修正 補充
年 月 日

申請日期	86. 3. 10
案　　號	86102943
類　　別	G10L 5/02

公告本

A4 434528
C4

中文說明書修正頁(88年2月)

（以上各欄由本局填註）

發明 新型 專利說明書

一、發明新型名稱	中　文	基於分音節成寬音階，限制序列向量量化及隱藏馬可夫模型，自動將語音分節成用於語音處理應用的音素單元的方法及裝置
	英　文	~~A METHOD AND APPARATUS FOR AUTOMATIC SPEECH SEGMENTATION~~ INTO PHONEME-LIKE UNITS FOR USE IN SPEECH PROCESSING APPLICATIONS, AND BASED ON SEGMENTATIN INTO BROAD PHONETIC CLASSES, SEQUENCE-CONSTRAINED VECTOR QUANTIZATION, AND HIDDEN-MARKOV-MODELS
二、發明創作人	姓　名	1. 史帝芬 克拉倫斯 保斯 2. 伊文斯 吉曼尼 查理 坎普 3. 李昂那杜斯 法蘭西庫斯 威廉
	國　籍	1. 3. 荷蘭　2. 比利時
	住、居所	1. 荷蘭恩特荷芬市米蘭克街28號 2. 比利時李瑪市瑪西奧路20號 3. 荷蘭恩特荷芬市李尼吉哈夫區47號
三、申請人	姓　名（名稱）	荷蘭商皇家飛利浦電子股份有限公司
	國　籍	荷蘭
	住、居所（事務所）	荷蘭愛因和文市格羅尼渥街1號
	代表人姓　名	J．G．A．羅夫斯

434528

（由本局填寫）

承辦人代碼：	
大　　類：	
IPC分類：	

修正補充 88年1月26日

A6
B6

本案已向：

國（地區）	申請專利，申請日期：	案號：	，□有　□無主張優先權
歐盟	1996.2.27	96200509.6	

有關微生物已寄存於：　　，寄存日期：　　，寄存號碼：

（請先閱讀背面之注意事項再填寫本頁各欄）

裝　訂　線

五、發明說明（1）

發明背景

本發明是關於將語音自動分節以用於語音處理應用之方法。在各種可能的應用當中，一特定之應用是語音合成，特別是以雙音(diphones)串連的語音合成。雙音是短的語音分節，各分節主要包括二相鄰音素之間的一個轉換，再分別加上前一音素之最後部分及其接續音素的最初部分。雙音可根據若干本身已知之規則從已分節成單一音素的一資料庫中摘錄出來。通常，該資料庫包括一控制之下環境中，從一特定單一說話者錄製而得之分離字，而且亦包括經驗證拼音字(phonetic transcription)與聲音實現(acoustic realization)之間的相似性。1992年於加拿大Alberta省Banff市舉行的國際語音及語言處理會議中，O. Boëfard等人發表Automatic Generation of Optimized Unit Dictionaries for Text to Speech Synthesis（自動轉換文字爲語音合成之最佳辭典）一文，其序文與第1211-1215頁中即揭示一根據音素隱藏馬可夫模型(HMM)的簡明而可自動實現分節的方法。然而，如今已發現該已知方法的品質不足，因爲該方法所發現之界限一般而言過於偏離以一手動程序放置的相關界限位置。當然，如該等音素隱藏馬可夫模型(HMMs)在開始時，以一分離及手動分節之資料庫校對時，則可改善分節的準確度。但該手動分節資料庫的設定成本經常過高，因爲在一語音合成系統中需要一個新的說話者時，每次均須重複設定。因此，本發明最主要的目的之一在提出一項育於語音分節的方法，該方法爲完全自動、毋須手動分節語音資料，而且可提供比所提及之

五、發明說明（2）

方法爲佳的成果。

發明概述

根據本發明一種現象，本發明提供了一種以供語音處理應用使用的自動分節語音的方法，該方法包括下述步驟：

• 將一語音資料庫的發音予以分類並分節成發聲(voiced)、不發聲(unvoiced)及無聲(silence)三類寬音階(BPC)，用以獲得初步分節的位置；

• 利用初步的分節位置當做爲定點，俾在一SCVQ(限制序列向量量化)步驟中以限制序列向量量化進一步分節成類似音素之單元；

• 以該SCVQ(限制序列向量量化)步驟所提供的分節標注音素隱藏馬可夫模型，並進一步以巴姆-韋爾契估算法(Baum-Welch estimation)調整該等隱藏馬可夫模型(HMM)之參數；

• 最後，利用該等經過充分校準的隱藏馬可夫模型執行Viterb之發音錄寫的校整，而以此方法獲得該等最後的分節點。

所引用之該方法有一額外優點，即僅需要如發音錄寫中主要因素最少的啓始資料。特別是，不需要分別以手動方式分節資料庫以供估算該等隱藏馬可夫模型(HMM)參數。

有利的是，在校準之後可建立一雙音組以供進一步之使用，例如用於語音合成中。本發明已提供一種簡明及便宜的多數說話者之系統。

本發明亦有關於一於語音處理應用中使用之語音分節的裝置，該裝置包括：

五、發明說明（3）

• 由一語音資料庫所提供之寬音階(BPC)分節方法，其用以將所收到的語調予以分類及分節成發聲(voiced)、不發聲(unvoiced)及無聲(silence)等三類寬音階(BPC)，

• 由該寬音階(BPC)分節方法所提供之SCVQ(限制序列向量量化)分節方法，其利用初步的分節位置作爲固定點而藉由限制序列向量量化(SCVQ)執行進一步做類似音素單元之分節工作，

• 由該SCVQ(限制序列向量量化)分節方法所提供之語音隱藏馬可夫方法(HMM)，其用以開始音素隱藏馬可夫模型(HMM)及進一步調整隱藏馬可夫方法(HMM)之參數；

• 由該隱藏馬可夫方法(HMM)所控制之最後分節方法。

該種裝置將可使未經訓練之人員在短時間內用以訓練隨便一個新的說話者使用之。本發明進一步之優點詳述於隨後所申請專利範圍中。

圖式簡單說明

本發明上述及其它目標與優點將於揭示該較佳具體實施例時予以討論，尤其於所附圖示中更爲詳細，該等圖示顯示：

圖1：該裝置的整體方塊圖；

圖2：某一無意義字的五項測量數值；

圖3：一第一寬音階分節；

圖4：以向量量化之音素分節；以及

圖5：相同，但已經以HMM(隱藏馬可夫模型)分節予以改良。

五、發明說明（4）

較佳整體具體實施例之詳細說明

本發明意欲將一資料庫的各語調切割成一序列非重疊相鄰之音素分節，並形成該等分節與該音標拼字所提供的一序列音階標籤之間的一對一的對應。語音可適當地以一序列之聲音向量來說明，各聲音向量一致在通常爲10-20毫秒(ms)及每次轉變的2.5-10毫秒(ms)的時間內凸顯該語音之特性。任一時間t的p-次元聲音向量爲$o(t)=[o_1(t)\ldots o_p(t)]'$，而顯示向量轉換的聲調及一完整的時間系列則以$O(1,T)=o(1),\ldots o(T)$表示之。在本發明之具體實施例中，寬音階(BPC)分節之$p=5$、限制序列向量量化之$p=12$，而隱藏馬可夫模型之$p=51$。一寬音階元素或跨一分節l之類似音素單元可用一原型向量或以$\hat{c}_l$表示的質量中心表示之。分節l之分節內部距離$d_l(i,j)$可定界爲該跨越該分節之向量O(i,j)與該質量中心$\hat{c}_l$：$d_l(i,j)=\sum_{t=i}^{j} d(o(t),\hat{c}_l)$之間距離的總和。將變形：

$$\sum_{l=1}^{L} d_l(b_{l-1}+1,b_l)=\sum_{l=1}^{L}\ \sum_{t=b_{l-1}+1}^{b_l} d(o(t),\hat{c}_l)$$

減至最小後，其分節點b_l($l=1\ldots L-1$)。該一般公式此後用於測量距離、程序減少，以及用於決定該等質量中心。

圖1顯示一系統具體實施例之整體方塊圖。此處，項目20是一麥克風以用於接收來自一特定說話者的語音，該說話者受命說出該組預定之獨立字，各獨立字必須分節。該等字組已提議用於各種語言中。某些或所有該等字可能是無意義的字。方塊22中之第一階處理程序是定期取樣、數位化及過濾。其結果是儲存於中間記憶體24以用於從該處理程序中切斷語音接收。一般控制是位於方塊34，該方塊34

五、發明說明（5）

可爲一以適當方式經程式設計過的標準電腦，且該方塊34控制方塊24、31、36。該控制暗含該應用程式的同步化、供應，或還包括各處理步驟間的中間儲存器之同步化、供應。爲簡化起見，僅繪出一單一的單方向連接。接著，讀出記憶體24的發音之後，在方塊26中著手進行第一階段的分階及分節成發聲(VOIced)、不發聲(UNVoiced)及無聲(SILence)三類寬音階(BPC)。該初步寬音階(BPC)分節的結果進而在方塊28中予以處理，在該方塊28中執行第二階段之限制序列向量量化(SCVQ)以進一步分節成音素。方塊30實行第三階段。該方塊30利用先前SCVQ階段所傳來的分節以設定該等音素隱藏馬可夫模型(HMMs)之初值，然後藉由在欲分節的資料庫上執行巴姆-韋爾契估算法(Baum - Welch estimation)以進一步調整該等音素隱藏馬可夫模型(HMMs)。有關極佳及可得之隱藏馬可夫模型(HMMs)的論述與巴姆-韋爾契估算法(Baum - Welch)，請參閱L.R. Rabiner的文章'A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition'，該文發表於1989年2月之IEEE(電機及電子工程師學會)會報，第77冊，第2號，第257 - 286頁。最後，在方塊31中利用該等隱藏馬可夫模型(HMMs)以執行該等資料庫發音及其音標拼字之間的維特比暫時對準(Viterbi temporal alignment)動作。維特比對準(Viterbi alignment)本身是標準技術，請參考同一文件。

分節的結果是儲存於記憶體32中。雙音產生裝置39從每一發音所產生的序列音素中設立產生實際語音所需之各種雙音，該等雙音被寫回記憶體中當作一資料庫以用於依實

五、發明說明（6）

際從說話者收到的有限數量之語音為基礎而產生可能具有任意內容的語音，或用於其他語音處理應用。此外該控制是衍生自(觀念上的)控制／校正系統36，該系統36經由連線42而獲得該資料庫，經由連線40而接收該等雙音，且可進一步經由雙向連線38而與記憶體32進行互動。為簡潔起見，未顯示經由連線41之實際語音輸出裝置。該語音處理之流程圖一般是以相似方式設定，因為各個分別說出的字是在取得下一字之前先處理完全。

第一階段：分節成寬音階

第一階段分節應提供所謂固定點以用於後續階段。該等三類音階是未存在有任何語音波形的無聲(SIL)、語音波形為半週期性的發聲(VOI)及波形為非週期性或隨機之不發聲(UNV)。在該具體實施例中，分節是以下列五項不同的測量為基礎：

－ E_N：1-(正規化的短時間能量)，其中該正規化係使得無聲(SIL)具有趨近於1之 E_N；

－ E_{low}：(正規化的低頻能量)，範圍為50 - 1200 Hz；

－ E_{high}：(正規化的高頻能量)，範圍為2000 - 4000 Hz；

－ 零交越率(zero - crossing rate) Z_N；

－ 一次方LPC模型的第一LPC係數a1。

在該具體實施例中，該取樣波形x(k), k = 0 . . . N - 1已先藉由一過濾函數 $(1 - 0.95z^{-1})$ 施以頻應預矯(pre - emphasized)、封鎖並以漢明框選法(Hamming - windowed)分成20毫秒(ms)的時間格(frames)，而整個時間格位移2.5毫秒(ms)。圖2顯示一以荷蘭

五、發明說明（7）

語發音之無意義字'kekakke'的五項前述測量。緊接於該波形之下標示有該等相關音素記號。如圖所示，母音'A'的主要週期約爲10毫秒(msec)。該圖中可清楚地看到三個發聲(voiced)間隔、三個不發聲(unvoiced)間隔及四個無聲(silence)間隔。該五項前述測量的結果是以個別線圖顯示。某些轉變比其他更清晰可見。實際上，連續音素之間的轉變不一定暗示寬音階有改變。

接下來，首先以寬音階方式將該等發音之可用音標拼字對映至一譯音(transcription)；事實上，系統知道實際應已收到何字，因爲該說話者必須以規定順序輸入該等字。假設某一發音存在於一序列L'寬音階元件中，使用者必須在該觀察O_l^T中找出對應的連續分節與把定義於上之總距離減至最小的一組質量中心$\{\hat{c}_l\}$。此處所使用的距離是加權之歐氏距離(weighted Euclidean distance)，而該等測量的估算差異被用於當作加權因數。該最小化工作是在十分類似於下列極爲知名之階層建立法(level - building approach)的一動態程式設計結構中執行，請參照C.S. Myers及L.R. Rabiner之文章A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition，發表於1981年4月ASSP(美國語音處理協會)的IEEE(電機及電子工程師學會)會報，第29冊(Vol. 29)，編號2(No. 2)，第284 - 297頁。此處所使用之方法基本上是反覆性的且一直持續直到無法達成進一步改良。該程序之各次反覆均包括連續步驟。第一步針對一組已知的質量中心而找出一組最佳界線b_l。第二步利用最新獲取之分節點將該等

五、發明說明（8）

質量中心{$\hat{c}_l$}及該等差異予以更新。由於此處使用加權之歐氏距離，代表一寬音階的最新質量中心即等於該組已知分節點之全部觀察向量的數學平均。第一步從單元差異及代表下列理想寬音階的質量中心開始進行：

$\hat{c}_{SIL}=(1,0,0,1,1)'$、$\hat{c}_{UNV}=(0,0,1,1,0)'$、$\hat{c}_{VOI}=(0,1,0,0,1)'$

一寬音階元件的最小、平均及最大歷時時間相等於組成一寬音階分節之音素的最小、平均與最大歷時時間和(sum)。有關該等音素之最小及最大歷時時間可從文獻中得知。該等最小及最大歷時時間資訊是用於提供一寬音階(BPC)分節可存在的時間間隔上限與下限以限制對於寬音階(BPC)界線之搜尋。最後，平均歷時時間是用於按該發音的實際長度比例來縮放該等最小及最大歷時時間。

圖3顯示第一寬音階分節結果，而以與手工決定之轉變點相差小於一特定差的累積量來表示。如圖所示，因爲其差異小於20毫秒(msec)，故82%的轉變令人滿意地局限於一局部區域。

第二階段：限制序列向量量化(SCVQ)

上述分節產生固定點以用於進一步處理語音。該處理程序一般會縮減爲將各寬音階元件分節成其組成音素單元。更特別的是，該等發音被分節成連續性擬似無變化之元件。由於要求必須無變化，故雙母音及破裂音進一步劃分成其組成音素。再者，必須以類似於前一階段所使用之反覆方式找出將總體扭曲減至最小之分節界線與質量中心。對於第一步反覆步驟而言，該質量中心是定義成透過所有

五、發明說明（9）

附帶相同標籤並聚積成一單一觀察序列 $O_{b_{l-1}+1}^{b_l} = o_{(b_{l-1}+1)}, \ldots o(b_l)$ 的假設將扭曲減至最小之向量。

$$\hat{c}_l = arg_{c_l} \min \frac{1}{b_l - b_{l-1}} \sum_{t=b_{l-1}+1}^{b_l} d(o(t), c_l)$$

質量中心是以該方式從手頭上的觀察中獲得，其明顯的優點在於事先不需任何知識。第一次限制序列向量量化(SCVQ)反覆是藉由動態階層建立程式設計(level - building dynamic programming)來完成。第二次反覆之方案相同於早先的寬音階(BPC)分節階段者。第一次反覆所發現之音素質量中心被使用於下一步驟中當作參考，使程序縮減爲樣板比對。第三及最後一步驟針對產生之各標籤而計算質量中心。進一步的反覆不會再產生任何改善。圖4顯示利用由第一階段所產生之固定點而進行的音素分節。完全不同於該寬音階(BPC)階段的是，此處該等標籤大約與50個不同音階相關。該分節在考量如此眾多音階之下必然比第一階段之分節更易爲人所接受：在全部案例當中，有70%的案例與手工分節之誤差小於20毫秒(msec)。

第三階段：以隱藏馬可夫模型(HMM)進行分節

使用隱藏馬可夫模型可更進一步藉由闡述統計學上的語音變化性而改善限制序列向量量化(SCVQ)所獲致之結果。再者，總共約使用50個不同的音素單元，各自獲得其自己的隱藏馬可夫模型(HMM)。爲確保據實進行完全自動程序，故以該限制序列向量量化(SCVQ)階段所獲致之結果設定該等隱藏馬可夫模型(HMMs)的初值。除該脈衝束(burst)單元以外，各音素單元均獲得一簡單的由左至右而含有六種

五、發明說明（10）

狀態之隱藏馬可夫模型(HMM)拓樸，且包括自我迴路及直接往下一階段的轉變；因此，歷時時間至少爲六個時間格，各時間格爲5毫秒(msec)，而該等時間格較短於早先考量者。一擬似脈衝束的分節是以具有同類自我迴路及轉變之2-階段模型來代表。

放射分佈(emission distributions)屬於連續類型。各觀察向量被分割成4子向量，分別代表

－瞬間特性(16波道的過濾束分析)；

－比對前一時間格而決定的該等特性之第一及第二有限的差異；

－能量數據。

各子向量之放射機率密度是一多變常態混合分配(multivariate Gaussian mixture distribution)。

該等隱藏馬可夫模型(HMM)參數是按以下方式調整。請考慮一特定音素單元及在該限制序列向量量化(SCVQ)階段爲此切割出的該組觀察序列。該音素單元之隱藏馬可夫模型(HMM)的初值設定是藉由將各觀察序列之聲音向量均勻分散於該單元的6(或2)種隱藏馬可夫模型(HMM)狀態之間來完成。結果，一觀察的特定聲音向量以均勻分散於該等狀態之中的方式分派給該模型之各狀態。接下來使用一k-平均演算法(k-means algorithm)將該等聲音向量分割成叢集(clusters)，從該等叢集中可計算出該混合分配的參數之初值。亦請見J.G. Wilpon及L.R. Rabiner的文章A Modified K-Means Clustering Algorithm for Use in Isolated Word Recognition，發表於1985

五、發明說明 (11)

年6月ASSP(美國語音處理協會)的IEEE(電機及電子工程師學會)會報，第33冊(Vol. 33)，編號3 (No. 3)，第587 - 594頁。

該等參數是平均向量、該等組成密度的斜度矩陣(covariance matrices)及混合加權(mixture weights)。該等轉變機率的初值被設定爲非0之任意值。一旦獲得該等初始隱藏馬可夫模型(HMMs)，即施以一受監督之維特比校準(Viterbi training)，結果造成一新的分節。從該處可計算出最新的隱藏馬可夫模型(HMM)參數。後者包括利用從該比對路徑收集而得的狀態與狀態之間的轉變統計數字計算出之轉變機率。接下來，針對各完整發音，依照巴姆-韋爾契估算法(Baum - Welch)進行參數估算，藉以調整該等模型。

最後，該等完全經過校正之隱藏馬可夫模型(HMMs)被用以執行該資料庫中的各字及其音標拼字之間的維特比對準(Viterbi alignment)動作。此舉產生最後的分節點。圖5顯示該結果：其改善效果相對於圖4之限制序列向量量化(SCVQ)分節而言是驚人的：接近90%的轉變具有小於20毫秒(msec)之誤差。

434528

A5
B5

四、中文發明摘要（發明之名稱：基於分音節成寬音階，限制序列向量量化及隱藏馬可夫模型，自動將語音分節成用於語音處理應用的音素單元的方法及裝置）

爲進行機器語音分節，首先將一已知語音字串資料庫中的發音予以分類及分節成發聲(voiced)、不發聲(unvoiced)及無聲(silence)三類寬音階(BPC)。其次，利用初步的分節位置當作固定點，再以限制序列向量量化進一步分節成音素單元。最後，藉由隱藏馬可夫模型精準調整爲該等分節音素，且在校準之後組成一雙音可供進一步之使用。

英文發明摘要（發明之名稱：A METHOD AND APPARATUS FOR AUTOMATIC SPEECH SEGMENTATION INTO PHONEME-LIKE UNITS FOR USE IN SPEECH PROCESSING APPLICATIONS, AND BASED ON SEGMENTATIN INTO BROAD PHONETIC CLASSES, SEQUENCE-CONSTRAINED VECTOR QUANTIZATION, AND HIDDEN-MARKOV-MODELS）

For machine segmenting of speech, first utterances from a database of known spoken words are classified and segmented into three broad phonetic classes (BPC) voiced, unvoiced, and silence. Next, using preliminary segmentation positions as anchor points, sequence-constrained vector quantization is used for further segmentation into phoneme-like units. Finally, exact tuning to the segmented phonemes is done through Hidden-Markov Modelling and after training a diphone set is composed for further usage.

（請先閱讀背面之注意事項再填寫本頁各欄）

裝 訂 線

A8
B8
C8
D8

六、申請專利範圍

1. 一種自動將語音分節俾以用於語音處理應用之方法，其包括下述步驟：

• 將一語音資料庫的發音予以分類及分節成發聲(voiced)、不發聲(unvoiced)及無聲(silence)三類寬音階(BPC)，以便獲得初步的分節位置；

• 利用初步的分節位置作爲固定點，在一SCVQ(限制序列向量量化)步驟中再以限制序列向量量化進一步分節成音素單元；

• 以該SCVQ(限制序列向量量化)步驟所提供的分節設定音素隱藏馬可夫模型之初值，並進一步藉由巴姆-韋爾契估算法(Baum - Welch estimation)調整該等隱藏馬可夫模型(HMM)參數；

• 最後，利用該等完全經過校正的隱藏馬可夫模型以執行該等發音及其音標拼字之間的維特比對準(Viterbi alignment)動作，且以該方式獲得最後之分節點。

2. 如申請專利範圍第1項之方法，在該分節之後組成一組雙音以供進一步使用。

3. 如申請專利範圍第1項或第2項之方法，其中該語音處理是語音合成。

4. 一種將語音分節以用於語音處理應用之裝置，其包括：

• 由一語音資料庫所提供之寬音階(BPC)分節方法，用以將收到的發音予以分類並分節成發聲(voiced)、不發聲(unvoiced)及無聲(silence)三類寬音階(BPC)，

• 由該寬音階(BPC)分節方法所提供之SCVQ(限制序列向

六、申請專利範圍

量量化)分節方法，用以利用初步的分節位置作爲固定點以藉由限制序列向量量化(SCVQ)執行進一步分節成音素單元，

• 由該SCVQ(限制序列向量量化)分節方法所提供之語音隱藏馬可夫方法(HMM)，用以設定音素隱藏馬可夫模型(HMM)的初值，並進一步調整隱藏馬可夫模型(HMM)參數；

• 由該隱藏馬可夫方法(HMM)所控制的最後分節方法。

5. 如申請專利範圍第4項之裝置，其包括該分節方法所提供之雙音產生方法以用於組成一組雙音。

6. 如申請專利範圍第4項或第5項之裝置，其尚包括一輸出控制階段以用於透過一介於該調整方法與該語音合成輸出階段之間的中間儲存階段來控制一語音合成輸出階段。

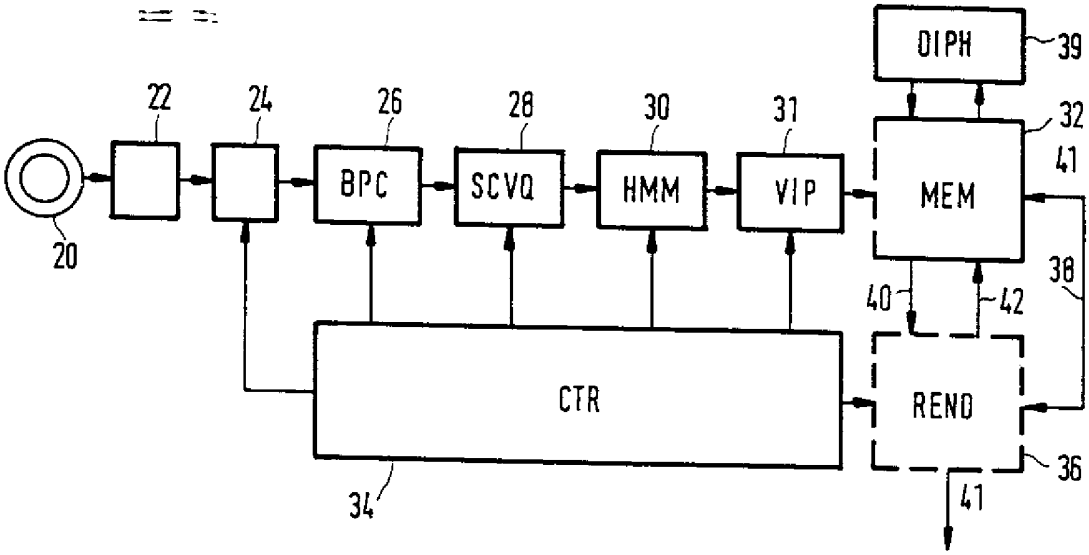
DIPH
39
22
24
26
28
30
31
32
41
BPC
SCVQ
HMM
VIP
MEM
20
38
40
42
CTR
REND
36
34
41

圖 1

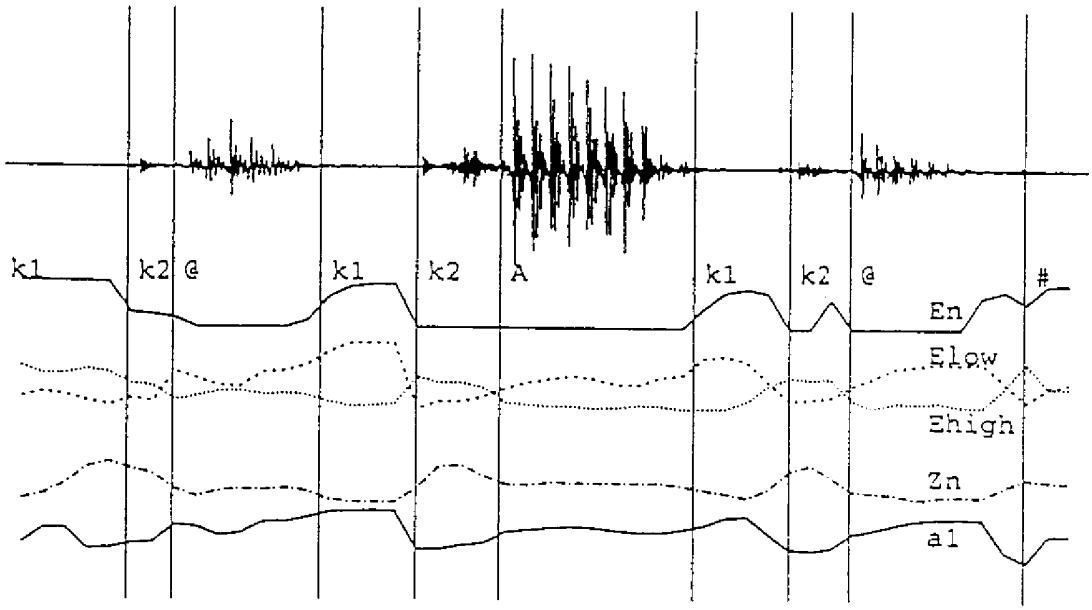
k1
k2
@
k1
k2
A
k1
k2
@
#
En
Elow
Ehigh
Zn
a1

圖 2

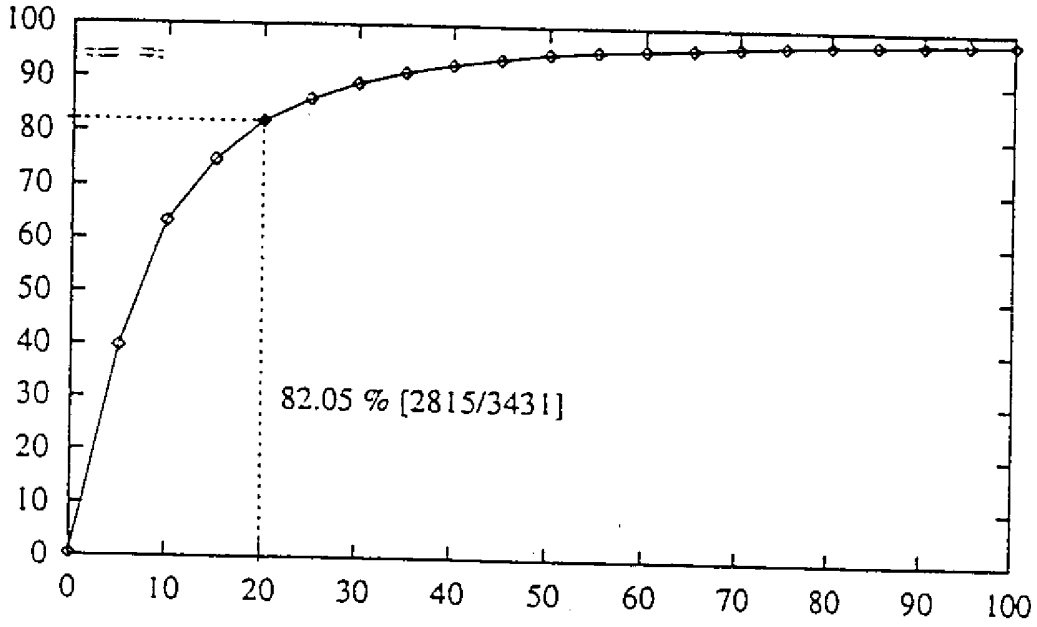
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
82.05 % [2815/3431]

圖 3

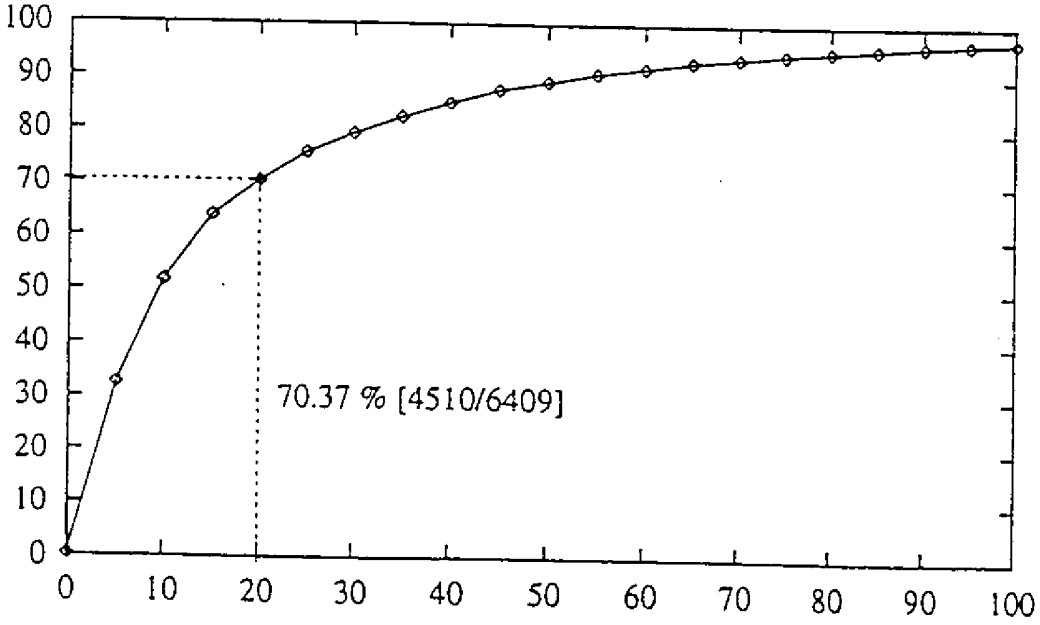
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
70.37 % [4510/6409]

圖 4

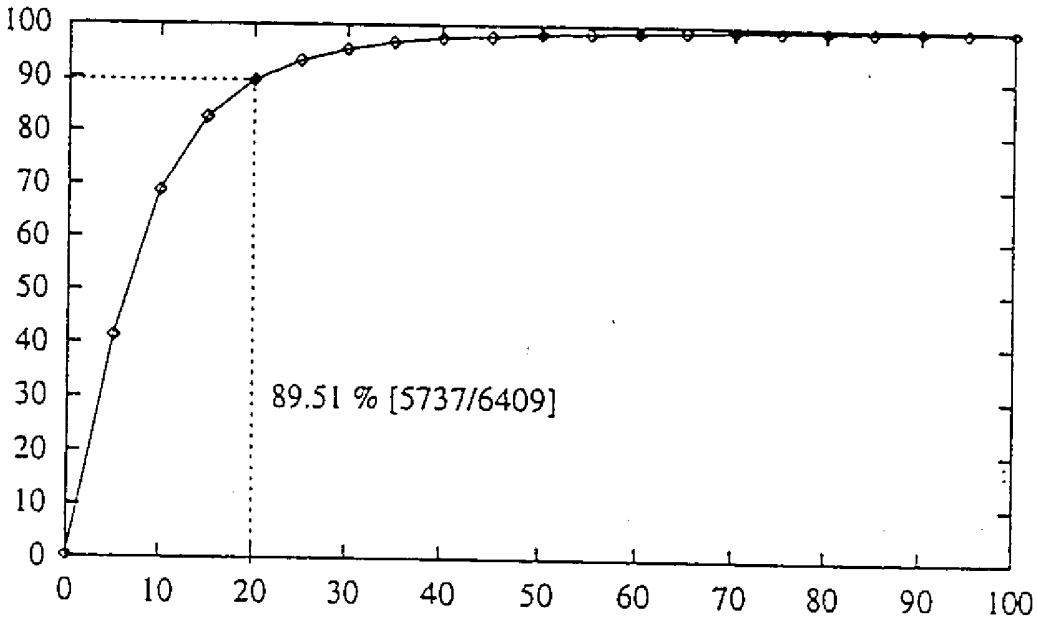
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
89.51 % [5737/6409]

圖 5

修正 補充
年 月 日

申請日期	86. 3. 10
案　　號	86102943
類　　別	G10L 5/02

公告本

A4 434528
C4

中文說明書修正頁(88年2月)

（以上各欄由本局填註）

發明 新型 專利說明書

一、發明新型名稱	中　文	基於分音節成寬音階，限制序列向量量化及隱藏馬可夫模型，自動將語音分節成用於語音處理應用的音素單元的方法及裝置
	英　文	~~A METHOD AND APPARATUS FOR AUTOMATIC SPEECH SEGMENTATION~~ INTO PHONEME-LIKE UNITS FOR USE IN SPEECH PROCESSING APPLICATIONS, AND BASED ON SEGMENTATIN INTO BROAD PHONETIC CLASSES, SEQUENCE-CONSTRAINED VECTOR QUANTIZATION, AND HIDDEN-MARKOV-MODELS
二、發明創作人	姓　名	1. 史帝芬 克拉倫斯 保斯 2. 伊文斯 吉曼尼 查理 坎普 3. 李昂那杜斯 法蘭西庫斯 威廉
	國　籍	1. 3. 荷蘭　2. 比利時
	住、居所	1. 荷蘭恩特荷芬市米蘭克街28號 2. 比利時李瑪市瑪西奧路20號 3. 荷蘭恩特荷芬市李尼吉哈夫區47號
三、申請人	姓　名（名稱）	荷蘭商皇家飛利浦電子股份有限公司
	國　籍	荷蘭
	住、居所（事務所）	荷蘭愛因和文市格羅尼渥街1號
	代表人姓　名	J．G．A．羅夫斯

434528

（由本局填寫）

承辦人代碼：	
大　　類：	
IPC分類：	

修正補充 88年1月26日

A6
B6

本案已向：

國（地區）	申請專利，申請日期：	案號：	，□有　□無主張優先權
歐盟	1996.2.27	96200509.6	

有關微生物已寄存於：　　，寄存日期：　　，寄存號碼：

（請先閱讀背面之注意事項再填寫本頁各欄）

裝　訂　線

五、發明說明（1）

發明背景

本發明是關於將語音自動分節以用於語音處理應用之方法。在各種可能的應用當中，一特定之應用是語音合成，特別是以雙音(diphones)串連的語音合成。雙音是短的語音分節，各分節主要包括二相鄰音素之間的一個轉換，再分別加上前一音素之最後部分及其接續音素的最初部分。雙音可根據若干本身已知之規則從已分節成單一音素的一資料庫中摘錄出來。通常，該資料庫包括一控制之下環境中，從一特定單一說話者錄製而得之分離字，而且亦包括經驗證拼音字(phonetic transcription)與聲音實現(acoustic realization)之間的相似性。1992年於加拿大Alberta省Banff市舉行的國際語音及語言處理會議中，O. Boëfard等人發表Automatic Generation of Optimized Unit Dictionaries for Text to Speech Synthesis（自動轉換文字爲語音合成之最佳辭典）一文，其序文與第1211-1215頁中即揭示一根據音素隱藏馬可夫模型(HMM)的簡明而可自動實現分節的方法。然而，如今已發現該已知方法的品質不足，因爲該方法所發現之界限一般而言過於偏離以一手動程序放置的相關界限位置。當然，如該等音素隱藏馬可夫模型(HMMs)在開始時，以一分離及手動分節之資料庫校對時，則可改善分節的準確度。但該手動分節資料庫的設定成本經常過高，因爲在一語音合成系統中需要一個新的說話者時，每次均須重複設定。因此，本發明最主要的目的之一在提出一項育於語音分節的方法，該方法爲完全自動、毋須手動分節語音資料，而且可提供比所提及之

五、發明說明（2）

方法爲佳的成果。

發明概述

根據本發明一種現象，本發明提供了一種以供語音處理應用使用的自動分節語音的方法，該方法包括下述步驟：

• 將一語音資料庫的發音予以分類並分節成發聲(voiced)、不發聲(unvoiced)及無聲(silence)三類寬音階(BPC)，用以獲得初步分節的位置；

• 利用初步的分節位置當做爲定點，俾在一SCVQ(限制序列向量量化)步驟中以限制序列向量量化進一步分節成類似音素之單元；

• 以該SCVQ(限制序列向量量化)步驟所提供的分節標注音素隱藏馬可夫模型，並進一步以巴姆-韋爾契估算法(Baum-Welch estimation)調整該等隱藏馬可夫模型(HMM)之參數；

• 最後，利用該等經過充分校準的隱藏馬可夫模型執行Viterb之發音錄寫的校整，而以此方法獲得該等最後的分節點。

所引用之該方法有一額外優點，即僅需要如發音錄寫中主要因素最少的啓始資料。特別是，不需要分別以手動方式分節資料庫以供估算該等隱藏馬可夫模型(HMM)參數。

有利的是，在校準之後可建立一雙音組以供進一步之使用，例如用於語音合成中。本發明已提供一種簡明及便宜的多數說話者之系統。

本發明亦有關於一於語音處理應用中使用之語音分節的裝置，該裝置包括：

五、發明說明（3）

• 由一語音資料庫所提供之寬音階(BPC)分節方法，其用以將所收到的語調予以分類及分節成發聲(voiced)、不發聲(unvoiced)及無聲(silence)等三類寬音階(BPC)，

• 由該寬音階(BPC)分節方法所提供之SCVQ(限制序列向量量化)分節方法，其利用初步的分節位置作爲固定點而藉由限制序列向量量化(SCVQ)執行進一步做類似音素單元之分節工作，

• 由該SCVQ(限制序列向量量化)分節方法所提供之語音隱藏馬可夫方法(HMM)，其用以開始音素隱藏馬可夫模型(HMM)及進一步調整隱藏馬可夫方法(HMM)之參數；

• 由該隱藏馬可夫方法(HMM)所控制之最後分節方法。

該種裝置將可使未經訓練之人員在短時間內用以訓練隨便一個新的說話者使用之。本發明進一步之優點詳述於隨後所申請專利範圍中。

圖式簡單說明

本發明上述及其它目標與優點將於揭示該較佳具體實施例時予以討論，尤其於所附圖示中更爲詳細，該等圖示顯示：

圖1：該裝置的整體方塊圖；

圖2：某一無意義字的五項測量數值；

圖3：一第一寬音階分節；

圖4：以向量量化之音素分節；以及

圖5：相同，但已經以HMM(隱藏馬可夫模型)分節予以改良。

五、發明說明（4）

較佳整體具體實施例之詳細說明

本發明意欲將一資料庫的各語調切割成一序列非重疊相鄰之音素分節，並形成該等分節與該音標拼字所提供的一序列音階標籤之間的一對一的對應。語音可適當地以一序列之聲音向量來說明，各聲音向量一致在通常爲10-20毫秒(ms)及每次轉變的2.5-10毫秒(ms)的時間內凸顯該語音之特性。任一時間t的p-次元聲音向量爲$o(t)=[o_1(t)\ldots o_p(t)]'$，而顯示向量轉換的聲調及一完整的時間系列則以$O(1,T)=o(1),\ldots o(T)$表示之。在本發明之具體實施例中，寬音階(BPC)分節之$p=5$、限制序列向量量化之$p=12$，而隱藏馬可夫模型之$p=51$。一寬音階元素或跨一分節l之類似音素單元可用一原型向量或以$\hat{c}_l$表示的質量中心表示之。分節l之分節內部距離$d_l(i,j)$可定界爲該跨越該分節之向量O(i,j)與該質量中心$\hat{c}_l$：$d_l(i,j)=\sum_{t=i}^{j} d(o(t),\hat{c}_l)$之間距離的總和。將變形：

$$\sum_{l=1}^{L} d_l(b_{l-1}+1,b_l)=\sum_{l=1}^{L}\ \sum_{t=b_{l-1}+1}^{b_l} d(o(t),\hat{c}_l)$$

減至最小後，其分節點b_l($l=1\ldots L-1$)。該一般公式此後用於測量距離、程序減少，以及用於決定該等質量中心。

圖1顯示一系統具體實施例之整體方塊圖。此處，項目20是一麥克風以用於接收來自一特定說話者的語音，該說話者受命說出該組預定之獨立字，各獨立字必須分節。該等字組已提議用於各種語言中。某些或所有該等字可能是無意義的字。方塊22中之第一階處理程序是定期取樣、數位化及過濾。其結果是儲存於中間記憶體24以用於從該處理程序中切斷語音接收。一般控制是位於方塊34，該方塊34

五、發明說明（5）

可爲一以適當方式經程式設計過的標準電腦，且該方塊34控制方塊24、31、36。該控制暗含該應用程式的同步化、供應，或還包括各處理步驟間的中間儲存器之同步化、供應。爲簡化起見，僅繪出一單一的單方向連接。接著，讀出記憶體24的發音之後，在方塊26中著手進行第一階段的分階及分節成發聲(VOIced)、不發聲(UNVoiced)及無聲(SILence)三類寬音階(BPC)。該初步寬音階(BPC)分節的結果進而在方塊28中予以處理，在該方塊28中執行第二階段之限制序列向量量化(SCVQ)以進一步分節成音素。方塊30實行第三階段。該方塊30利用先前SCVQ階段所傳來的分節以設定該等音素隱藏馬可夫模型(HMMs)之初值，然後藉由在欲分節的資料庫上執行巴姆-韋爾契估算法(Baum - Welch estimation)以進一步調整該等音素隱藏馬可夫模型(HMMs)。有關極佳及可得之隱藏馬可夫模型(HMMs)的論述與巴姆-韋爾契估算法(Baum - Welch)，請參閱L.R. Rabiner的文章'A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition'，該文發表於1989年2月之IEEE(電機及電子工程師學會)會報，第77冊，第2號，第257 - 286頁。最後，在方塊31中利用該等隱藏馬可夫模型(HMMs)以執行該等資料庫發音及其音標拼字之間的維特比暫時對準(Viterbi temporal alignment)動作。維特比對準(Viterbi alignment)本身是標準技術，請參考同一文件。

分節的結果是儲存於記憶體32中。雙音產生裝置39從每一發音所產生的序列音素中設立產生實際語音所需之各種雙音，該等雙音被寫回記憶體中當作一資料庫以用於依實

五、發明說明（6）

際從說話者收到的有限數量之語音為基礎而產生可能具有任意內容的語音，或用於其他語音處理應用。此外該控制是衍生自(觀念上的)控制／校正系統36，該系統36經由連線42而獲得該資料庫，經由連線40而接收該等雙音，且可進一步經由雙向連線38而與記憶體32進行互動。為簡潔起見，未顯示經由連線41之實際語音輸出裝置。該語音處理之流程圖一般是以相似方式設定，因為各個分別說出的字是在取得下一字之前先處理完全。

第一階段：分節成寬音階

第一階段分節應提供所謂固定點以用於後續階段。該等三類音階是未存在有任何語音波形的無聲(SIL)、語音波形為半週期性的發聲(VOI)及波形為非週期性或隨機之不發聲(UNV)。在該具體實施例中，分節是以下列五項不同的測量為基礎：

－ E_N：1-(正規化的短時間能量)，其中該正規化係使得無聲(SIL)具有趨近於1之 E_N；

－ E_{low}：(正規化的低頻能量)，範圍為50 - 1200 Hz；

－ E_{high}：(正規化的高頻能量)，範圍為2000 - 4000 Hz；

－ 零交越率(zero - crossing rate) Z_N；

－ 一次方LPC模型的第一LPC係數a1。

在該具體實施例中，該取樣波形x(k), k = 0 . . . N - 1已先藉由一過濾函數 $(1 - 0.95z^{-1})$ 施以頻應預矯(pre - emphasized)、封鎖並以漢明框選法(Hamming - windowed)分成20毫秒(ms)的時間格(frames)，而整個時間格位移2.5毫秒(ms)。圖2顯示一以荷蘭

五、發明說明（7）

語發音之無意義字'kekakke'的五項前述測量。緊接於該波形之下標示有該等相關音素記號。如圖所示，母音'A'的主要週期約爲10毫秒(msec)。該圖中可清楚地看到三個發聲(voiced)間隔、三個不發聲(unvoiced)間隔及四個無聲(silence)間隔。該五項前述測量的結果是以個別線圖顯示。某些轉變比其他更清晰可見。實際上，連續音素之間的轉變不一定暗示寬音階有改變。

接下來，首先以寬音階方式將該等發音之可用音標拼字對映至一譯音(transcription)；事實上，系統知道實際應已收到何字，因爲該說話者必須以規定順序輸入該等字。假設某一發音存在於一序列L'寬音階元件中，使用者必須在該觀察O_l^T中找出對應的連續分節與把定義於上之總距離減至最小的一組質量中心$\{\hat{c}_l\}$。此處所使用的距離是加權之歐氏距離(weighted Euclidean distance)，而該等測量的估算差異被用於當作加權因數。該最小化工作是在十分類似於下列極爲知名之階層建立法(level - building approach)的一動態程式設計結構中執行，請參照C.S. Myers及L.R. Rabiner之文章A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition，發表於1981年4月ASSP(美國語音處理協會)的IEEE(電機及電子工程師學會)會報，第29冊(Vol. 29)，編號2(No. 2)，第284 - 297頁。此處所使用之方法基本上是反覆性的且一直持續直到無法達成進一步改良。該程序之各次反覆均包括連續步驟。第一步針對一組已知的質量中心而找出一組最佳界線b_l。第二步利用最新獲取之分節點將該等

五、發明說明（8）

質量中心{$\hat{c}_l$}及該等差異予以更新。由於此處使用加權之歐氏距離，代表一寬音階的最新質量中心即等於該組已知分節點之全部觀察向量的數學平均。第一步從單元差異及代表下列理想寬音階的質量中心開始進行：

$\hat{c}_{SIL}=(1,0,0,1,1)'$、$\hat{c}_{UNV}=(0,0,1,1,0)'$、$\hat{c}_{VOI}=(0,1,0,0,1)'$

一寬音階元件的最小、平均及最大歷時時間相等於組成一寬音階分節之音素的最小、平均與最大歷時時間和(sum)。有關該等音素之最小及最大歷時時間可從文獻中得知。該等最小及最大歷時時間資訊是用於提供一寬音階(BPC)分節可存在的時間間隔上限與下限以限制對於寬音階(BPC)界線之搜尋。最後，平均歷時時間是用於按該發音的實際長度比例來縮放該等最小及最大歷時時間。

圖3顯示第一寬音階分節結果，而以與手工決定之轉變點相差小於一特定差的累積量來表示。如圖所示，因爲其差異小於20毫秒(msec)，故82%的轉變令人滿意地局限於一局部區域。

第二階段：限制序列向量量化(SCVQ)

上述分節產生固定點以用於進一步處理語音。該處理程序一般會縮減爲將各寬音階元件分節成其組成音素單元。更特別的是，該等發音被分節成連續性擬似無變化之元件。由於要求必須無變化，故雙母音及破裂音進一步劃分成其組成音素。再者，必須以類似於前一階段所使用之反覆方式找出將總體扭曲減至最小之分節界線與質量中心。對於第一步反覆步驟而言，該質量中心是定義成透過所有

五、發明說明（9）

附帶相同標籤並聚積成一單一觀察序列 $O_{b_{l-1}+1}^{b_l} = o_{(b_{l-1}+1)}, \ldots o(b_l)$ 的假設將扭曲減至最小之向量。

$$\hat{c}_l = arg_{c_l} \min \frac{1}{b_l - b_{l-1}} \sum_{t=b_{l-1}+1}^{b_l} d(o(t), c_l)$$

質量中心是以該方式從手頭上的觀察中獲得，其明顯的優點在於事先不需任何知識。第一次限制序列向量量化(SCVQ)反覆是藉由動態階層建立程式設計(level - building dynamic programming)來完成。第二次反覆之方案相同於早先的寬音階(BPC)分節階段者。第一次反覆所發現之音素質量中心被使用於下一步驟中當作參考，使程序縮減爲樣板比對。第三及最後一步驟針對產生之各標籤而計算質量中心。進一步的反覆不會再產生任何改善。圖4顯示利用由第一階段所產生之固定點而進行的音素分節。完全不同於該寬音階(BPC)階段的是，此處該等標籤大約與50個不同音階相關。該分節在考量如此眾多音階之下必然比第一階段之分節更易爲人所接受：在全部案例當中，有70%的案例與手工分節之誤差小於20毫秒(msec)。

第三階段：以隱藏馬可夫模型(HMM)進行分節

使用隱藏馬可夫模型可更進一步藉由闡述統計學上的語音變化性而改善限制序列向量量化(SCVQ)所獲致之結果。再者，總共約使用50個不同的音素單元，各自獲得其自己的隱藏馬可夫模型(HMM)。爲確保據實進行完全自動程序，故以該限制序列向量量化(SCVQ)階段所獲致之結果設定該等隱藏馬可夫模型(HMMs)的初值。除該脈衝束(burst)單元以外，各音素單元均獲得一簡單的由左至右而含有六種

五、發明說明（10）

狀態之隱藏馬可夫模型(HMM)拓樸，且包括自我迴路及直接往下一階段的轉變；因此，歷時時間至少爲六個時間格，各時間格爲5毫秒(msec)，而該等時間格較短於早先考量者。一擬似脈衝束的分節是以具有同類自我迴路及轉變之2-階段模型來代表。

放射分佈(emission distributions)屬於連續類型。各觀察向量被分割成4子向量，分別代表

－瞬間特性(16波道的過濾束分析)；

－比對前一時間格而決定的該等特性之第一及第二有限的差異；

－能量數據。

各子向量之放射機率密度是一多變常態混合分配(multivariate Gaussian mixture distribution)。

該等隱藏馬可夫模型(HMM)參數是按以下方式調整。請考慮一特定音素單元及在該限制序列向量量化(SCVQ)階段爲此切割出的該組觀察序列。該音素單元之隱藏馬可夫模型(HMM)的初值設定是藉由將各觀察序列之聲音向量均勻分散於該單元的6(或2)種隱藏馬可夫模型(HMM)狀態之間來完成。結果，一觀察的特定聲音向量以均勻分散於該等狀態之中的方式分派給該模型之各狀態。接下來使用一k-平均演算法(k-means algorithm)將該等聲音向量分割成叢集(clusters)，從該等叢集中可計算出該混合分配的參數之初值。亦請見J.G. Wilpon及L.R. Rabiner的文章A Modified K-Means Clustering Algorithm for Use in Isolated Word Recognition，發表於1985

五、發明說明 (11)

年6月ASSP(美國語音處理協會)的IEEE(電機及電子工程師學會)會報，第33冊(Vol. 33)，編號3 (No. 3)，第587 - 594頁。

該等參數是平均向量、該等組成密度的斜度矩陣(covariance matrices)及混合加權(mixture weights)。該等轉變機率的初值被設定爲非0之任意值。一旦獲得該等初始隱藏馬可夫模型(HMMs)，即施以一受監督之維特比校準(Viterbi training)，結果造成一新的分節。從該處可計算出最新的隱藏馬可夫模型(HMM)參數。後者包括利用從該比對路徑收集而得的狀態與狀態之間的轉變統計數字計算出之轉變機率。接下來，針對各完整發音，依照巴姆-韋爾契估算法(Baum - Welch)進行參數估算，藉以調整該等模型。

最後，該等完全經過校正之隱藏馬可夫模型(HMMs)被用以執行該資料庫中的各字及其音標拼字之間的維特比對準(Viterbi alignment)動作。此舉產生最後的分節點。圖5顯示該結果：其改善效果相對於圖4之限制序列向量量化(SCVQ)分節而言是驚人的：接近90%的轉變具有小於20毫秒(msec)之誤差。

434528

A5
B5

四、中文發明摘要（發明之名稱：基於分音節成寬音階，限制序列向量量化及隱藏馬可夫模型，自動將語音分節成用於語音處理應用的音素單元的方法及裝置）

爲進行機器語音分節，首先將一已知語音字串資料庫中的發音予以分類及分節成發聲(voiced)、不發聲(unvoiced)及無聲(silence)三類寬音階(BPC)。其次，利用初步的分節位置當作固定點，再以限制序列向量量化進一步分節成音素單元。最後，藉由隱藏馬可夫模型精準調整爲該等分節音素，且在校準之後組成一雙音可供進一步之使用。

英文發明摘要（發明之名稱：A METHOD AND APPARATUS FOR AUTOMATIC SPEECH SEGMENTATION INTO PHONEME-LIKE UNITS FOR USE IN SPEECH PROCESSING APPLICATIONS, AND BASED ON SEGMENTATIN INTO BROAD PHONETIC CLASSES, SEQUENCE-CONSTRAINED VECTOR QUANTIZATION, AND HIDDEN-MARKOV-MODELS）

For machine segmenting of speech, first utterances from a database of known spoken words are classified and segmented into three broad phonetic classes (BPC) voiced, unvoiced, and silence. Next, using preliminary segmentation positions as anchor points, sequence-constrained vector quantization is used for further segmentation into phoneme-like units. Finally, exact tuning to the segmented phonemes is done through Hidden-Markov Modelling and after training a diphone set is composed for further usage.

（請先閱讀背面之注意事項再填寫本頁各欄）

裝 訂 線

A8
B8
C8
D8

六、申請專利範圍

1. 一種自動將語音分節俾以用於語音處理應用之方法，其包括下述步驟：

• 將一語音資料庫的發音予以分類及分節成發聲(voiced)、不發聲(unvoiced)及無聲(silence)三類寬音階(BPC)，以便獲得初步的分節位置；

• 利用初步的分節位置作爲固定點，在一SCVQ(限制序列向量量化)步驟中再以限制序列向量量化進一步分節成音素單元；

• 以該SCVQ(限制序列向量量化)步驟所提供的分節設定音素隱藏馬可夫模型之初值，並進一步藉由巴姆-韋爾契估算法(Baum - Welch estimation)調整該等隱藏馬可夫模型(HMM)參數；

• 最後，利用該等完全經過校正的隱藏馬可夫模型以執行該等發音及其音標拼字之間的維特比對準(Viterbi alignment)動作，且以該方式獲得最後之分節點。

2. 如申請專利範圍第1項之方法，在該分節之後組成一組雙音以供進一步使用。

3. 如申請專利範圍第1項或第2項之方法，其中該語音處理是語音合成。

4. 一種將語音分節以用於語音處理應用之裝置，其包括：

• 由一語音資料庫所提供之寬音階(BPC)分節方法，用以將收到的發音予以分類並分節成發聲(voiced)、不發聲(unvoiced)及無聲(silence)三類寬音階(BPC)，

• 由該寬音階(BPC)分節方法所提供之SCVQ(限制序列向

六、申請專利範圍

量量化)分節方法，用以利用初步的分節位置作爲固定點以藉由限制序列向量量化(SCVQ)執行進一步分節成音素單元，

• 由該SCVQ(限制序列向量量化)分節方法所提供之語音隱藏馬可夫方法(HMM)，用以設定音素隱藏馬可夫模型(HMM)的初值，並進一步調整隱藏馬可夫模型(HMM)參數；

• 由該隱藏馬可夫方法(HMM)所控制的最後分節方法。

5. 如申請專利範圍第4項之裝置，其包括該分節方法所提供之雙音產生方法以用於組成一組雙音。

6. 如申請專利範圍第4項或第5項之裝置，其尚包括一輸出控制階段以用於透過一介於該調整方法與該語音合成輸出階段之間的中間儲存階段來控制一語音合成輸出階段。