I866178

(19)中華民國智慧財產局

(12)發明說明書公告本 **(11)證書號數：TW I866178 B**

(45)公告日： 中華民國 113 (2024) 年 12 月 11 日

(21)申請案號：112115390 (22)申請日： 中華民國 112 (2023) 年 04 月 25 日

(51)Int. Cl.： *G10L13/033 (2013.01)* *G10L13/04 (2013.01)*
G10L17/04 (2013.01)

(71)申請人：旭瑞文化傳媒股份有限公司 (中華民國) SHINERAY CO., LTD. (TW)
臺北市內湖區瑞光路 70 號 5 樓

(72)發明人：林世海 (TW) ；葉韋麟 (TW) ；周敬程 (TW)

(74)代理人：郭雨嵐；林發立

(56)參考文獻：

TW	202119391A	CN	109788345A
CN	114025186A	US	2022/0122593A1

審查人員：施孝欣

申請專利範圍項數：20 項 圖式數：6 共 30 頁

(54)名稱

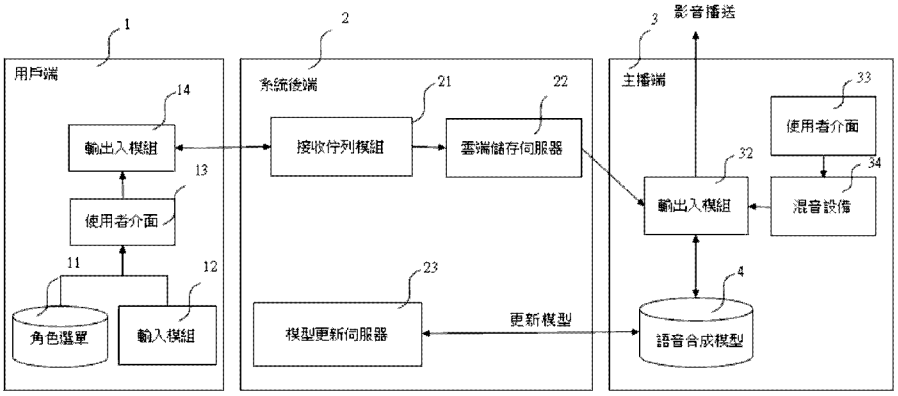
影音播送系統及其方法

(57)摘要

本發明影音播送系統包含：一語音合成模型，包含複數個特色參數；以及一系統後端，允許至少一用戶端與至少一主播端連線登入，並接收該用戶端所傳送的主播號、用戶號、輸入訊息與一預設語音編號或一客製化語音；其中，該語音合成模型將該預設語音編號對應至少部分的特色參數或根據該客製化語音選用至少一特色參數，並根據該對應或選用的特色參數與該輸入訊息產生一效果語音，俾使該主播端提供的影音播送含有一主播端影像、該用戶號與該效果語音。

The present invention provides an audio-visual broadcasting system comprising: a speech synthesis model including a plurality of characteristic parameters; and a system backend allowing at least one user end and at least one host end to connect and log in thereto and receive a host number, a user number, an input message and a preset voice number or a customized voice sent by the user end, wherein the speech synthesis model matches the preset voice number with at least a portion of the characteristic parameters or selects at least one characteristic parameter based on the customized voice, and then generates an effect voice based on the matched or selected characteristic parameters and the input message, so that an audio-visual broadcast provided by the host end includes a video of the host end, the user number and the effect voice.

指定代表圖：



【圖 2A】

符號簡單說明：

1:用戶端

11:角色選單

12:輸入模組

13:使用者介面

14:輸出入模組

2:系統後端

21:接收佇列模組

22:雲端儲存伺服器

23:模型更新伺服器

3:主播端

32:輸出入模組

33:使用者介面

34:混音設備

4:語音合成模型

公告本

I866178

【發明摘要】

【中文發明名稱】 影音播送系統及其方法

【英文發明名稱】 Audio-visual broadcasting system and method thereof

【中文】本發明影音播送系統包含：一語音合成模型，包含複數個特色參數；以及一系統後端，允許至少一用戶端與至少一主播端連線登入，並接收該用戶端所傳送的主播號、用戶號、輸入訊息與一預設語音編號或一客製化語音；其中，該語音合成模型將該預設語音編號對應至少部分的特色參數或根據該客製化語音選用至少一特色參數，並根據該對應或選用的特色參數與該輸入訊息產生一效果語音，俾使該主播端提供的影音播送含有一主播端影像、該用戶號與該效果語音。

【英文】The present invention provides an audio-visual broadcasting system comprising: a speech synthesis model including a plurality of characteristic parameters; and a system backend allowing at least one user end and at least one host end to connect and log in thereto and receive a host number, a user number, an input message and a preset voice number or a customized voice sent by the user end, wherein the speech synthesis model matches the preset voice number with at least a portion of the characteristic parameters or selects at least one characteristic parameter based on the customized voice, and then generates an effect voice based on the matched or selected characteristic parameters and the input message, so that an audio-visual broadcast provided by the host end includes a video of the host end, the user number and the effect voice.

【指定代表圖】　　圖2A

【代表圖之符號簡單說明】

1 用戶端

11 角色選單

12 輸入模組

13 使用者介面

14 輸出入模組

2 系統後端

21 接收佇列模組

22 雲端儲存伺服器

23 模型更新伺服器

3 主播端

32 輸出入模組

33 使用者介面

34 混音設備

4 語音合成模型

【發明說明書】

【中文發明名稱】 影音播送系統及其方法

【英文發明名稱】 Audio-visual broadcasting system and method thereof

【技術領域】

【0001】 本發明係關於一種影音播送系統及其方法，特別是一種根據輸入訊息產生一效果語音之影音播送系統及其方法。

【先前技術】

【0002】 關於語音合成TTS(Text to Speech/文字轉聲音)技術，大部分的人聯想到的都是Google小姐或Siri，是網路影音產業熟知之聲音播送方式，如實況直播中，為了刺激不同客戶來捐贈禮物給直播主，直播平台除了推廣圖片/動畫相關禮物之外，漸漸開始有音效相關的禮物。

【0003】 然而，傳統的語音合成技術通常是基於單一聲音來源的合成，且產生的語音效果可能會顯得不自然或不符合特定角色的語音特徵。這些音效禮物往往只能是固定的內容，例如欲錄製好的音檔，往往難以表達用戶真正想表達。如果是自己喜歡或熟悉的聲音，能夠被播送至直播平台給對方聽，也能夠更影響共鳴。現有之直播平台注重於文字、貼圖與禮物的樣式，積極開發不同禮物的種類、音效，或是有聲貼圖，都是以單一語音搭配單一畫面的對應呈現，而往往這些音效並不一定能呈現觀眾想表達的話語。

【0004】故現有語音合成技術仍有以下缺點：缺乏更多元化的表達方式，與語音合成模型只是將文字轉換為語音，缺乏針對不同角色與情境設計專屬的聲音效果。

【0005】因此，本技術領域之產業需要一種新的影音播送系統及其方法，該系統結合了語音合成技術和即時傳輸技術，允許用戶間或複數用戶與主播之間進行即時互動交流，並支持多角色聲音特徵的語音合成技術，以提供更加自然和流暢的語音效果。利用更少的語句與不必透過人工標註以降低作業成本，並同時提供可供選擇之客製化語音之訊息交流平台系統，為該產業亟待解決之問題。

【發明內容】

【0006】為解決上述問題，本發明的目的在於提供一種影音播送系統，將輸入訊息產生一效果語音應用於用戶間或用戶與主播之間進行即時互動交流。

【0007】為了達到上述的目的，本發明影音播送系統包含：一語音合成模型，包含複數個特色參數；以及一系統後端，允許至少一用戶端與至少一主播端連線登入，並接收該至少一用戶端所傳送的一主播號、一用戶號、一預設語音編號與一輸入訊息，其中該預設語音編號對應一角色選單所提供複數個角色之一；其中該語音合成模型將該預設語音編號對應至少部分的特色參數，並根據該對應的特色參數與該輸入訊息產生一效果語音，俾使該主播端提供的一影音播送含有一主播端影像、該用戶號與該效果語音。

【0008】根據本發明影音播送系統所提供之一種態樣，該語音合成模型配置於該系統後端；該系統後端根據該主播號傳送該用戶號與該效果語音至對應該主播號的主播端，俾使該主播端提供的該影音播送含有該主播端影像、該用戶

號與該效果語音。該系統後端根據該主播號傳送該輸入訊息至對應該主播號的主播端，俾使該主播端提供的該影音播送含有該主播端影像、該用戶號、該輸入訊息與該效果語音。

【0009】根據本發明影音播送系統所提供之另一種態樣，該語音合成模型配置於該主播端。該系統後端根據該主播號傳送該用戶號與該預設語音編號至對應該主播號的主播端，俾使該主播端提供的該影音播送含有該主播端影像、該用戶號與該效果語音。該系統後端根據該主播號傳送該輸入訊息至對應該主播號的主播端，俾使該主播端提供的該影音播送含有該主播端影像、該用戶號、該輸入訊息與該效果語音。該主播端提供的該影音播送進一步含有對應該預設語音編號的一角色圖像。該輸入訊息為一文字訊息。其中該語音合成模型是根據監督式類神經網路訓練產生該等特色參數，分別用以產生對應的效果語音。

【0010】為了達到上述的目的，本發明再提供一種影音播送系統，包含：一語音合成模型，包含複數個特色參數；以及一系統後端，允許至少一用戶端與至少一主播端連線登入，並接收該至少一用戶端所傳送的一主播號、一用戶號、一客製化語音與一輸入訊息；其中該語音合成模型將根據該客製化語音選用至少一特色參數，並根據該選用的至少一特色參數與該輸入訊息產生一效果語音，俾使該主播端提供的一影音播送含有一主播端影像、該用戶號與該效果語音。

【0011】根據本發明影音播送系統所提供之一種態樣，該語音合成模型配置於該系統後端或該主播端且該輸入訊息為一文字訊息。

【0012】為解決上述問題，本發明又提供一種影音播送系統，包含：一語音合成模型，包含複數個特色參數；以及一系統後端，允許至少兩用戶端連線登入，並從發話之該用戶端接收一用戶號與一輸入訊息，以及接收一客製化語音與

一預設語音編號其中之一；其中當該系統後端從發話之該用戶端接收該預設語音編號時，該語音合成模型將該預設語音編號對應複數個特色參數的其中之一，並根據該對應的特色參數與該輸入訊息產生一第一效果語音，俾使該用戶號對應的用戶端接收該輸入訊息與該第一效果語音；其中當該系統後端從發話之該用戶端接收該客製化語音時，該語音合成模型將根據該客製化語音選用數個特色參數，並根據該選用的數個特色參數與該輸入訊息產生一第二效果語音，俾使該用戶號對應的用戶端接收該輸入訊息與該第二效果語音。

【0013】本發明進一步提供該影音播送系統之一種影音播送方法，使至少一主播端提供的一影音播送含有一主播端影像，包含：建立一語音合成模型，該語音合成模型包含複數個特色參數；從至少一用戶端接收一主播號、一用戶號與一輸入訊息，且接收一客製化語音與一預設語音編號其中之一；當接收該預設語音編號時，該語音合成模型將該預設語音編號對應至少部分的特色參數，並根據該對應的特色參數與該輸入訊息產生一第一效果語音，俾使該主播號對應的主播端提供的該影音播送含有該主播號對應的該主播端影像與該第一效果語音；以及當接收該客製化語音時，該語音合成模型將根據該客製化語音選用數個特色參數，並根據該選用的數個特色參數與該輸入訊息產生一第二效果語音，俾使該主播號對應的主播端提供的該影音播送含有該主播號對應的該主播端影像與該效果語音。

【0014】本發明進一步提供該影音播送系統之一種影音播送方法，其中該語音合成模型配置於一系統後端或該主播號對應的主播端，且該輸入訊息為一文字訊息。

【0015】本發明進一步提供該影音播送系統之另一種影音播送方法，包含：建立一語音合成模型，該語音合成模型包含複數個特色參數；從至少一用戶端接收一用戶號與一輸入訊息，且接收一客製化語音與一預設語音編號其中之一；當接收該預設語音編號時，該語音合成模型將該預設語音編號對應複數個特色參數的其中之一，並根據該對應的特色參數與該輸入訊息產生一第一效果語音，俾使該用戶號對應的用戶端接收該輸入訊息與該第一效果語音；其中當該系統後端從發話之該用戶端接收該客製化語音時，該語音合成模型將根據該客製化語音選用數個特色參數，並根據該選用的數個特色參數與該輸入訊息產生一第二效果語音，俾使該用戶號對應的用戶端接收該輸入訊息與該第二效果語音。

【0016】綜上所述，根據本發明所實施的影音播送系統及其方法，結合語音合成技術和即時傳輸技術，允許用戶間或複數用戶與主播之間進行即時互動，並支持多角色聲音特徵的語音合成，以提供更加自然和流暢的語音效果，並同時提供可供選擇之客製化語音之訊息交流平台系統。

【圖式簡單說明】

【0017】圖1為本發明影音播送系統之架構圖，係顯示該影音播送系統之系統後端架構。

【0018】圖2A為本發明影音播送系統之再詳細架構圖，係顯示該影音播送系統之系統後端結合主播端與用戶端之架構。

【0019】圖2B為本發明影音播送系統之另一實施例架構圖，其中該語音合成模型位於系統後端。

【0020】 圖2C為本發明影音播送系統之再一實施例架構圖，顯示兩用戶端透過系統後端進行影音播送。

【0021】 圖3A為本發明影音播送系統之語音合成模型之方塊圖。

【0022】 圖3B為本發明影音播送系統之語音合成模型以預設編號取得特色參數之方塊圖。

【0023】 圖3C為本發明影音播送系統之語音合成模型以客製化語音取得特色參數之方塊圖。

【0024】 圖4為本發明影音播送系統使用預設角色選單進行影音播送之流程圖。

【0025】 圖5為本發明影音播送系統使用客製化語音進行影音播送之流程圖。

【0026】 圖6為本發明影音播送系統之使用者介面之一實施例示意圖。

【0027】 圖7為本發明影音播送系統之使用者介面之另一實施例示意圖。

【實施方式】

【0028】 下面藉由特定的具體實施例加以說明本發明之實施方式。

【0029】 首先參考圖1，係顯示本發明影音播送系統之架構圖。本發明影音播送系統之系統後端2允許至少一用戶端1與至少一主播端3連線登入，其中該系統後端2至少包括：接收佇列模組21、雲端儲存伺服器22、模型更新伺服器23。該系統後端2透過接收佇列模組21接收來自用戶端1所傳輸之資料，該資料包括但不限於：欲傳遞之主播端3之主播號、用戶端1之用戶號、輸入文字訊息以及預設語音編號與客製化語音其中之一者，系統後端2之雲端儲存伺服器22紀錄該預

設語音編號與客製化語音其中之一者，並藉由一語音合成模型4根據該預設語音編號與客製化語音其中之一者與輸入文字訊息產生一效果語音，俾使該主播號對應的主播端提供的一影音播送含有該主播端影像與該效果語音。

【0030】 請配合圖1參考圖2A，該圖2A顯示本發明影音播送系統之詳細架構圖。在本發明的一種實施例中，影音播送系統之系統後端2會接受來自一或複數個用戶端1的連線，其中該用戶端1包括：角色選單11、輸出入模組12、使用者介面13、輸出入模組14。用戶可透過用戶端1之使用者介面13操作選擇角色選單11，角色選單11為可選擇式選單，其中包括數個預設角色之預設角色編號，預設角色的語音被提供作為訓練語音合成模型4產生一效果語音的參考語音。用戶透過輸入模組12輸入傳送語音之文字訊息，再透過操作用戶端1之使用者介面13將預設角色編號與文字訊息透過輸出入模組14傳輸至系統後端2之接收佇列模組21。

【0033】 系統後端2允許至少一用戶端1與至少一主播端3連線登入，並透過接收佇列模組21接收來自用戶端1所傳輸之資料，該資料包括但不限於：欲傳遞之主播端3之主播號、用戶端1之用戶號、預設語音編號與輸入文字訊息，系統後端2之雲端儲存伺服器22紀錄該預設語音編號，並將該預設語音編號與輸入文字訊息，透過欲傳遞之主播號，傳送至指定之主播端3。其中該主播端3包括：輸出入模組32、使用者介面33、混音設備34。

【0034】 在本發明的此一實施例中，一語音合成模型4配置於各主播端3，該系統後端2包含模型更新伺服器23。主播端3之輸出入模組32接收系統後端2傳輸之預設語音編號與輸入文字訊息，將其傳送至該語音合成模型4。其中該語音合成模型4是根據類神經網路訓練不同角色的參考語音以產生複數個特色參數，

而選用不同的特色參數可使該語音合成模型4根據輸入的文字訊息產生一對應的效果語音，以下圖3A與3B將進一步說明該語音合成模型4。在本發明的一具體實施例中，當該系統後端2完成該語音合成模型4的訓練後，該模型更新伺服器23可以更新各主播端3的語音合成模型4，而下載最新訓練產生的複數個特色參數至各語音合成模型4，俾使該語音合成模型4可以使用最新訓練產生的複數個特色參數來產生對應的效果語音。

【0035】 例如，語音合成模型4事先以唐老鴨或米老鼠為角色的參考語音進行訓練後，該語音合成模型4可以根據預設語音編號對應選用唐老鴨為角色，並根據輸入文字訊息語音合成出唐老鴨說出文字訊息的效果語音。一主播可控透過使用者介面33進入混音設備34將此效果語音結合主播端影像透過輸出入模組32進行影音播送至用戶端1其中該主播端影像顯示主播實況畫面。請同時參考圖6，此影音播送會在主播實況畫面51包含由選用角色說出文字訊息的效果語音播送511，並進一步顯示表示該選用角色之圖案R2與用戶號，讓主播與用戶皆可看見是由哪個用戶號的用戶，送出由選用角色說出文字訊息的效果語音播送511。

【0036】 在本發明進一步具體實施例中，繼續參考圖2A，其中影音播送系統之系統後端2會接受來自一或複數個用戶端1的連線，用戶可透過用戶端1之使用者介面13操作選擇角色選單11，角色選單11的其中一選項提供用戶可選擇自行上傳客製化語音檔案。該客製化語音檔案可以為各種影音檔案格式，不局限於mp3、mp4等。例如，用戶希望使用周杰倫的聲音產生一效果語音傳送給一主播，但是周杰倫的聲音並非為角色選單11的其中一預設角色，所以不會存在於預設角色編號中，用戶可自行上傳周杰倫的語音檔案。用戶透過輸入模組12輸入傳送

語音之文字訊息，再透過操作用戶端1之使用者介面13將客製化語音檔案與文字訊息透過輸出入模組14傳輸至系統後端2之接收佇列模組21。後續可使用本發明語音合成模型4產生一效果語音來模擬周杰倫說出該文字訊息的聲音。

【0037】 在本發明的進一步實施例中，系統後端2允許至少一用戶端1與至少一主播端3連線登入，並透過接收佇列模組21接收來自用戶端1所傳輸之資料，該資料包括但不限於：欲傳遞之主播端3之主播號、用戶端1之用戶號、客製化語音檔案與輸入文字訊息，系統後端2之雲端儲存伺服器22紀錄該客製化語音檔案，並將該客製化語音檔案與輸入文字訊息，透過欲傳遞之主播號，傳送至指定之主播端3。

【0038】 主播端3之輸出入模組32接收系統後端2傳輸之客製化語音檔案與輸入文字訊息，將其傳送至語音合成模型4。其中該語音合成模型4是根據類神經網路訓練不同角色的參考語音以產生複數個特色參數，而選用不同的特色參數可使該語音合成模型4根據輸入的文字訊息產生一對應的客製化效果語音，以下圖3C將進一步說明該語音合成模型4如何處理客製化語音。一主播可控透過使用者介面33進入混音設備34將此客製化效果語音結合主播端影像透過輸出入模組32進行影音播送至用戶端1其中該主播端影像顯示主播實況畫面。請同時參考圖6，此影音播送會在主播實況畫面51包含由選用角色說出文字訊息的效果語音播送511，並進一步顯示用戶號，讓主播與用戶皆可看見是由哪個用戶號的用戶，送出說出文字訊息的效果語音播送511。

【0039】 請參考圖2B，該圖2B顯示本發明影音播送系統之另一實施例，其中該用戶端1與圖2A所述一致，該語音合成模型4配置於系統後端2，語音合成模型4連接雲端儲存伺服器22存取預設語音編號與輸入文字訊息以進行語音合成，

生成效果語音回傳至雲端儲存伺服器22，以下圖3A與3B將進一步說明該語音合成模型4。雲端儲存伺服器22再透過欲傳遞之主播號，將該效果語音與輸入文字訊息傳送至指定之主播端3。一主播可控透過使用者介面33進入混音設備34將效果語音結合主播端影像透過輸出入模組32進行影音播送至用戶端1，其中該主播端影像顯示主播實況畫面。

【0040】 在本發明進一步具體實施例中，繼續參考圖2B，其中該用戶端1與圖2A所述一致，該語音合成模型4位於系統後端2，語音合成模型4連接雲端儲存伺服器22存取客製化語音檔案，與輸入文字訊息進行語音合成，生成效果語音傳輸至雲端儲存伺服器22，以下圖3C將進一步說明該語音合成模型4如何處理客製化語音。雲端儲存伺服器22再透過欲傳遞之主播號，將效果語音傳送至指定之主播端3。一主播可控透過使用者介面33進入混音設備34將效果語音結合主播端影像透過輸出入模組32進行影音播送至用戶端1其中該主播端影像顯示主播實況畫面。

【0041】 請參考圖2C，圖2C顯示本發明影音播送系統之用戶端間透過雲端伺服器24接收效果語音之一實施例示意圖，影音播送系統之系統後端2會接受來自一或複數個用戶端1的連線，用戶可透過用戶端1之使用者介面13操作選擇角色選單11，角色選單11中為可選擇式選單，其中包括數個預設角色之預設角色編號。用戶透過輸入模組12輸入傳送語音之文字訊息，再透過操作用戶端1之使用者介面13將預設角色編號與文字訊息透過輸出入模組14傳輸至系統後端2之雲端伺服器24。

【0043】 系統後端2之雲端伺服器24接收到用戶端1上傳之預設角色編號與輸入的文字訊息，語音合成模型4將預設角色編號對應特色參數提取並將輸入

文字訊息進行語音合成，生成效果語音，以下圖3A與3B將進一步說明該語音合成模型4。在本發明一具體實施例中，當該系統後端2完成該語音合成模型4的訓練後，該模型更新伺服器23可以更新該語音合成模型4，而下載最新訓練產生的複數個特色參數至該語音合成模型4，俾使該語音合成模型4可以使用最新訓練產生的複數個特色參數來產生對應的效果語音。

【0044】 在本發明進一步具體實施例中，繼續參考圖2C，影音播送系統之系統後端2會接受來自一或複數個用戶端1的連線，用戶可透過用戶端1之使用者介面13操作選擇角色選單11，角色選單11中用戶可選擇自行上傳客製化語音，其中該選擇可供用戶選擇之客製化語音檔案。該客製化語音檔案可以為各種影音檔案格式，不局限於mp3、mp4等。用戶透過輸入模組12輸入傳送語音之文字訊息，再透過操作用戶端1之使用者介面13將客製化語音檔案與文字訊息透過輸出入模組14傳輸至系統後端2之雲端伺服器24。

【0045】 系統後端2之雲端伺服器24接收到用戶端1上傳之客製化語音檔案與輸入的文字訊息，語音合成模型4連接雲端儲存伺服器24存取客製化語音檔案，與輸入文字訊息進行語音合成，生成效果語音，以下圖3C將進一步說明該語音合成模型4如何處理客製化語音。另一用戶可透過系統後端2之雲端伺服器24取得該效果語音，並在用戶端1播放與使用。

【0046】 本發明語音合成模型4的架構可參考至圖3A所示方塊圖，在本發明的一具體實施例中，語音合成模型4的組成方塊包含音標嵌入、依一語者狀況的音標編碼器41、差異化適配器42以及依該語者狀況的Mel解碼器43，其中，該音標編碼器41與Mel解碼器43的語者狀況是以語音特徵提取44根據不同角色所提取的複數個特色參數所表示。本發明利用不同角色的參考語音來訓練語音合

成模型4及其語音特徵提取(Speech Extraction)44，該語音合成模型4可基於監督式或非監督式類神經網路來訓練。訓練後語音合成模型4的語音特徵提取44，可以根據預設角色編號指定一角色來提取出對應的複數個特色參數，或根據非預設角色之一客製化語音來提取出選用的複數個特色參數。

【0047】本發明語音合成模型4的各組成方塊可由根據Yihan Wu等人於2022年4月發表之論文所揭露內容來實施，該論文名稱為「AdaSpeech 4: Adaptive Text to Speech in Zero-Shot Scenarios」，公開網址：https://www.isca-speech.org/archive/interspeech_2022/index.html。該論文揭露一種語音合成技術，可利用參考語音(Reference Speech)來訓練一語音合成模型，該語音合成模型是基於類神經網路來訓練。

【0048】請繼續參考圖3A，在本發明的一具體實施例中，本發明利用不同角色的參考語音來訓練語音合成模型4，該語音合成模型4可基於監督式或非監督式類神經網路來訓練。訓練後語音合成模型4的語音特徵提取44，可以根據不同角色的參考語音提取出複數個特色參數。

【0049】當本發明系統使用預設語音編號與輸入文字訊息時，參考圖3B，該預設語音編號與輸入文字訊息的音標輸入至本發明語音合成模型4，透過預先訓練的語音特徵提取44將指定語音編號所對應角色的特色參數用來表示音標編碼器41與Mel解碼器43的語者狀況，該預先訓練的語音合成模型4會根據文字訊息的音標(phoneme)經音標嵌入，透過音標編碼器41進行音標編碼將輸入文字訊息的音標與特色參數編碼合成。接著，透過差異化適配器42進行音軌適配性調整，最後透過Mel解碼器43依該語者狀況解碼生成一效果語音，而該效果語音為指定語音編號所對應角色說出該文字訊息的語音。

【0050】接著參考圖3C，圖3C為當用戶使用客製化檔案與輸入文字訊息輸入至本發明語音合成模型4時，該語音合成模型4會透過語音特徵提取44對該客製化語音進行前處理與內差運算之特徵提取，透過前處理與內差運算的客製化語音將選用已訓練完成所產生之一或複數個特色參數來表示音標編碼器41與Mel解碼器43的語者狀況，該預先訓練的語音合成模型4會根據文字訊息的音標(phoneme)經音標嵌入，透過音標編碼器41進行音標編碼將輸入文字訊息的音標與特色參數編碼合成。接著，透過差異化適配器42進行音軌適配性調整，最後透過Mel解碼器43依該語者狀況解碼生成一效果語音，而該效果語音為客製化語音者說出該文字訊息的語音。

【0051】接著看到圖4，可配合參考圖2A，圖4顯示本發明使用預設之角色編號進行影音播送之一具體實施例流程圖。用戶於用戶端1透過使用者介面13接收一角色選單11之一預設語音編號與輸入模組12之一輸入訊息(S11)，例如：文字訊息，輸出入模組14將主播號 、用戶號、該預設語音編號與該輸入訊息輸出至系統後端(S12)，此兩步驟完成於用戶端1。

【0052】接著系統後端2之接收佇列模組21接收主播號、用戶號、該預設語音編號與該輸入訊息(S13)，雲端儲存伺服器22將接收之該預設語音編號與該輸入訊息，並傳送至一主播端3之語音合成模型4(S14)，此兩步驟完成於系統後端2。

【0053】接著主播端3之語音合成模型4以該預設語音編號對應角色的特色參數表示音標編碼器41與Mel解碼器43的語者狀況，並與該輸入訊息進行語音合成，而生成一效果語音(S15)。主播端3透過混音設備34將效果語音進行混音(S18)，混音設備會將該效果語音伴隨影音畫面一同輸出，如選擇的禮物等影音

效果，混音後之該效果語音透過輸出入模組32進行影音播送(S19)，此三步驟完成於主播端3。

【0054】當語音合成模型4接收預設語音編號並且以該預設語音編號對應角色的特色參數表示一語者狀況，並與該輸入訊息進行語音合成，而生成一效果語音時，若語音生成之參數有所優化，語音合成模型4會將優化之該特色參數同步更新至系統後端2之模型更新伺服器23(S16)，並且模型更新伺服器23會更新該特色參數，以利下次進行該對應之特色參數(S17)。

【0055】接著看到圖5，可配合參考圖2A，圖5顯示本發明使用客製化語音檔案進行影音播送之另一具體實施例流程圖。用戶於用戶端1透過使用者介面13接收一角色選單11之一客製化語音檔案與輸入模組12之一輸入訊息(S21)，輸出入模組14將主播號、用戶號、客製化語音檔案與輸入訊息輸出至系統後端(S22)，此兩步驟完成於用戶端1。

【0056】接著系統後端2之接收佇列模組21接收主播號、用戶號、客製化檔案與輸入訊息(S23)，雲端儲存伺服器22將接收之客製化語音檔案與該輸入訊息，並傳送至一主播端3之語音合成模型4(S24)，此兩步驟完成於系統後端2。

【0057】接著主播端3之語音合成模型4會透過語音特徵提取44對該客製化語音進行前處理與內差運算之特徵提取，透過前處理與內差運算的客製化語音將選用已訓練完成所產生之一或複數個特色參數來表示音標編碼器41與Mel解碼器43的語者狀況，該預先訓練的語音合成模型4會根據文字訊息的音標(phoneme)經音標嵌入，透過音標編碼器41進行音標編碼將輸入文字訊息的音標與特色參數編碼合成。接著，透過差異化適配器42進行音軌適配性調整，最後透過Mel解碼器43依該語者狀況解碼生成一效果語音(S25)，並透過混音設備34將效果語音進行混音(S26)，混音設備會將該效果語音伴隨影音畫面一同輸出，如選

擇的禮物等影音效果，混音後之該效果語音透過輸出入模組32進行影音播送(S27)，此三步驟完成於主播端3。

【0058】在本發明一具體實施例中，透過客製化語音檔案選用特色參數並不會被儲存於語音合成模型4中成為預設語音，每一次進行客製化效果語音皆須重新進行語音特徵之提取步驟。

【0059】接著看到圖6，圖6顯示本發明使用者介面之一具體實施例示意圖。其中畫面分為三區域，分別為左側之輔助列表52、中間之主播實況畫面51與右邊之聊天室畫面53。中間之主播實況畫面51會撥放主播實況之互動畫面與本發明之效果語音之效果語音播送511，該效果語音播送511包括效果語音與其對應之文字、選擇之選單角色之頭像與送出效果語音之用戶號，與相對應的禮物圖式等。

【0060】右側之聊天室畫面53會顯示複數個用戶與主播文字聊天之內容，在一特定實施例中，所送出的效果語音播送511的文字亦會在聊天室畫面53展示，並搭配對應之特殊顏色、外觀等。聊天室畫面53下方包含角色選單531與文字輸入模組532，圖6僅展示一種特定實施例中之角色選單531之呈現，其亦可以選單式、彈出式視窗等方式進行設計。例如：角色選單531可包含複數個頭像圖案R1、R2…，便於用戶識別選用。

【0061】左側之輔助列表52中會存放著相關互動功能等列表，左上方提供各種功能的選單列表54，每日任務列表521可設計每日需達成之任務，如發送五次對話等增加用戶與主播的互動性，與特定節日可舉辦活動，並登載在活動列表522中。左下方為其他主播清單531，可藉由畫面點選切換至不同主播的直播間，進行互動。

【0062】請參考圖7，圖7顯示本發明使用者介面之另一實施例示意圖，特別如手機用戶間傳遞訊息。本實施例中畫面顯示的為聊天畫面63，其中用戶位於

聊天畫面63之右側，而對象用戶位於聊天畫面63的左側，可見於圖7，在一特定實施例中，對象用戶傳送了文字訊息「我愛你」，於畫面右側顯示用戶頭像64，而用戶透過角色選單631選擇R2的角色，並同時透過文字輸入模組632發送「我愛妳」的文字，透過本發明之影音播送系統，該預設語音編號與輸入文字訊息進行一效果語音輸出，進行效果語音播送611給對象用戶。

【符號說明】

【0063】

1 用戶端

11 角色選單

12 輸入模組

13 使用者介面

14 輸出入模組

2 系統後端

21 接收佇列模組

22 雲端儲存伺服器

23 模型更新伺服器

24 雲端伺服器

3 主播端

32 輸出入模組

33 使用者介面

34 混音設備

4 語音合成模型

41 音標編碼器

42 差異化適配器

43 Mel解碼器

5 使用者介面

51 主播實況畫面

511 效果語音播送

52 輔助列表

521 每日任務清單

522 活動列表清單

523 其他主播清單

53 聊天室畫面

531 角色選單

532 文字輸入模組

54 選單列表

6 使用者介面

611 效果語音播送

63 聊天畫面

631 角色選單

632 文字輸入模組

64 用戶頭像

【發明申請專利範圍】

【請求項1】一種影音播送系統，包含：

一語音合成模型，包含複數個特色參數；以及

一系統後端，允許至少一用戶端透過一用戶端輸出入模組與至少一主播端透過一主播端輸出入模組連線登入，並接收該至少一用戶端所傳送的一主播號、一用戶號、一預設語音編號與一輸入訊息，其中該預設語音編號對應一角色選單所提供複數個角色之一；

其中該語音合成模型將該預設語音編號對應至少部分的特色參數，並根據該對應的特色參數與該輸入訊息產生一效果語音，俾使該主播端提供的一影音播送含有一主播端影像、該用戶號與該效果語音。

【請求項2】如請求項1所述之影音播送系統，其中該語音合成模型配置於該系統後端。

【請求項3】如請求項2所述之影音播送系統，其中該系統後端透過該主播端輸出入模組根據該主播號傳送該用戶號與該效果語音至對應該主播號的主播端，俾使該主播端提供的該影音播送含有該主播端影像、該用戶號與該效果語音。

【請求項4】如請求項3所述之影音播送系統，其中該系統後端透過該主播端輸出入模組根據該主播號傳送該輸入訊息至對應該主播號的主播端，俾使該主播端提供的該影音播送含有該主播端影像、該用戶號、該輸入訊息與該效果語音。

【請求項5】如請求項1所述之影音播送系統，其中該語音合成模型配置於該主播端。

【請求項6】 如請求項5所述之影音播送系統，其中該系統後端透過該主播端輸出入模組根據該主播號傳送該用戶號與該預設語音編號至對應該主播號的主播端，俾使該主播端提供的該影音播送含有該主播端影像、該用戶號與該效果語音。

【請求項7】 如請求項6所述之影音播送系統，其中該系統後端透過該主播端輸出入模組根據該主播號傳送該輸入訊息至對應該主播號的主播端，俾使該主播端提供的該影音播送含有該主播端影像、該用戶號、該輸入訊息與該效果語音。

【請求項8】 如請求項1所述之影音播送系統，其中該主播端提供的該影音播送進一步含有對應該預設語音編號的一角色圖像。

【請求項9】 如請求項1所述之影音播送系統，其中該輸入訊息為一文字訊息。

【請求項10】 如請求項1所述之影音播送系統，其中該語音合成模型是根據類神經網路訓練產生該等特色參數，用以產生對應的效果語音。

【請求項11】 一種影音播送系統，包含：

一語音合成模型，包含複數個特色參數；以及

一系統後端，允許至少一用戶端透過一用戶端輸出入模組與至少一主播端透過一主播端輸出入模組連線登入，並接收該至少一用戶端所傳送的一主播號、一用戶號、一客製化語音與一輸入訊息；

其中該語音合成模型將根據該客製化語音選用至少一特色參數，並根據該選用的至少一特色參數與該輸入訊息產生一效果語音，俾使該主播端提供的一影音播送含有一主播端影像、該用戶號與該效果語音。

【請求項12】如請求項11所述之影音播送系統，其中該語音合成模型配置於該系統後端或該主播端。

【請求項13】如請求項11所述之影音播送系統，其中該輸入訊息為一文字訊息。

【請求項14】一種影音播送系統，包含：

一語音合成模型，包含複數個特色參數；以及

一系統後端，允許至少兩用戶端透過一用戶端輸出入模組連線登入，並從發話之該用戶端接收一用戶號與一輸入訊息，以及接收一客製化語音與一預設語音編號其中之一；

其中當該系統後端從發話之該用戶端接收該預設語音編號時，該語音合成模型將該預設語音編號對應至少部分的特色參數，並根據該對應的特色參數與該輸入訊息產生一第一效果語音，俾使該用戶號對應的用戶端接收該輸入訊息與該第一效果語音；

其中當該系統後端從發話之該用戶端接收該客製化語音時，該語音合成模型將根據該客製化語音選用數個特色參數，並根據該選用的數個特色參數與該輸入訊息產生一第二效果語音，俾使該用戶號對應的用戶端接收該輸入訊息與該第二效果語音。

【請求項15】如請求項14所述之影音播送系統，其中該語音合成模型配置於該系統後端。

【請求項16】如請求項14所述之影音播送系統，其中該輸入訊息為一文字訊息。

【請求項17】一種影音播送方法，使至少一主播端透過一主播端輸出入模組提供的一影音播送含有一主播端影像，包含：

建立一語音合成模型，該語音合成模型包含複數個特色參數；

從至少一用戶端之一用戶端輸出入模組接收一主播號、一用戶號與一輸入訊息，且接收一客製化語音與一預設語音編號其中之一；

當接收該預設語音編號時，該語音合成模型將該預設語音編號對應至少部分的特色參數，並根據該對應的特色參數與該輸入訊息產生一第一效果語音，俾使該主播號對應的主播端提供的該影音播送含有該主播號對應的該主播端影像與該第一效果語音；以及

當接收該客製化語音時，該語音合成模型將根據該客製化語音選用數個特色參數，並根據該選用的數個特色參數與該輸入訊息產生一第二效果語音，俾使該主播號對應的主播端提供的該影音播送含有該主播號對應的該主播端影像與該效果語音。

【請求項18】如請求項17所述之影音播送方法，其中該語音合成模型配置於一系統後端或該主播號對應的主播端。

【請求項19】如請求項17所述之影音播送方法，其中該輸入訊息為一文字訊息。

【請求項20】一種影音播送方法，包含：

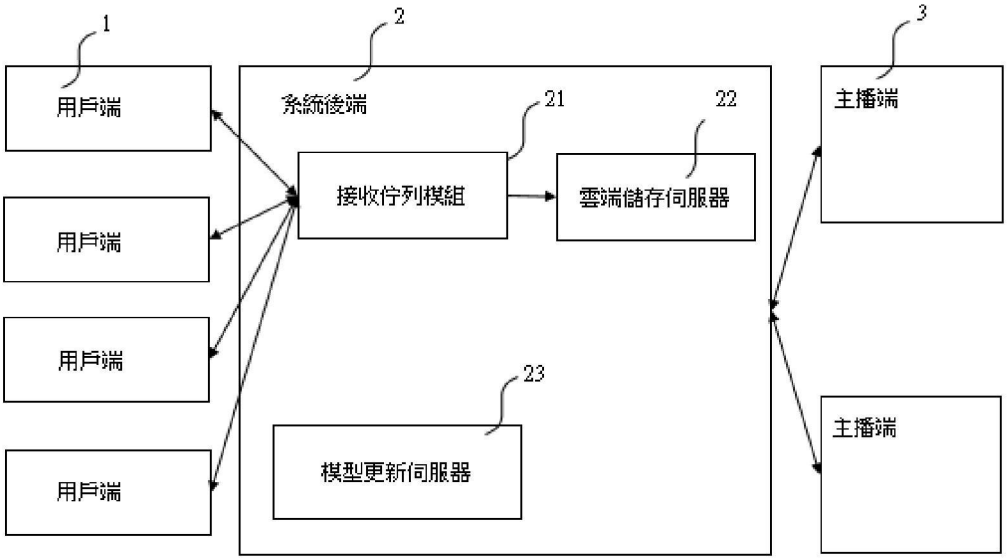
建立一語音合成模型，該語音合成模型包含複數個特色參數；

從至少一用戶端之一用戶端輸出入模組接收一用戶號與一輸入訊息，且接收一客製化語音與一預設語音編號其中之一；

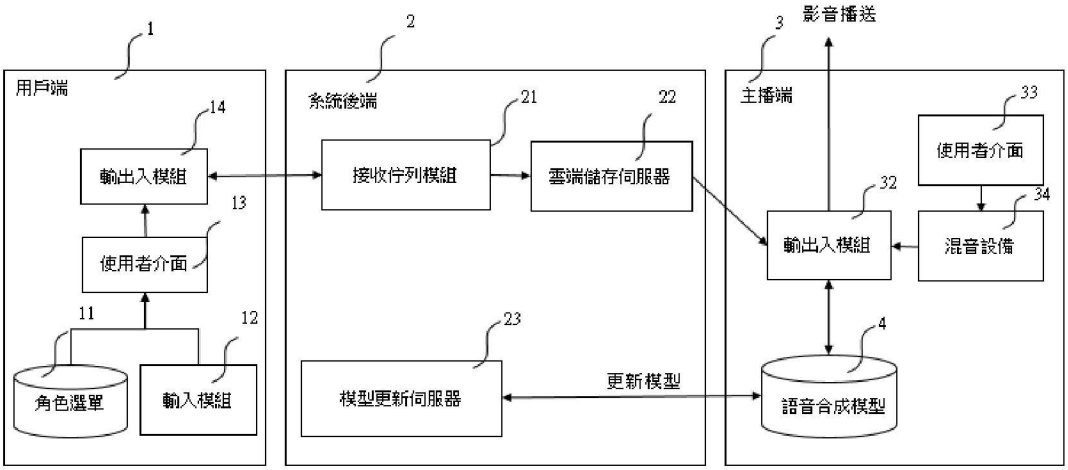
當接收該預設語音編號時，該語音合成模型將該預設語音編號對應至少部分的特色參數，並根據該對應的特色參數與該輸入訊息產生一第一效果語音，俾使該用戶號對應的用戶端接收該輸入訊息與該第一效果語音；

其中當該系統後端從發話之該用戶端接收該客製化語音時，該語音合成模型將根據該客製化語音選用數個特色參數，並根據該選用的數個特色參數與該輸入訊息產生一第二效果語音，俾使該用戶號對應的用戶端接收該輸入訊息與該第二效果語音。

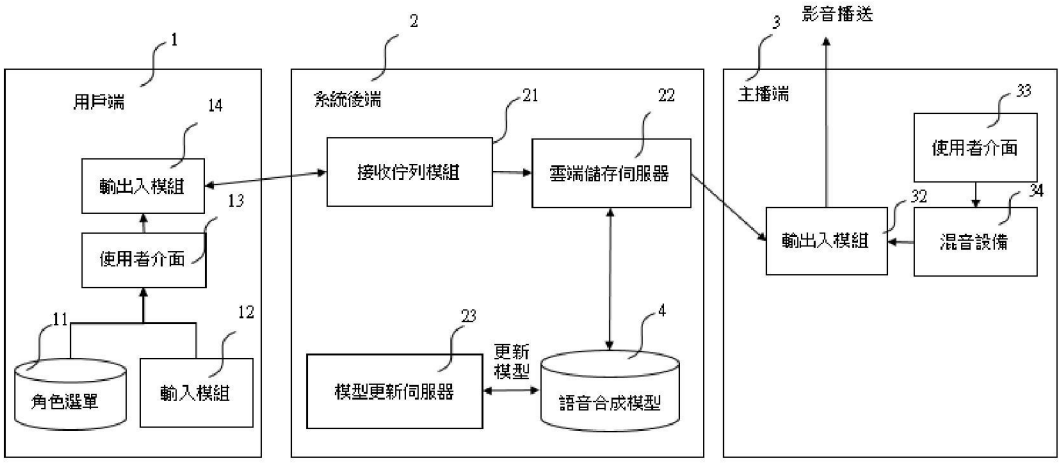
【發明圖式】


1
用戶端
用戶端
用戶端
用戶端
2
系統後端
21
接收佇列模組
22
雲端儲存伺服器
23
模型更新伺服器
3
主播端
主播端

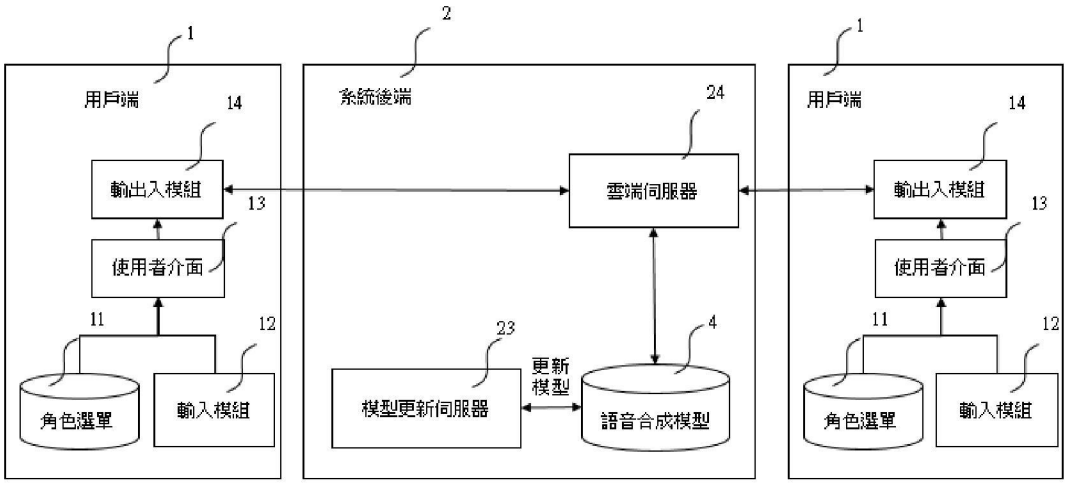
【圖 1】


1
用戶端
14
輸出入模組
13
使用者介面
11
12
角色選單
輸入模組
2
系統後端
21
22
接收佇列模組
雲端儲存伺服器
23
模型更新伺服器
更新模型
3
影音播送
主播端
33
使用者介面
32
34
輸出入模組
混音設備
4
語音合成模型

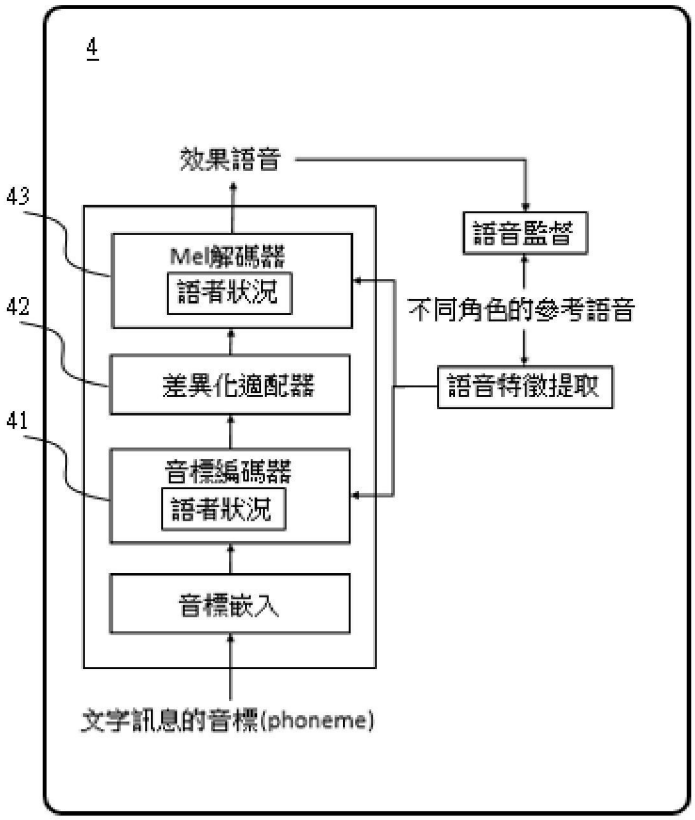
【圖 2A】


1
用戶端
14
輸出入模組
13
使用者介面
11
12
角色選單
輸入模組
2
系統後端
21
接收佇列模組
22
雲端儲存伺服器
23
模型更新伺服器
更新模型
4
語音合成模型
3
影音播送
主播端
33
使用者介面
32
34
輸出入模組
混音設備

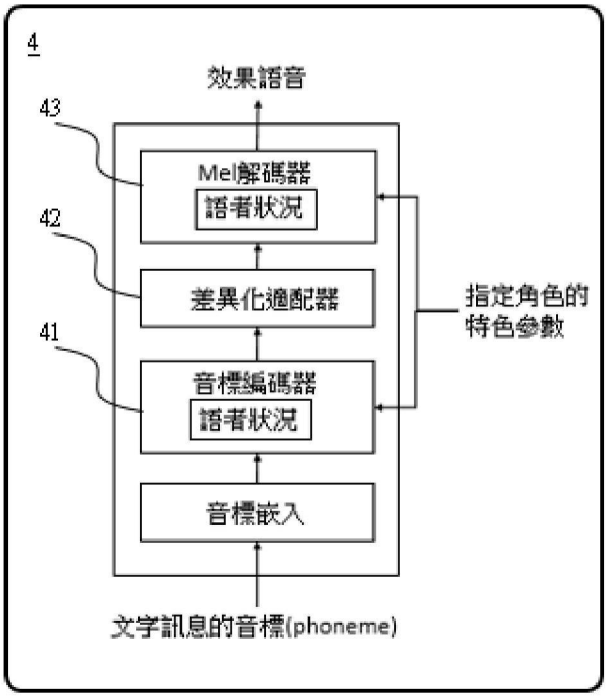
【圖 2B】


1
用戶端
14
輸出入模組
13
使用者介面
11
12
角色選單
輸入模組
2
系統後端
24
雲端伺服器
23
模型更新伺服器
更新模型
4
語音合成模型
1
用戶端
14
輸出入模組
13
使用者介面
11
12
角色選單
輸入模組

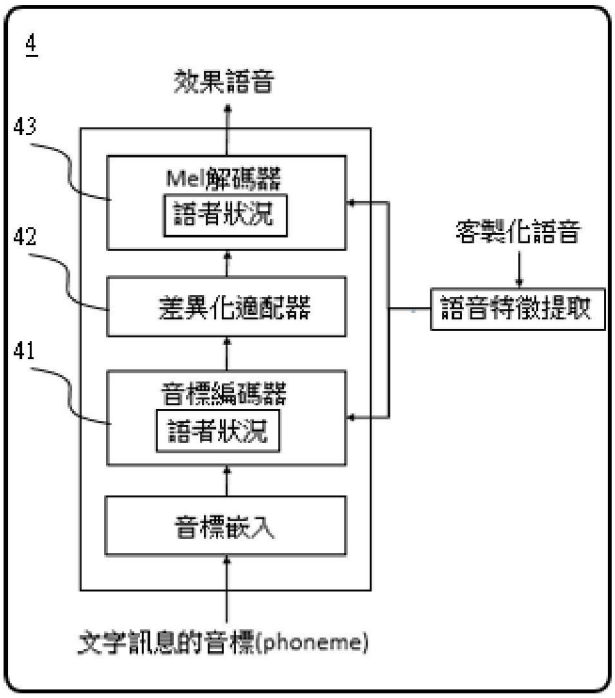
【圖 2C】


4
效果語音
43
Mel解碼器
語者狀況
42
差異化適配器
41
音標編碼器
語者狀況
音標嵌入
文字訊息的音標(phoneme)
語音監督
不同角色的參考語音
語音特徵提取

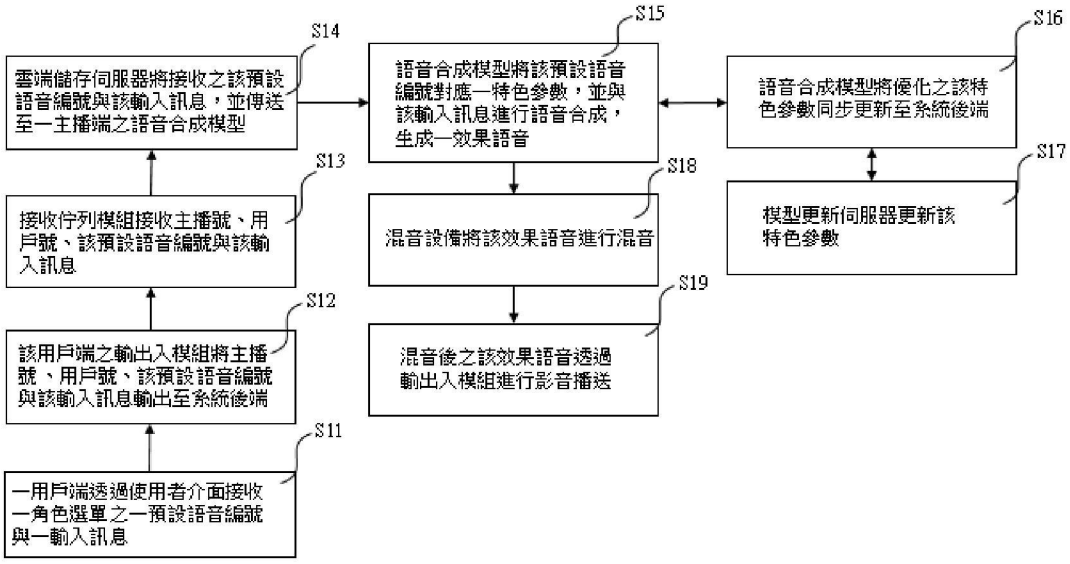
【圖 3A】


4
效果語音
43
Mel解碼器
語者狀況
42
差異化適配器
41
音標編碼器
語者狀況
音標嵌入
文字訊息的音標(phoneme)
指定角色的
特色參數

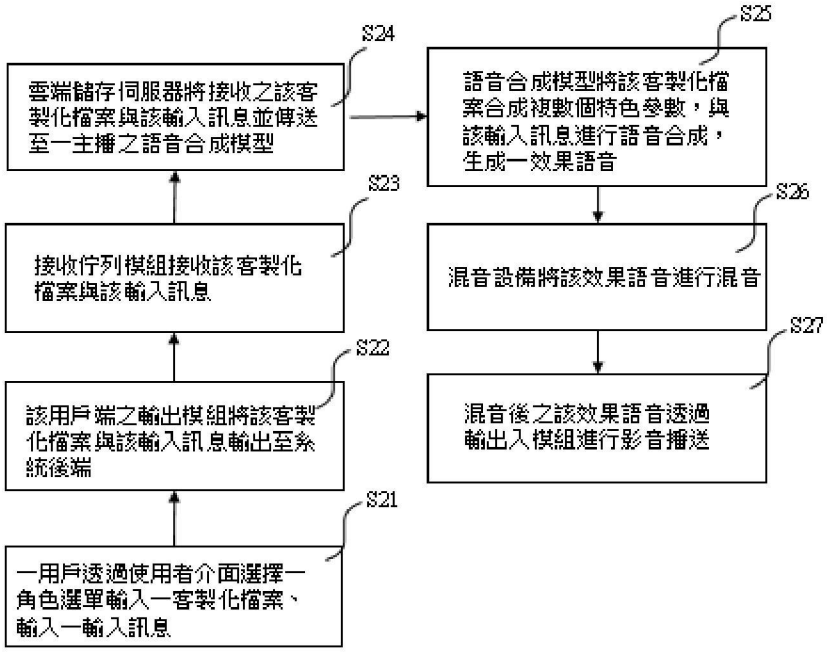
【圖 3B】


4
效果語音
43
Mel解碼器
語者狀況
42
差異化適配器
41
音標編碼器
語者狀況
音標嵌入
文字訊息的音標(phoneme)
客製化語音
語音特徵提取

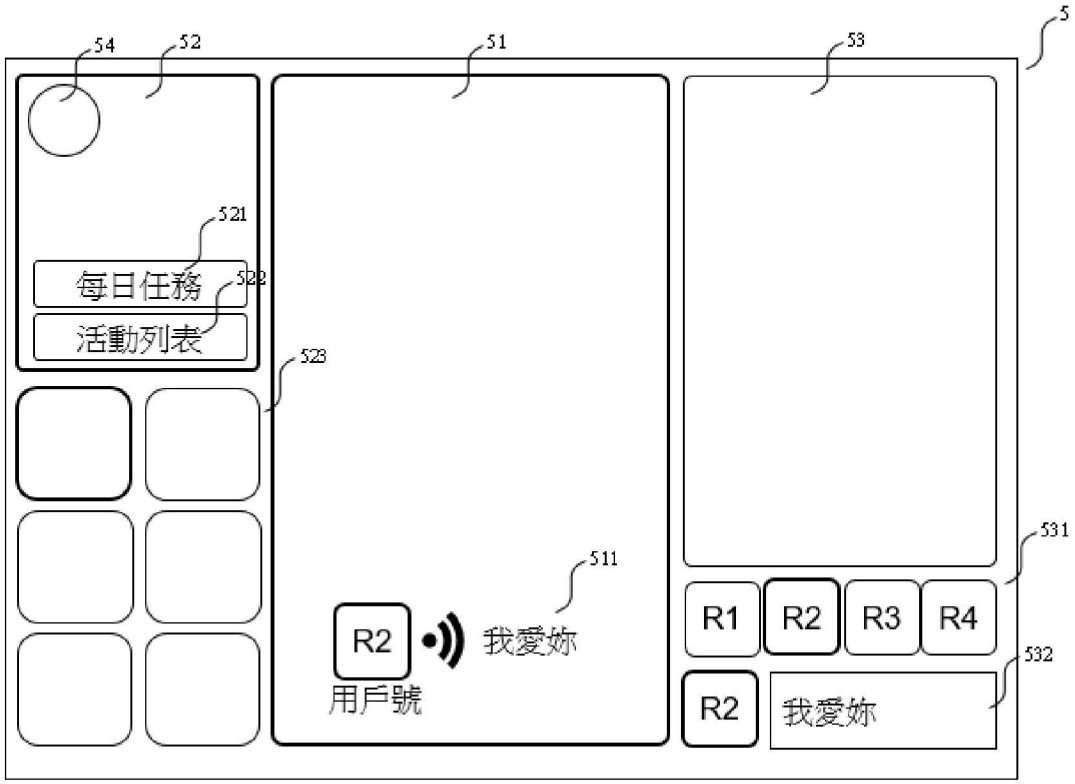
【圖 3C】


S11
一用戶端透過使用者介面接收一角色選單之一預設語音編號與一輸入訊息
S12
該用戶端之輸出入模組將主播號、用戶號、該預設語音編號與該輸入訊息輸出至系統後端
S13
接收佇列模組接收主播號、用戶號、該預設語音編號與該輸入訊息
S14
雲端儲存伺服器將接收之該預設語音編號與該輸入訊息，並傳送至一主播端之語音合成模型
S15
語音合成模型將該預設語音編號對應一特色參數，並與該輸入訊息進行語音合成，生成一效果語音
S16
語音合成模型將優化之該特色參數同步更新至系統後端
S17
模型更新伺服器更新該特色參數
S18
混音設備將該效果語音進行混音
S19
混音後之該效果語音透過輸出入模組進行影音播送

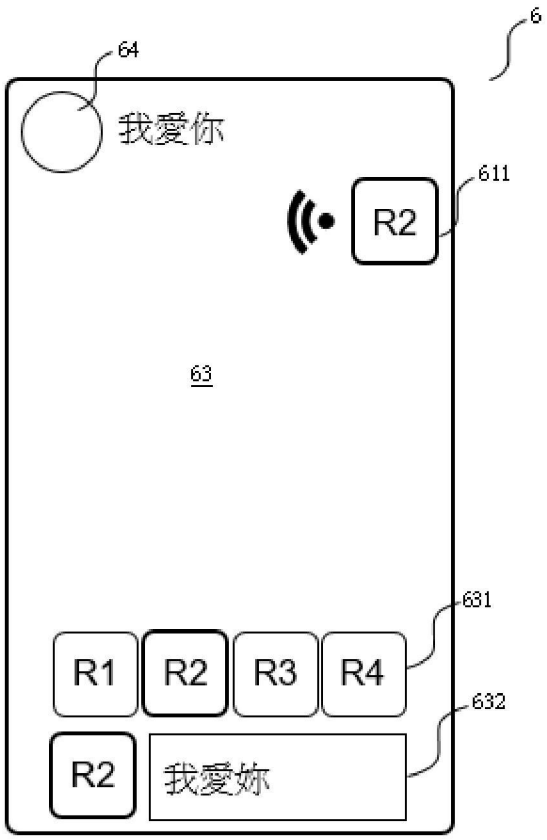
【圖 4】


S21
一用戶透過使用者介面選擇一角色選單輸入一客製化檔案、輸入一輸入訊息
S22
該用戶端之輸出模組將該客製化檔案與該輸入訊息輸出至系統後端
S23
接收佇列模組接收該客製化檔案與該輸入訊息
S24
雲端儲存伺服器將接收之該客製化檔案與該輸入訊息並傳送至一主播之語音合成模型
S25
語音合成模型將該客製化檔案合成複數個特色參數，與該輸入訊息進行語音合成，生成一效果語音
S26
混音設備將該效果語音進行混音
S27
混音後之該效果語音透過輸出入模組進行影音播送

【圖 5】


5
51
52
53
54
521
每日任務
522
活動列表
523
511
R2
我愛妳
用戶號
531
R1
R2
R3
R4
532
R2
我愛妳

【圖 6】



【圖 7】