that the extraction of the signal variation is based on a spectral feature derived from the speech input signal.

13. A method according to claim 12,
characterised in
that the extraction of the signal variation is further based on a phoneme alignment of a Hidden Markov Model based speech recognition hypothesis.

14. A Computer-software-product for defining a confidence measure for a recognition hypothesis concerning a speech input signal 2 in a speech recognition system, the computer-software-product comprising a series of state elements which are adapted to be processed by a data processing means such, that a method according to one of the claims 1 to 13 may be executed thereon.

Abstract

Based on the insight that variations in a speech input signal (2) are different depending on the origin of the signal (2) being a speech or a non-speech event, the present invention proposes method for predicting a misrecognition in a speech recognition system (2) with steps for receiving a speech input signal (S1), extracting at least one signal variation feature (S2) of the speech input signal (2), and applying a signal variation meter to the speech input signal for deriving a signal variation measure (S5).

(Figure 2)

【図１】

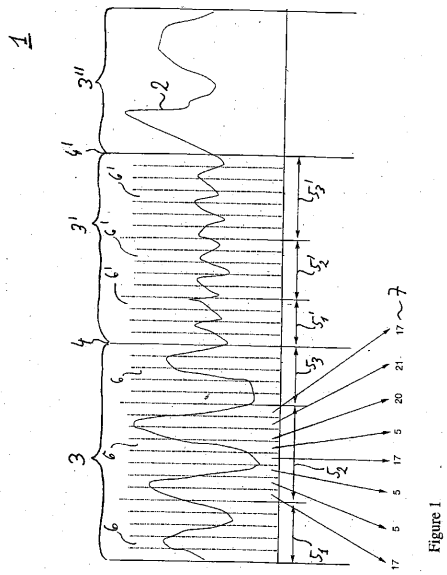


Figure 1

【図２】

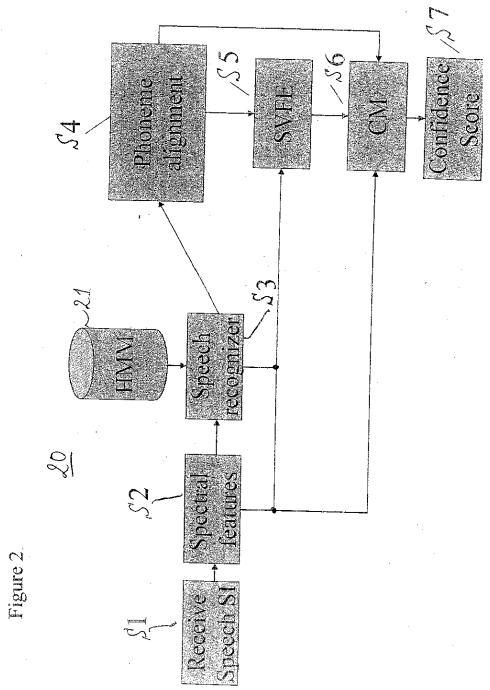


Figure 2