(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4631076号
(P4631076)
(45)発行日　平成23年2月16日(2011.2.16)　(24)登録日　平成22年11月26日(2010.11.26)

(51)Int.Cl.　F I
G10L 15/06 (2006.01)　G10L 15/06 3OOD
G10L 15/06 4OOV

請求項の数 5　外国語出願　(全 10 頁)

(21)出願番号　特願2004-318208 (P2004-318208)
(22)出願日　平成16年11月1日 (2004.11.1)
(65)公開番号　特開2006-126730 (P2006-126730A)
(43)公開日　平成18年5月18日 (2006.5.18)
審査請求日　平成19年10月12日 (2007.10.12)

（出願人による申告）平成１６年度独立行政法人情報通信研究機構、研究テーマ「大規模コーパスベース音声対話翻訳技術の研究開発」に関する委託研究、産業活力再生特別措置法第３０条の適用を受ける特許出願

(73)特許権者　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(74)代理人　100099933
弁理士　清水　敏
(72)発明者　張　勁松
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　フランク・スーン
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　中村　哲
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内

最終頁に続く

(54)【発明の名称】音素単位セットを最適化する方法及びシステム

(57)【特許請求の範囲】
【請求項１】
予め定められた言語の音素単位セットを最適化する方法であって、コンピュータに、
コンピュータ読出可能なフォーマットで基本単位セットを準備するステップと、
前記基本単位セットにリーブ・ワン・アウト法を適用することによって複数個の基本単位サブセットを生成するステップと、
前記基本単位サブセットの各々について言語的識別力の所定の尺度を計算するステップと、
前記基本単位セットを、前記基本単位サブセットのうち最も高い言語的識別力を備えたもので置換えるステップと、
前記生成するステップ、計算するステップ、及び置換えるステップを、所定の基準が満たされるまで繰返すステップとを実行させる、予め定められた言語の音素単位セットを最適化する方法。
【請求項２】
前記計算するステップが、前記基本単位セットと、前記基本単位サブセットの各々との間の相互情報量を計算するステップを含む、請求項１に記載の方法。
【請求項３】
前記置換えるステップが、前記基本単位セットを、前記基本単位サブセットのうち前記計算するステップで計算された相互情報量の最も高い値を有するもので置換えるステップを含む、請求項２に記載の方法。

【請求項４】
前記基本単位セットは前記予め定められた言語のための基本音素セットである、請求項１～請求項３のいずれかに記載の方法。
【請求項５】
予め定められた言語の単位セットを最適化するシステムであって、
基本単位セットをコンピュータ読出可能なフォーマットで記憶するための記憶手段と、
前記基本単位セットにリーブ・ワン・アウト法を適用することによって複数個の基本単位サブセットを生成するための生成手段と、
前記基本単位サブセットの各々について言語的識別力の所定の尺度を計算するための計算手段と、
前記記憶手段に記憶された前記基本単位セットを、最も高い言語的識別力を有する基本単位サブセットで置換えるための置換手段と、
前記記憶手段、生成手段、計算手段及び置換手段を、所定の基準が満たされるまで繰返し動作するよう制御するための制御手段とを含む、システム。
【発明の詳細な説明】
【技術分野】
【０００１】
この発明は自動音声認識（Ａｕｔｏｍａｔｉｃ Ｓｐｅｅｃｈ Ｒｅｃｏｇｎｉｔｉｏｎ：ＡＳＲ）に関し、特に、ＡＳＲで用いられる音素セット等の音素単位セットの最適化に関する。
【背景技術】
【０００２】
ＡＳＲはマン－マシン－インタラクションにおける必須のツールである。ＡＳＲによって、コンピュータは自然言語によるオペレータの指令を理解することができ、オペレータはコンピュータのための複雑なコマンドシステムを学ぶ必要がなくなる。
【０００３】
図６は基本的なＡＳＲの機構を示す。図６を参照して、ＡＳＲシステム１６２は、入力音声Ｘ１６０をデコードし、認識された（デコードされた）単語＾Ｗ１６４（文中「＾」の記号は本来文字Wの上に付されるものである。）を、以下の式１６６を用いて出力する。
【０００４】
【数１】

$$\hat{W} = \arg\max_{W} P(X,W)$$

$$= \arg\max_{W} P(X \mid W) \cdot P(W)$$

ここでＰ（Ｘ｜Ｗ）は音響モデル確率を示し、Ｐ（Ｗ）は言語モデル確率を示す。これらのモデルは対象となる言語の単語を、それぞれの音素のシーケンスと共に記載するレキシコンを用いて構築される。音素は予め定められた基本音素セットのうちから選択される。
【０００５】
大語彙連続音声認識（Ｌａｒｇｅ Ｖｏｃａｂｕｌａｒｙ Ｃｏｎｔｉｎｕｏｕｓ Ｓｐｅｅｃｈ Ｒｅｃｏｇｎｉｔｉｏｎ：ＬＶＣＳＲ）システムでは、広く受入れられた音素セットが用いられる。
【発明の開示】
【発明が解決しようとする課題】
【０００６】
簡単なＬＶＣＳＲタスクと、より複雑なＬＶＣＳＲタスクとで同じ音素セットを用いるべきか、という問題がある。数字認識タスク等の小さな語彙のタスクでは、数字等の単語が基本単位として用いられる。同様に、簡単なＬＶＣＳＲタスクでは、簡単な音素セット

を用いることが有利かもしれない。

【０００７】

ＡＳＲに関する多くの研究では、いくつかの発見的手法により決定された音素セットが試され、ＡＳＲ認識性能に基づいて、１セットが選択される。

【０００８】

音素セットにより多くの単位が含まれれば、音素学的により識別性のある情報を提供するであろう。しかしこれは、より詳細な音響的差異を使用するという意味でもある。音声認識の場合、より詳細な、またはより小さい音響差異をモデル化する必要が生じると、ＡＭ（Ａｃｏｕｓｔｉｃ　Ｍｏｄｅｌ：音響モデル）の頑健性が低下する傾向がある。

【０００９】

音素セットに含まれる単位数が少なければ、より大きな音素セットに比べて、各音素ＡＭは、より多くのトレーニングデータを有することが通常である。さらに、音素の数が少ない場合、音素間での差異は多くの音素間での差異より大きくなる傾向がある。この結果、音素セットが小さくなればＡＭはより頑健になり得る。しかし、音素セットサイズを小さくすると別の問題が生じる。すなわち、言語空間内における識別力が失われることである。例えば、日本語の長母音「Ａ」と短母音「ａ」とが一つの母音にマージされるので、単語間の混同が増加するであろう。

【００１０】

この点に関して、最新のＡＳＲ最適化は以下の考え方により行なわれる。上述の式を以下の形に書くことができる。

【００１１】

【数２】

$$\begin{aligned}\arg\max_W P(X,W) &= \arg\max_W P(X\mid W)P(W)\\ &= \arg\max_W \sum_F P(X\mid F,W)P(F,W)\\ &\approx \arg\max_W \sum_F P(X\mid F)P(F\mid W)P(W)\\ &\approx \arg\max_W \max_F P(X\mid F)P(F\mid W)P(W),\end{aligned}$$

ここでＦは基本単位シーケンスを示し、Ｐ（Ｘ｜Ｆ）は頑健な音響モデル化の優勢なトピックを示し、Ｐ（Ｆ｜Ｗ）は発音モデル化の注目のトピックを示し、Ｐ（Ｗ）は顕著な言語モデル化を示す。多くの場合、Ｆは音素セットである。

【００１２】

しかし、先行技術では、種々の基本単位のセットを用いた場合に関する比較についてはヒューリスティックな試みがいくつかあったものの、特に確率を用いたＡＳＲの枠組み全体を考慮して基本単位セットの最適化を行なうことはほとんど全くされていないといえる。

【００１３】

従って、この発明の目的の一つは、特定のＡＳＲタスクのための基本単位セットを最適化する方法と装置とを提供することである。

【００１４】

この発明の別の目的は特定のＡＳＲタスクのための音素セットを最適化する方法と装置とを提供することである。

【課題を解決するための手段】

【００１５】

この発明の一局面によれば、予め定められた言語の音素単位セットを最適化する方法は、コンピュータに、コンピュータ読出可能なフォーマットで基本単位セットを準備するス

テップと、基本単位セットにリーブ・ワン・アウト法を適用することによって複数個の基本単位サブセットを生成するステップと、基本単位サブセットの各々について言語的識別力の所定の尺度を計算するステップと、基本単位セットを、基本単位サブセットのうち最も高い言語的識別力を備えたもので置換えるステップと、生成するステップ、計算するステップ、及び置換えるステップを、所定の基準が満たされるまで繰返すステップとを実行させる。

【００１６】

好ましくは、計算するステップは、基本単位セットと、基本単位サブセットの各々との間の相互情報量を計算するステップを含む。

【００１７】

より好ましくは、置換えるステップは、基本単位セットを、基本単位サブセットのうち計算するステップで計算された相互情報量の最も高い値を有するもので置換えるステップを含む。

【００１８】

さらに好ましくは、基本単位セットは予め定められた言語のための基本音素セットである。

【００１９】

この発明の別の局面によれば、予め定められた言語の単位セットを最適化するシステムは、基本単位セットをコンピュータ読出可能なフォーマットで記憶するための記憶手段と、基本単位セットにリーブ・ワン・アウト法を適用することによって複数個の基本単位サブセットを生成するための生成手段と、基本単位サブセットの各々について言語的識別力の所定の尺度を計算するための計算手段と、記憶手段に記憶された基本単位セットを、最も高い言語的識別力を有する基本単位サブセットで置換えるための置換手段と、記憶手段、生成手段、計算手段及び置換手段を、所定の基準が満たされるまで繰返し動作するよう制御するための制御手段とを含む。

【発明を実施するための最良の形態】

【００２０】

ＡＳＲの場合、二つの単語を識別するのに２種類の識別のための手段がある。一つは発音であり、他方は単語の文脈、すなわち言語モデル（Ｌａｎｇｕａｇｅ　Ｍｏｄｅｌ：ＬＭ）である。一対の単語を音響スコアで識別することが困難な場合、例えば、同音語や類音語の場合、文脈的な単語情報があれば識別が容易になるであろう。例えば、「橋」と「箸」とは明らかに異なる文脈の単語である。

【００２１】

上述の議論に基づき、この実施例は特定のＡＳＲタスクのための音素セットの最適な設計、すなわちタスクに基づく音素設計を提案する。基本的な考え方は、ある大きな音素セットから１音素を削除しても言語的識別力が大きく減じられることがなければ、音素セットサイズを減じるためにその音素を削除してもよい、というものである。

【００２２】

この実施例では、最大相互情報量（Ｍｕｔｕａｌ　Ｉｎｆｏｒｍａｔｉｏｎ：ＭＩ）基準に基づく音素セット設計を採用する。すなわち、ＭＩを基本単位サブセットの言語的識別力の尺度として用いる。この実施例は中国語の最適化された音素セットを設計することに関するものである。

【００２３】

基本単位セットΦは二つの具体的な局面で重要となる。すなわち、これは音響空間全体の主たる分類を規定し、さらに、言語空間の分類の重要な手がかりを提供する。

【００２４】

図１は異なる基本単位セット、$\Phi_1=\{f_1, f_2, \cdots f_N\}$ 及び $\Phi_2=\{p_1, p_2, \cdots p_M\}$ による直観的な影響力を示す。図１を参照して、Φ_2 の音素数Ｍは Φ_1 の音素数Ｎよりはるかに大きいと仮定する（すなわち、Ｎ＜＜Ｍ）。Φ_1 は音響空間２０をＮ個のサブスペース $f_1, f_2, \cdots f_N$ に分割し、Φ_2 は同じ音響空間２２をより小さいサブ

スペースｐ$_1$，ｐ$_2$，…ｐ$_M$に分割する。従って、Φ$_1$は頑健な音響モデルを提供することができるが、その一方で、識別力はΦ$_2$のそれに比して弱い。

【００２５】

図２はこの実施例の単位セットのトレーニングのための構成全体を示す。図２を参照して、トレーニングシステムは、トレーニング用の最新のＡＳＲシステム４０と、言語モデルのための記憶部４２と、レキシコンベースのデコードシステム４４とを含む。

【００２６】

トレーニング用ＡＳＲシステム４０は、入力されたテキストWを音素シーケンスＦに変換するための音声生成及びＡＳＲモジュール５０と、音素シーケンスＦによって形成される単語ラティス内のデコードされた単語テキストのうちで最も確率の高い単語テキスト＾Wを、言語モデル４２を参照しつつラティスの各経路をスコアリングすることによって選択するための単語ラティススコアリングモジュール５２とを含む。

【００２７】

レキシコンベースのデコードシステム４４は、見出し語の各々を、それぞれの音素セットΦ$_1$及びΦ$_2$を用いて記述する辞書６２及び６４と、辞書６２及び６４をそれぞれ用いて、入力テキストWを音素シーケンスＦ$_1$及びＦ$_2$に変換するためのレキシコンベースの変換モジュール６０と、音素シーケンスＦ$_1$及びＦ$_2$によって形成される単語ラティス内の単語テキストのうちで最も確率の高い単語テキストＷ$_1$及びＷ$_2$を、言語モデル４２を参照しつつラティスの各経路をスコアリングすることによって選択するための単語ラティススコアリングモジュール６６とを含む。図２では説明を簡潔にするため、二つの辞書のみを示す。この実施例は中国語のＡＳＲシステムに関し、音素セットΦ$_1$は声調情報を含み、一方音素セットΦ$_2$はこれを含まない。

【００２８】

トレーニング用ＡＳＲシステム４０はトレーニングテキストWのコーパスを受け、以下の最大化式に従って、デコードされた単語＾Wを出力する。

【００２９】

【数３】

$$\hat{W} = \arg\max_{W} \max_{F} P_A(X \mid F) P_P(F \mid W) P_L(W).$$

確率Ｐ（Ｗ｜Ｆ）を最大にする音素セットが最適な音素セット＾Φとして選択される。すなわち、

【００３０】

【数４】

$$\hat{\Phi} = \arg\max_{\Phi} P(W \mid F)$$

$$= \arg\max_{\Phi} \frac{P(F \mid W)P(W)}{P(F)}$$

$$= \arg\max_{\Phi} \frac{P(F \mid W)P(W)}{\sum_{W_i} p(F \mid W_i)\, p(W_i)}$$

トレーニング用ＡＳＲシステム４０とレキシコンベースのデコードシステム４４との動作により、上述の式に従って、Ｐ（Ｗ｜Ｆ）の要素を計算し、最適な音素セット＾Φを選択することができる。

【００３１】

図３はこの実施例の音素セット最適化システム８０の全体構造を示す図である。図３を

参照して、音素セット最適化システム８０は、基本単位セット９０の記憶装置と、トレーニングテキスト９２の記憶装置と、基本単位セット９０及びトレーニングテキスト９２を用いて音素セットを最適化し、最適化された音素セット９４を出力するための音素セット最適化モジュール９６とを含む。

【００３２】

音素セット最適化モジュール９６は、コンピュータ上で実行されるソフトウェアで実現可能である。ソフトウェアの制御の流れを図４のフロー図で示す。図４を参照して、音素セット最適化モジュール９６は以下のステップを実行する。初期音素セットΦ_0（すなわち基本単位セット９０）で作業中の音素セットΦを置換える（ステップ１００）。音素サブセットΦ_i（ｉ＝１からΦの要素数まで；$\Phi_i=\Phi-\{e_i\}$；e_iはΦ中のｉ番目の音素）を生成する（ステップ１０２）。作業中のセットΦとサブセットΦ_iの各々との間の相互情報量MI_iを計算する（ステップ１０４）。以下の式を満たす指数Ｍを特定する（ステップ１０６）。

【００３３】

【数５】

$$M=\arg\max_i MI_i;$$

その後Ｍ番目の音素サブセットΦ_Mを選択し、選択されたサブセットΦ_M中の音素を用いてレキシコン及びテキストコーパスを作り変える（ステップ１０８）。作り変える過程において、レキシコンとテキストコーパスとは、レキシコンとテキストコーパス中で用いられている削除された音素を、それぞれ最も近い音素とマージするように更新される。

【００３４】

音素セット最適化モジュール９６はさらに、予め定められた停止条件が満たされたか否かを判断するステップを実行する（ステップ１１０）。もし条件が満たされれば、音素セット最適化モジュール９６は動作を停止する。さもなければ、制御はステップ１１２に進み、ここで選択されたサブセットΦ_Mで作業中のセットΦを置換え、その後制御はステップ１０２に戻る。

【００３５】

予め定められた数だけ繰返したあと、動作は停止する。これに代えて、相互情報量の減少が予め定められたしきい値を超えた場合に動作を停止することもできる。

【００３６】

音素セット最適化モジュール９６は以下のように動作する。始めに、ステップ１００で、基本単位セット９０が作業用セットΦとして選択される。ステップ１０２で音素サブセットΦ_1からΦ_Nまでが生成される。サブセットΦ_iは作業中のセットΦから音素e_iを除くことで生成される。言換えれば、Φ_iは作業中のセットΦにリーブ・ワン・アウト法を適用することによって生成される。

【００３７】

ステップ１０４で、作業中のセットΦとサブセットΦ_1からΦ_Nの各々との間の相互情報量MI_iが計算される。ステップ１０６で、相互情報量MI_i中で対応の相互情報量MI_Mを最大にする指数Ｍが選択される。

【００３８】

ステップ１０８で、Ｍ番目の音素サブセット（サブセットΦ_M）が選択され、選択された音素サブセットΦ_Mを用いてレキシコンとテキストコーパスとが作り変えられる。

【００３９】

ステップ１１０で、停止条件が満たされたか否かが判断される。もし条件が満たされていなければ、制御はステップ１１２に進み、ここでΦがΦ_Mと置換される。その後、制御はステップ１０２に戻り、ステップ１０２から１０８までが繰返される。停止条件が満たされると、動作は停止する。

【００４０】

こうして、詳細な音素分類に基づいたものであってかつサイズの大きい初期単位セット

９０から始めて、音素セット最適化モジュール９６は何らかの基準に従って繰返しながら音素セットを減じることができる。

【００４１】

図５はこの実施例の検証実験の結果を示す。この実験では、声調情報を含む元の２０３単位からなるセットを減少させる。声調情報を含まない５９単位のセットを比較のために用いた。これら二つのセットは最新の中国語ＡＳＲシステムで広く用いられているものである。検証用テキストコーパスは１，６１４個の短文を含み、単語数は合計で９，４８４個である。

【００４２】

図５を参照して、５９の声調なしの単位セットＣ（ボックス１３２で示す）と比較して、元の２０３の声調付きセットは、ビット表現でより高い相互情報量を有する。線１３０で示す削減過程で、同じ５９単位の数を備えて生成された単位セットは、図５の点Ａで示すように、声調なしの単位セットに比べてより高い相互情報量を維持した。言換えれば、生成されたセットＡは、数が同じであるにもかかわらず、伝統的な５９の声調なし単位セットよりも良好な言語的識別力を有する。図５の点Ｂの単位セットは、声調なし単位セットＣとほぼ同じ量の相互情報量を維持しているが、単位数は遥かに少ない。数は４７であり、従ってこれはＣセットより効率が良い。

【００４３】

上述の通り、この実施例のシステムと方法とは、相互情報量を減じることなく、音素セット中の音素の数をうまく減じることができる。タスクを特定したテキストをトレーニングに用いれば、音素セットはそのタスク用に最適化でき、その音素セットを用いれば、そのタスクについて十分な識別力を有する頑健な音響モデルを得ることができる。また、十分に詳細な識別力を有する言語モデルを提供できる。

【００４４】

上述の実施例では音素セットを最適化したが、この発明は音素セットの最適化に限定されない。この発明は、ＡＳＲにおいて音素セットに置換可能ないずれの基本音素単位セットの最適化にも適用可能である。例えば、語彙が比較的小さい場合には、単位セットは語彙中の単語（単語発音）のセットであり得る。

【００４５】

今回開示された実施の形態は単に例示であって、本発明が上記した実施の形態のみに制限されるわけではない。本発明の範囲は、発明の詳細な説明の記載を参酌した上で、特許請求の範囲の各請求項によって示され、そこに記載された文言と均等の意味および範囲内でのすべての変更を含む。

【図面の簡単な説明】

【００４６】

【図１】異なる基本単位セットからの直観的な影響力を示す図である。

【図２】この実施例の単位セットのトレーニングの全体構成を示す図である。

【図３】この実施例の音素セット最適化システム８０の全体構造を例示する図である。

【図４】この実施例の音素セット最適化モジュール９６を実現するソフトウェアの制御フローを示す図である。

【図５】この実施例の検証実験結果をグラフの形で示す図である。

【図６】先行技術による基本ＡＳＲスキームを示す図である。

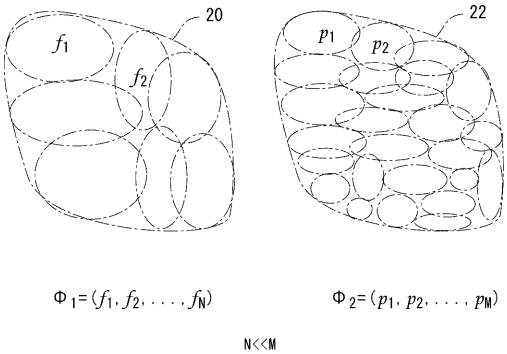
【符号の説明】

【００４７】

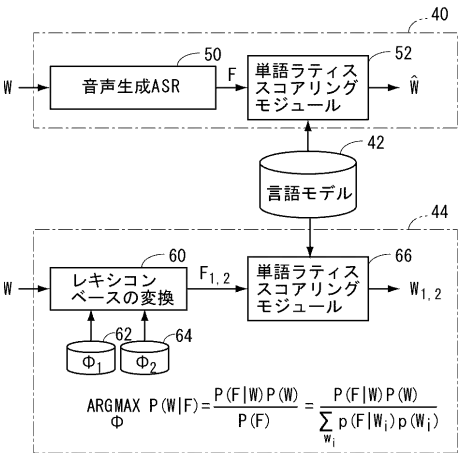
４０　トレーニング用ＡＳＲシステム
４２　言語モデル
４４　レキシコンベースのデコードシステム
５０　ＡＳＲモジュール
５２　単語ラティススコアリングモジュール
６０　レキシコンベースの変換モジュール

６２、６４　辞書
６６　単語ラティススコアリングモジュール
８０　音素セット最適化システム
９０　基本単位セット
９２　トレーニングテキスト
９４　最適化音素セット
９６　音素セット最適化モジュール

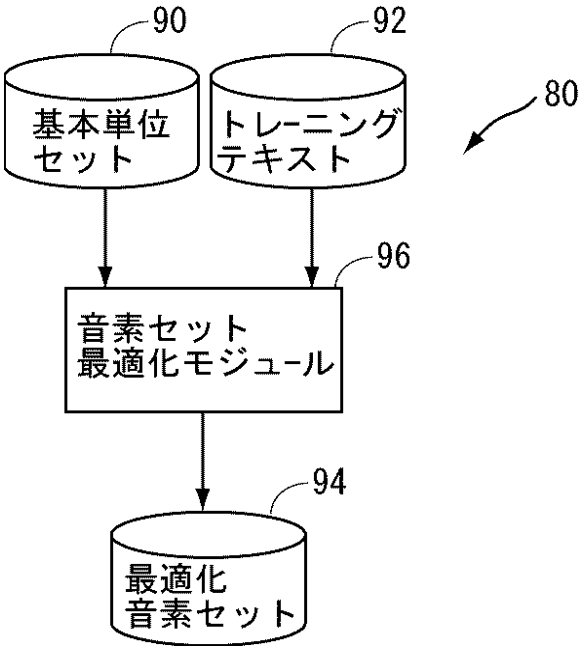
【図１】



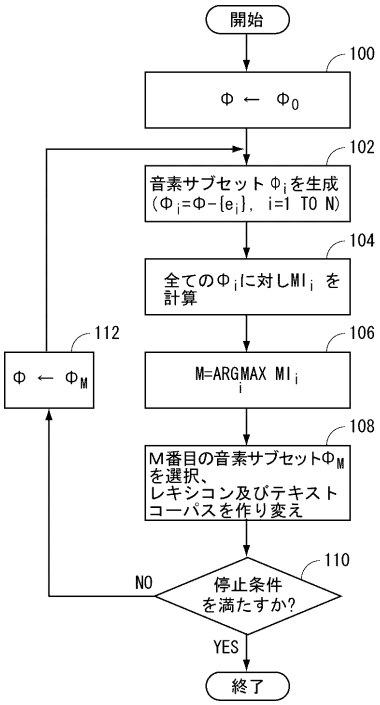
【図２】



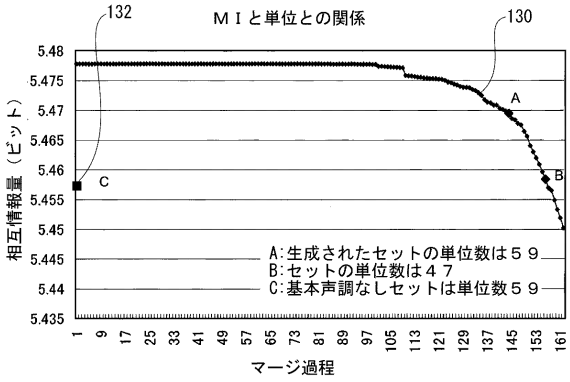
【図３】



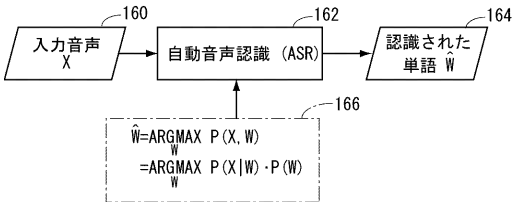
【図４】


開始
100
Φ ← Φ0
102
音素サブセットΦiを生成
(Φi=Φ-{ei}, i=1 TO N)
104
全てのΦiに対しMIi を
計算
106
M=ARGMAX MIi
i
108
M番目の音素サブセットΦM
を選択、
レキシコン及びテキスト
コーパスを作り変え
110
停止条件
を満たすか?
NO
YES
112
Φ ← ΦM
終了

【図５】


ＭＩと単位との関係
132
130
A
B
C
相互情報量（ビット）
5.48
5.475
5.47
5.465
5.46
5.455
5.45
5.445
5.44
5.435
1
9
17
25
33
41
49
57
65
73
81
89
97
105
113
121
129
137
145
153
161
マージ過程
A:生成されたセットの単位数は５９
B:セットの単位数は４７
C:基本声調なしセットは単位数５９

【図６】


160
入力音声
X
162
自動音声認識 (ASR)
164
認識された
単語 Ŵ
166
Ŵ=ARGMAX P(X,W)
W
=ARGMAX P(X|W)·P(W)
W

フロントページの続き

審査官　山下　剛史

(56)参考文献　特表平１０－５０１０７８（ＪＰ，Ａ）
特開平０９－２８８４９２（ＪＰ，Ａ）
特開平０４－２９５８９３（ＪＰ，Ａ）
特開平１１－３５２９９４（ＪＰ，Ａ）
特開平１１－２７２２９１（ＪＰ，Ａ）
上田　修功，最小分類誤り基準に基づくニューラルネットワーク識別機の最適線形統合法，電子情報通信学会論文誌，日本，電子情報通信学会，１９９９年　3月２５日，Ｖｏｌ．Ｊ８２－Ｄ－ＩＩ　Ｎｏ．３，ｐ．５２２－５３０

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１５／００－１７／００
ＪＳＴＰｌｕｓ／ＪＭＥＤＰｌｕｓ／ＪＳＴ７５８０（ＪＤｒｅａｍＩＩ）