公告本

申請日期	90 5 17
案號	90111865
類別	G10L 15/00

A4
C4

512308

（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	以聲音為驅動機制的嘴型即時動態模擬方法
	英文	
二、發明創作人	姓名	王學武
	國籍	中華民國
	住、居所	台北市敦化南路二段二一六號十三樓
三、申請人	姓名（名稱）	財團法人資訊工業策進會
	國籍	中華民國
	住、居所（事務所）	台北市和平東路二段一0六號十一樓
	代表人姓名	林逢慶

裝 訂 線

經濟部智慧財產局員工消費合作社印製

本紙張尺度適用中國國家標準（CNS）A4規格（210×297公釐）

（由本局填寫）

承辦人代碼：	A6 B6
大　　類：	
ＩＰＣ分類：	

本案已向：

國（地區）　申請專利，申請日期：　　案號：　　，□有　□無主張優先權

無

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

經濟部智慧財產局員工消費合作社印製

本紙張尺度適用中國國家標準（CNS）A4規格（210×297公釐）

五、發明說明（ | ）

【本發明之領域】

本發明係有關嘴型模擬之技術領域，尤指一種以聲音為驅動機制的嘴型即時動態模擬方法。

【本發明之背景】

按，隨著電腦技術的發展，各種造型的嘴型與說話時的搭配，無論在 3D 或是 2D 方面的應用，例如在現今的電影、電腦遊戲等視聽娛樂之應用上，已經成為不可或缺的一部分。然而在這些應用中，一般而言，造型的嘴型與聲音的搭配大都是以手工的方式調整，而以人工製作嘴形30秒約需要1.5小時，因此，其耗時極長而缺乏效率，而即使有提供語音的辨認來決定對應之嘴形，也都是將聲音轉成相對應的文字，然後再依照相對應文字的嘴型大小進行嘴型仿眞，惟此種仿眞方式都僅能限制與單一的語言，例如為純中文與純英文，而不能中英文混合。因此，以前述習知嘴形模擬方法來製作的動畫或是影片，通常非常的耗費人力與時間，而有予以改進之必要。

發明人爰因於此，本於積極發明之精神，亟思一種可以解決上述問題之「以聲音為驅動機制的嘴型即時動態模擬方法」，幾經研究實驗終至完成此項新穎進步之發明。

【本發明之概述】

本發明之目的係在提供一種以聲音為驅動機制的嘴型即時動態模擬方法，以達成即時的同步動態模擬，而無需用到複雜的語音辨認技術，且能打破單一語言的限制。

（請先閱讀背面之注意事項再填寫本頁）
裝
訂
線
經濟部智慧財產局員工消費合作社印製

本紙張尺度適用中國國家標準（CNS）A4 規格（210 × 297 公釐）

五、發明說明（２）

為達前述之目的，本發明之以聲音為驅動機制的嘴型即時動態模擬方法，主要包括下述之步驟：（Ａ）將輸入之影音資訊的聲音分成複數個連續而且有重疊的音框；（Ｂ）將每一個音框轉成複數個倒頻譜參數，並求取每個音框內嘴型的寬度與高度兩個參數，其中，每一個音框係由對應之倒頻譜參數及嘴型的寬度與高度參數所組成之一音頻-視覺向量所代表；（Ｃ）利用向量量化將該等音頻-視覺向量分成複數群，以使能量與嘴形大小相近之音頻-視覺向量在同一群；（Ｄ）以高斯混合模型作為每一群的表示基礎；以及，（Ｅ）對每一個群，根據向量量化所得到的結果，設定起始設定值，以利用最大預測演算法來求取每一群的最佳高斯混合模型的參數值，俾供模擬受測者之聲音。

由於本發明設計新穎，能提供產業上利用，且確有增進功效，故依法申請專利。

為使　貴審查委員能進一步瞭解本發明之結構、特徵及其目的，茲附以圖式及較佳具體實施例之詳細說明如后：

【圖式簡單說明】

第１圖：係為本發明之以聲音為驅動機制的嘴型即時動態模擬方法在訓練階段的流程圖。

第２圖：係為本發明之以聲音為驅動機制的嘴型即時動態模擬方法在求取訓練參數之組合示意圖。

（請先閱讀背面之注意事項再填寫本頁）
裝
訂
線
經濟部智慧財產局員工消費合作社印製
本紙張尺度適用中國國家標準（CNS）A4規格（210×297公釐）

五、發明說明（３）

第3圖：係為本發明之以聲音為驅動機制的嘴型即時動態模擬方法在模擬階段之流程圖。

【較佳具體實施例之詳細說明】

為說明本發明之以聲音為驅動機制的嘴型即時動態模擬方法，請先參照第1圖所示，其顯示本發明之方法在訓練階段之流程圖。本發明在訓練階段是以攝影機拍攝訓練者的朗誦事先設計好的數段文字，俾以求取訓練參數，併請參照第2圖所示所欲求取之訓練參數之組合示意圖，首先，將輸入之影音資訊（Video及Audio）的聲音分成複數個連續而且有重疊的音框（步驟S11），並以特徵分析（Feature Extraction）將每一個音框轉成複數個（例如13個）倒頻譜參數（Cepstrum coefficients）（以a表示）（步驟S12），且相對應於每一個音框，以透過嘴形追蹤程式（Lip-tracking program）取得這個音框內嘴型的寬度（Width）與高度（Height）兩個參數（以v表示）（步驟S13），而對於每一個音框，此15個參數便可組成為一個音頻-視覺向量（Audio-visual feature vector）（步驟S14），以作為該音框的代表。

在取得一系列的音頻-視覺向量v之後，再利用向量量化（Vector Quantization）將這些音頻-視覺向量分成N群（步驟S15），以使能量與嘴形大小相近之音頻-視覺向量在同一群，而每一群即對應有一個收斂後的中心向量（Center Vector）與共變異矩陣（Covariance

（請先閱讀背面之注意事項再填寫本頁）
裝 訂 線
經濟部智慧財產局員工消費合作社印製

本紙張尺度適用中國國家標準（CNS）A4規格（210×297公釐）

五、發明說明（4）

Matrix），步驟S16係以高斯混合模型（Gaussian Mixture Model，GMM）作為每一群的表示基礎，亦即，以GMM來表示音頻-視覺向量的機率分佈，其中，GMM是K個高斯函數（Gaussian function）的權重和（weighted sum），可由以下的公式所示：

$$p(o)=\sum_{i=i}^{k} w_i g[\mu_i,\Sigma_i](o)$$

其中，w_i 為混合權重，$g[\mu_i,\Sigma_i](o)$ 為具有平均值（mean）μ_i與共變異矩陣Σ_i 的高斯函數，如下所示。

$$g[\mu_i,\Sigma_i](o)=\frac{1}{\sqrt{(2\pi)^{15}|\Sigma_i|}}\exp\left\{-\frac{1}{2}(o-\mu_i)\Sigma^{-1}(o-\mu_i)\right\}$$。

於步驟S17中，對每一個群i，根據向量量化所得到的結果，取其中心向量作為初始平均值（initial mean）μ_i，以收斂後的共變異矩陣作為分群i之共變異矩陣Σ_i，而分群i中的音頻-視覺向量數目，佔所有音頻-視覺向量數目的比例則作為初始混合權重（initial mixture weight）w_i，而以前述之起始設定值，即可利用最大預測演算法（Expectation-Maximization algorithm）來求取每一群的最佳高斯混合模型的參數值：μ_i、Σ_i與w_i。

而在模擬階段，參照第3圖所示，係首先將受測者的聲音分成複數個連續而且有重疊的音框（步驟S31），再將每一個音框轉成複數個（例如13個）倒頻譜參數（以a表示）（步驟S32），也就是聲音特徵向量a。步驟S33則根據a出現在每一群中的機率值，取一個加權平均值而求出目前的嘴型大小 $\tilde{v}$。另為了加速求解的速度，可設定N

（請先閱讀背面之注意事項再填寫本頁）
裝
訂
線
經濟部智慧財產局員工消費合作社印製

本紙張尺度適用中國國家標準（CNS）A4規格（210×297公釐）

五、發明說明（5）

= K，亦即，將向量量化的分群數目設定為與表示 GMM 所使用的高斯函數的個數相同，而其求解之公式如下：

$$\tilde{v} = E[v \mid a] = \sum_{i=1}^{K} \frac{w_i p_{i,a}(a)}{p_a(a)} * \bar{v}_i$$

其中 $p_{i,a}(a) = g[\mu_i, \Sigma_i](a)$，

$$p_a(a) = \sum_{i=i}^{K} w_i g[\mu_i, \Sigma_i](a)$$，

$$\bar{v}_i = \int v g[u_i, \Sigma_i](v_i) dv$$ 。

由以上之說明可知，本發明之方法係以分群的方法，將聲音與嘴型大小依照高斯混合模型與向量量化做一個統計上的分群。以這些為分群的基礎，當有說話聲音輸入時，可根據聲音落在各分群的機率，算出該聲音所相對應的嘴型大小。而依照此嘴型大小，便可以針對造型的嘴型做即時的聲音與嘴型大小的同步動態模擬。因此，無需用到複雜的語音辨認技術，即可實現嘴型之模擬，同時亦可打破單一語言的限制，且能達成及時的同步動態模擬。

綜上所陳，本發明無論就目的、手段及功效，在在均顯示其迥異於習知技術之特徵，為嘴形模擬之設計上的一大突破，懇請 貴審查委員明察，早日賜准專利，俾嘉惠社會，實感德便。惟應注意的是，上述諸多實施例僅係為了便於說明而舉例而已，本發明所主張之權利範圍自應以申請專利範圍所述為準，而非僅限於上述實施例。

（請先閱讀背面之注意事項再填寫本頁）

裝 訂 線

經濟部智慧財產局員工消費合作社印製

本紙張尺度適用中國國家標準（CNS）A4 規格（210 × 297 公釐）

A5
B5

四、中文發明摘要（發明之名稱：以聲音為驅動機制的嘴型即時動態模擬方法）

本發明係為一種以聲音為驅動機制的嘴型即時動態模擬方法，其應用高斯混合模型與向量量化作為聲音與嘴型大小的分群基礎。在訓練階段，聲音與嘴型的同步資料從視訊中取得，聲音分成連續而且有重疊的音框，每一個音框轉成多個倒頻譜參數，嘴型的部分則擷取寬與高兩個參數，以構成一個向量。在取得一系列向量後，透過 向量量化將其分群，再以高斯混合模型作為每一群的描述基礎，透過最大預測演算法找出最佳的描述方式。而在聲音與嘴型大小的對應階段，首先將聲音分成連續而且有重疊的音框，每一個音框轉成多個倒頻譜參數，再算出這個參數在每一個分群中的出現機率，根據這個機率值與每一個分群中所相對應的嘴型大小，以加種平均的方式，而求得每一段聲音的嘴型大小。

英文發明摘要（發明之名稱：）

（請先閱讀背面之注意事項再填寫本頁各欄）

裝 訂 線

經濟部智慧財產局員工消費合作社印製

六、申請專利範圍

1. 一種以聲音為驅動機制的嘴型即時動態模擬方法，主要包括下述之步驟：

（A）將輸入之影音資訊的聲音分成複數個連續而且有重疊的音框；

（B）將每一個音框轉成複數個倒頻譜參數，並求取每個音框內嘴型的寬度與高度兩個參數，其中，每一個音框係由對應之倒頻譜參數及嘴型的寬度與高度參數所組成之一音頻-視覺向量所代表；

（C）利用向量量化將該等音頻-視覺向量分成複數群，以使能量與嘴形大小相近之音頻-視覺向量在同一群；

（D）以高斯混合模型作為每一群的表示基礎；以及

（E）對每一個群，根據向量量化所得到的結果，設定起始設定值，以利用最大預測演算法來求取每一群的最佳高斯混合模型的參數值，俾供模擬受測者之聲音。

2. 如申請專利範圍第1項所述之方法，其更包含下述之步驟：

（F）將受測者的聲音分成複數個連續而且有重疊的音框，再將每一個音框轉成複數個代表聲音特徵向量之倒頻譜參數；以及

（G）根據聲音特徵向量出現在每一群中的機率值，取一個加權平均值而求出對應於受測者聲音之嘴型大小。

（請先閱讀背面之注意事項再填寫本頁）
裝
訂
線

經濟部智慧財產局員工消費合作社印製

六、申請專利範圍

3. 如申請專利範圍第1項所述之方法，其中，於步驟（B）中，係以特徵分析將每一個音框轉成複數個倒頻譜參數。

4. 如申請專利範圍第1項所述之方法，其中，於步驟（B）中，係以透過嘴形追蹤程式取得音框內嘴型的寬度與高度兩個參數。

5. 如申請專利範圍第1項所述之方法，其中，於步驟（C）中，每一群具有一個收斂後的中心向量與共變異矩陣。

6. 如申請專利範圍第1項所述之方法，其中，於步驟（D）中，係以高斯混合模型來表示音頻-視覺向量的機率分佈。

7. 如申請專利範圍第6項所述之方法，其中，該高斯混合模型是K個高斯函數的權重和，可由以下的公式所示：

$$p(o)=\sum_{i=i}^{k} w_i g[\mu_i,\Sigma_i](o)\ ,$$

當中，w_i 為混合權重，$g[\mu_i,\Sigma_i](o)$ 為具有平均值 μ_i 與共變異矩陣 Σ_i 的高斯函數，其可表示為：

$$g[\mu_i,\Sigma_i](o)=\frac{1}{\sqrt{(2\pi)^{15}|\Sigma_i|}}\exp\left\{-\tfrac{1}{2}(o-\mu_i)\Sigma^{-1}(o-\mu_i)\right\}\ 。$$

8. 如申請專利範圍第7項所述之方法，其中，於步驟（E）中，對每一個群i，係取其中心向量作為初始平均值 μ_i，以收斂後的共變異矩陣作為分群i之共變異矩陣 Σ_i，而分群i中的音頻-視覺向量數目，佔所有音頻-視覺向量數

（請先閱讀背面之注意事項再填寫本頁）
裝
訂
線
經濟部智慧財產局員工消費合作社印製

六、申請專利範圍

目的比例則作為初始混合權重 w_i，俾供作為起始設定值而求取每一群的最佳高斯混合模型的參數值 μ_i、Σ_i 與 w_i。

9. 如申請專利範圍第8項所述之方法，其中，向量量化的分群數目係設定為與表示高斯混合模型所使用的高斯函數的個數相同，以依據以下之公式求解：

$$\tilde{v} = E[v \mid a] = \sum_{i=1}^{K} \frac{w_i p_{i,a}(a)}{p_a(a)} * \bar{v}_i ,$$

當中，$p_{i,a}(a) = g[\mu_i, \Sigma_i](a)$，

$$p_a(a) = \sum_{i=i}^{K} w_i g[\mu_i, \Sigma_i](a) ,$$

$$\bar{v}_i = \int v g[u_i, \Sigma_i](v_i) dv ,$$

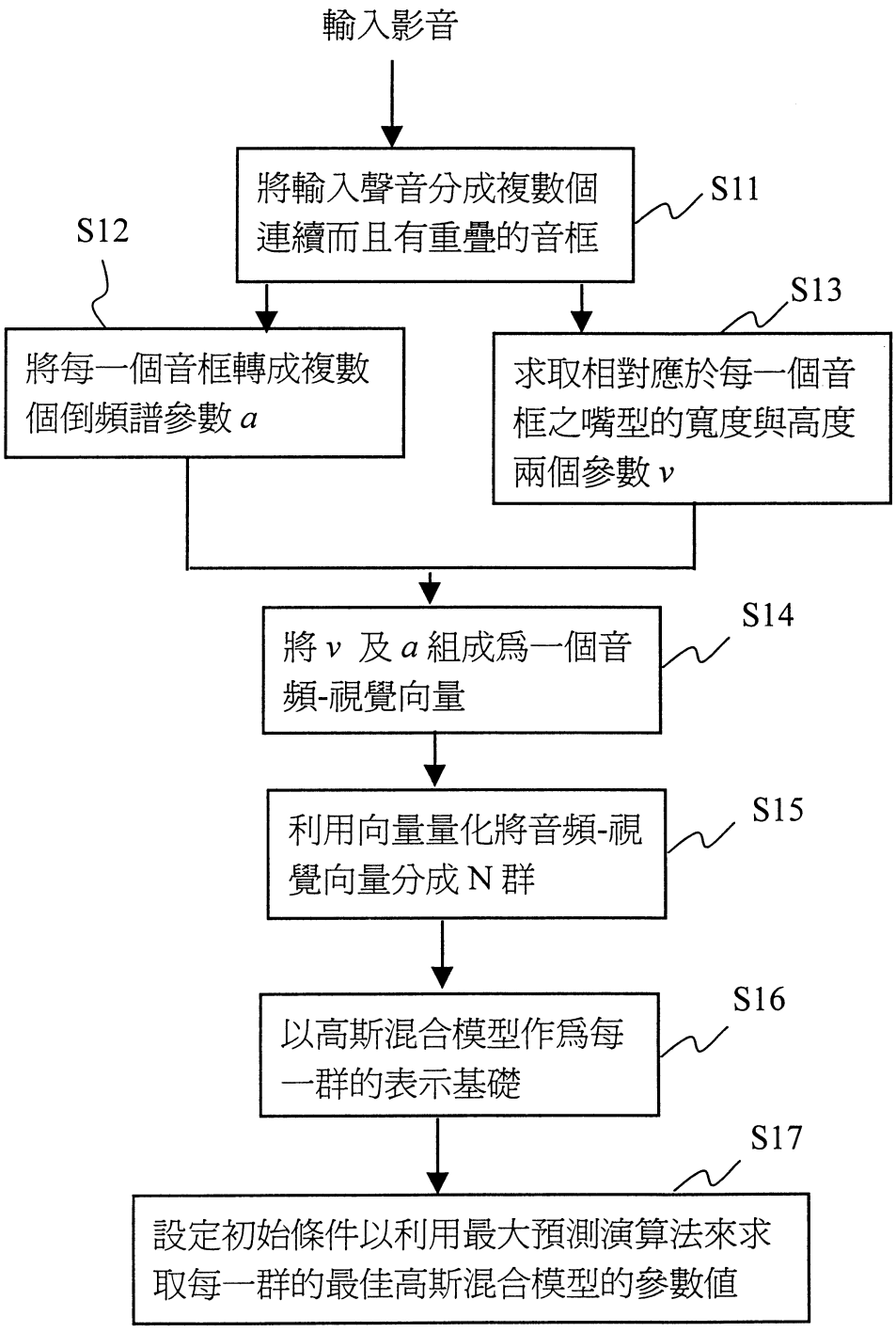
a 表示倒頻譜參數，v 表示嘴型的寬度與高度兩個參數，$\tilde{v}$ 表示嘴型大小。

（請先閱讀背面之注意事項再填寫本頁）

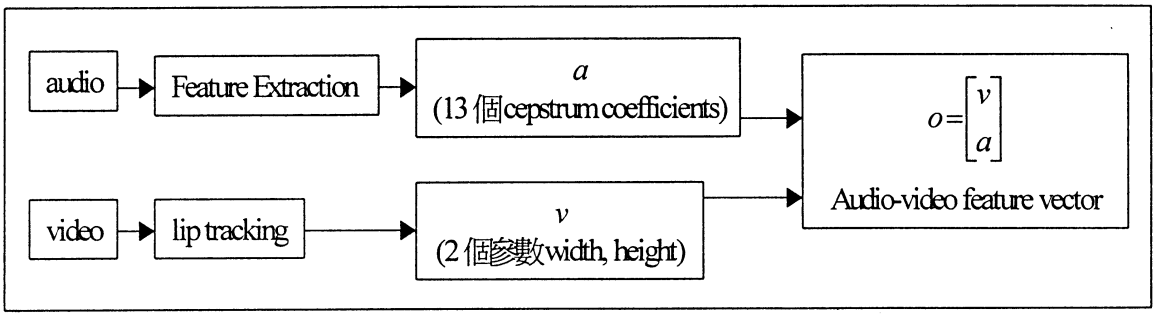
裝 訂 線

經濟部智慧財產局員工消費合作社印製

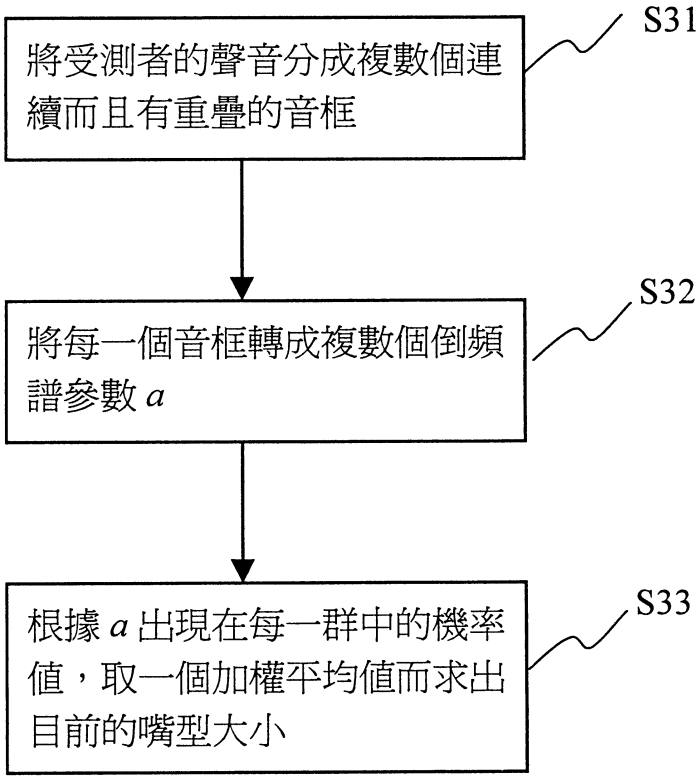
本紙張尺度適用中國國家標準（CNS）A4規格（210×297公釐）


輸入影音
將輸入聲音分成複數個連續而且有重疊的音框
S11
S12
S13
將每一個音框轉成複數個倒頻譜參數 a
求取相對應於每一個音框之嘴型的寬度與高度兩個參數 v
將 v 及 a 組成爲一個音頻-視覺向量
S14
利用向量量化將音頻-視覺向量分成 N 群
S15
以高斯混合模型作爲每一群的表示基礎
S16
S17
設定初始條件以利用最大預測演算法來求取每一群的最佳高斯混合模型的參數值

第 1 圖


audio
Feature Extraction
a
(13 個cepstrum coefficients)
o=
v
a
Audio-video feature vector
video
lip tracking
v
(2 個參數width, height)

第 2 圖


將受測者的聲音分成複數個連續而且有重疊的音框
S31
將每一個音框轉成複數個倒頻譜參數 a
S32
根據 a 出現在每一群中的機率值，取一個加權平均值而求出目前的嘴型大小
S33

第 3 圖