The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 35-48

歌詞演唱錯誤偵測

Automatic Sung Lyrics Verification

孔祥勳　Shiang-Shiun Kung
國立台北科技大學電子系
Department of Electronic Engineering
National Taipei University of Technology
squarprince@gmail.com

馬勤皓　Cin-Hao Ma
國立台北科技大學電子系
Department of Electronic Engineering
National Taipei University of Technology
t101419012@ntut.edu.tw

沈信甫　Sin-Fu Shen
國立台北科技大學電子系
Department of Electronic Engineering
National Taipei University of Technology
squarprince@gmail.com

蕭博元　Po-Yuan Hsiao
國立台北科技大學電子系
Department of Electronic Engineering
National Taipei University of Technology
ccmomcc@gmail.com

蔡偉和　Wei-Ho Tsai
國立台北科技大學電子系
Department of Electronic Engineering
National Taipei University of Technology
encorew56527@gmail.com

摘要

本研究嘗試發展一種唱詞確認系統，以自動判斷演唱者是否唱錯歌詞。雖然直覺上，唱詞確認相似於語句確認問題，可以利用語音辨認上所使用的方法來處理，但由於歌唱聲音訊號就像語音訊號的伸縮、變形過後版本，我們發現直接利用語句確認進行唱詞確認的效果並不如預期。有鑑於歌唱時常因母音被拉長若干倍而造成與說話時的訊號相差甚多，我們試圖找出歌唱中的母音位置，並對其長度壓縮或裁剪，使其接近語音訊號，以

使語句確認方法較能正常運作。經實驗結果顯示，透過母音長度壓縮或裁剪可大幅提升唱詞判斷的正確率。

Abstract

This study proposes a sung lyrics verification system for detecting if the lyrics sung by a performer are incorrect and further pointing out the potential mistake that the performer made. In essence, sung lyrics verification is similar to the problem of speech utterance verification in the speech recognition research community, and therefore the techniques in the letter can be applied to the former. However, our preliminary experiment found that a speech utterance verification system cannot handle singing data well, mainly because of the significant differences between singing and speech. To tackle this problem, we develop two strategies, respectively, from a signal processing perspective and from a model processing perspective. In the signal processing, recognizing that the vowels are often lengthened during singing, we propose vowel shrinking and vowel decimation to adjust the length of a vowel in singing to a normal length in speaking. In the model processing, we include a duration model concept in the acoustic modeling to reduce the differences between singing and speech. Our experiments show that the proposed methods can improve the performance of the sung lyrics verification to 72% and 90% accuracy using vowel shrinking, vowel decimation, and duration model approach, respectively, compared to 63% accuracy obtained with the baseline speech utterance verification system.

關鍵詞：唱詞確認，語句確認，母音壓縮，母音裁剪

Keywords: Singing Evaluation, Sung Lyrics Verification, Vowel Shrinking, Vowel Decimation, Duration Model

一、緒論

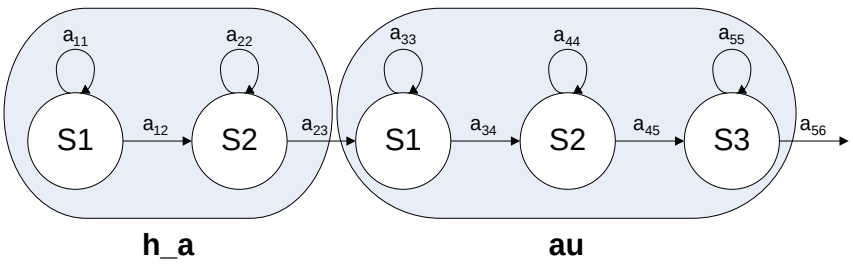
唱歌是人類的天賦，但要唱得好聽或有技巧則需要尋求管道來精進。通常，我們藉由別人口中得知自己唱歌是否好聽，甚至是聘請歌唱老師進行指導。然而，經由歌唱老師指導雖能夠讓學習者了解自身歌唱技巧上的缺點並加以改進，但並非所有人都有能力聘請專業人士來指導。因此，若有一套系統能夠在任何時間或是任何地點提供如專業人士般的指導，指出使用者在唱歌時所犯的錯誤，讓其提升歌唱實力，將會是一大助益。

綜觀目前市面上的卡拉 OK 伴唱系統中，具有自動歌唱評分功能的不在少數，但大多數仍以娛樂效果為主，並沒有實際評分或指導效果。在學術研究中，最完整的卡拉 OK 歌唱評分系統[1]採用「音高」、「動態音量」與「和諧度」三項依據進行評分，但卻忽略了「歌詞」這項依據。當演唱者沒有唱在歌曲的節奏上，或是唱成不同的字詞，便會產生唱錯歌詞的情形。並且，在真實歌唱比賽中，歌詞在評審評分時也佔了相當的比重。因此，「歌詞」是其中一項不可忽略的評分依據。有鑑於目前尚未有人針對「唱詞確認」進行探討，本研究嘗試評估自動唱詞確認的可行性。

二、應用語句確認系統於唱詞確認

一開始，本研究建立了一個以隱藏式馬可夫模型(Hidden Markov Model)為基礎的中文語句確認系統[2]，評估其用於中文唱詞確認問題的可能性與效能。我們透過 Hidden Markov Model Toolkit (HTK) [3]來實現語句確認系統，其中聲學模型是以次音節(Sub-syllable)為單位，共使用一百五十一個聲學模型(含靜音)，每一個模型皆為混合高斯機率密度之連續型隱藏式馬可夫模型。而用以訓練產生該模型的語音資料是 TCC-300 [4]。

考慮中文基本音節約有 411 個，我們利用次音節模型拼出此 411 個音節。舉例來說，圖一為中文音「好」的聲學模型圖，它包含子音模型「h_a」與母音模型「au」，其中「h_a」模型使用了兩個狀態，「au」模型則使用了三個狀態來描述，而 $\{a_{11}, a_{12}, a_{22}, a_{23}, a_{33}, a_{34}, a_{44}, a_{45}, a_{55}, a_{56}\}$ 為狀態轉移機率。

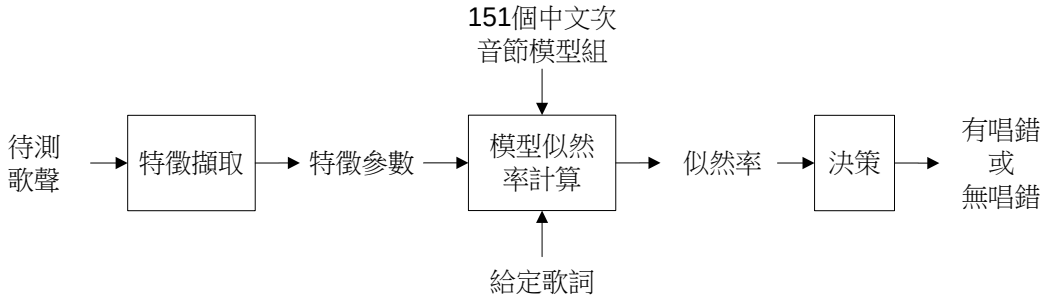


圖一、中文音「好」之聲學模型圖

如圖二所示，給定一段歌詞後，我們依其歌詞的次音節發音串接出模型Λ。則當一段歌唱聲音受測時，系統將其時域訊號轉成特徵參數 **O**，並利用維特比演算法(Viterbi Algorithm)計算特徵參數 **O** 相對於模型Λ的對數似然率 ln Pr(**O**|Λ)。理論上，似然率越大，代表該歌聲所唱的歌詞越正確；似然率越小，代表該歌聲所唱的歌詞越不正確。但為了量化正確性成為可判斷的數值，我們需要有一個基準似然率來做比較，亦即進行似然率的正規化。本論文採用類似文獻[5]所討論的方法，透過語音辨認法判斷受測歌聲 **O** 最可能是唱甚麼，例如Λ^*為維特比演算法所求出之最佳路徑所對應的模型串，則系統根據方程式(1)所得之分數判斷受測歌聲 **O** 是否唱錯

$$\text{分數} = \ln \Pr(\mathbf{O} \mid \Lambda) - \ln \Pr(\mathbf{O} \mid \Lambda^*) \begin{array}{c} \text{正確} \\ > \\ \leq \\ \text{不正確} \end{array} \delta \qquad (1)$$

其中 δ 為可調之臨界值(Threshold)。



圖二、使用語句確認系統進行唱詞確認

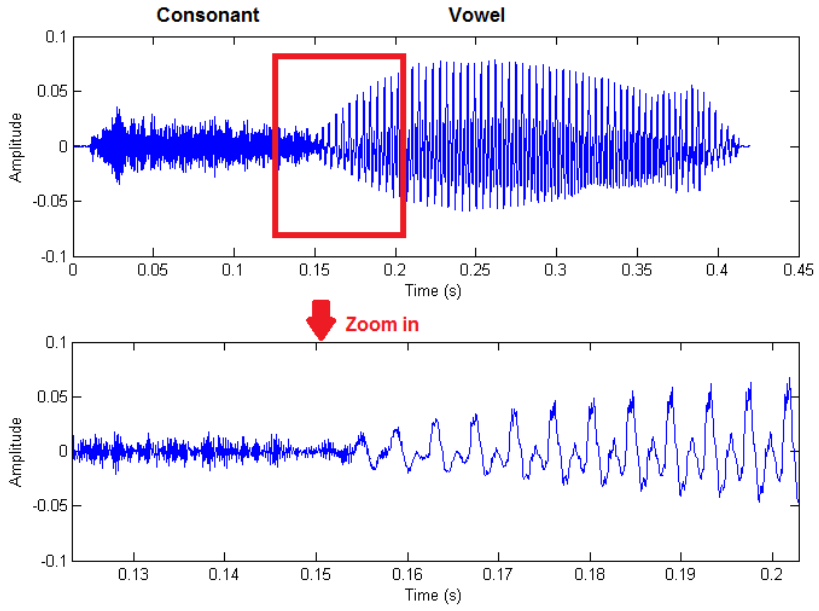
三、針對歌唱訊號特性來改善語句確認系統

由於歌唱聲音訊號可以視為語音訊號的伸縮、變形過後版本，我們發現利用上述語句確認方式進行唱詞確認結果並不如理想。為此，本研究從聲音訊號處理進行改善嘗試。主要想法是考慮歌唱時常因母音被拉長若干倍而造成與說話時的訊號相差甚多，我們因此試圖找出歌唱中的母音位置，並對其長度壓縮或裁剪，使其接近語音訊號，讓語句確認方法較能正常運作。

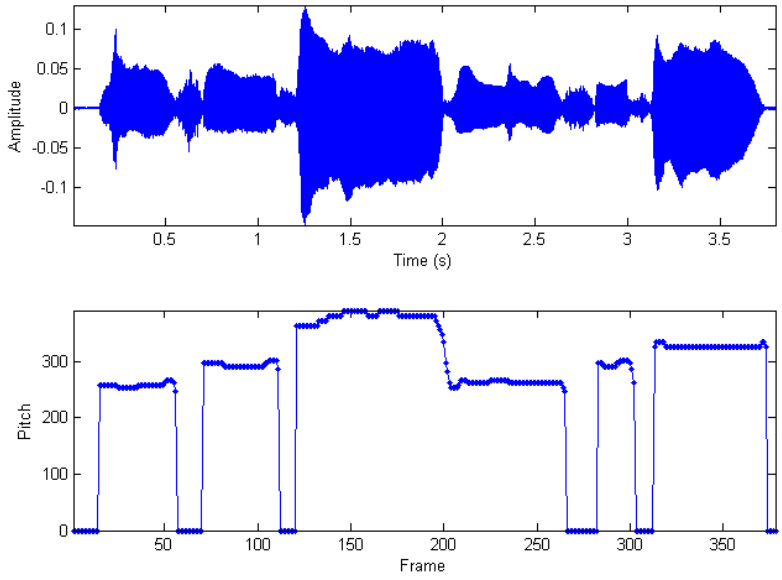
中文為一字一音節結構，每一個音節皆由子音(可能不包含)、母音與聲調所組成。

考慮一首歌曲大致包含歌詞和旋律兩部分，當依照歌詞內容進行朗讀所產生的聲音訊號為語音訊號；而若在同樣歌詞內容的情況下，加入旋律進行歌唱，所產生的聲音訊號即為歌唱聲音訊號。若與語音訊號相比，歌唱聲音訊號在同樣歌詞內容上的長度通常較長，一般是配合歌曲旋律將歌詞部分拉長，而拉長的聲音部分多為母音部分。因此首先，我們需要找到母音的所在位置。圖三為一中文字之子音(Consonant)與母音(Vowel)的位置圖。從聲音訊號波形圖上觀察，能發現母音部分具有週期性；反之，子音部分則大多無週期性。因此，我們尋找一段歌唱聲音訊號或語音訊號具有週期性的位置即相當於等於找到其母音所在位置。

週期的倒數為頻率，一段聲音訊號之頻率的高低對應到時域上的音高(Pitch)。因此，我們計算一段聲音訊號的音高值並設定一臨界值，當高於此臨界值即判定為母音，便可達到母音偵測的目標。為此，本研究利用 YAAPT (Yet Another Algorithm for Pitch Tracking) [6]方法進行音高的追蹤，以便達到母音的偵測。圖四為一段聲音訊號之音高追蹤示意圖。



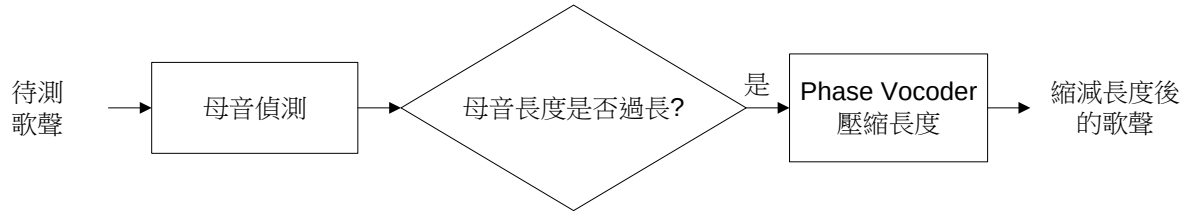
圖三、中文字「時」之子音與母音位置圖



圖四、演唱歌詞「有時候，有時候」之音高追蹤示意圖

（一）、母音壓縮

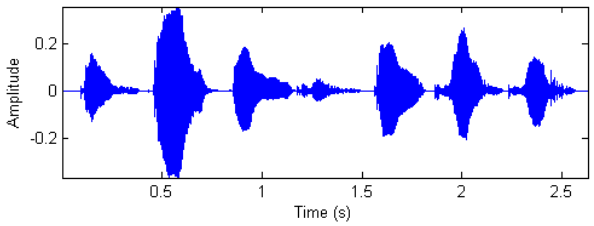
找到聲音訊號的母音位置後，我們將其壓縮，使其長度能夠接近一般的語音長度。本研究利用 Phase Vocoder [7][8]方法，針對超過一定長度的母音部分進行壓縮。圖五為母音壓縮流程圖，歌唱聲音訊號經由母音偵測後，從聲音訊號的起始位置依序計算母音音框(Frame)的數量。當母音音框超過某數量時(本研究設定為 10)，系統便藉由 Phase Vocoder 將此段母音部分進行壓縮，最後得到壓縮後的聲音訊號。圖六(a)為一段歌詞的語音訊號圖，而圖六(b)與圖六(c)為同樣一段歌詞之歌唱聲音壓縮前後的訊號圖。我們可以看到壓縮後的歌聲訊號長度接近說話的聲音訊號。



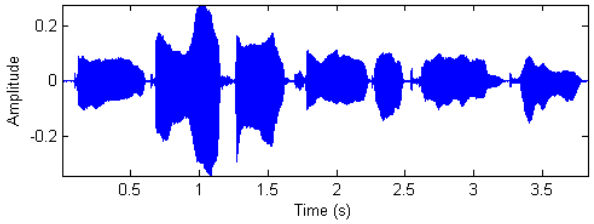
圖五、壓縮歌聲中的母音

（二）、母音裁剪

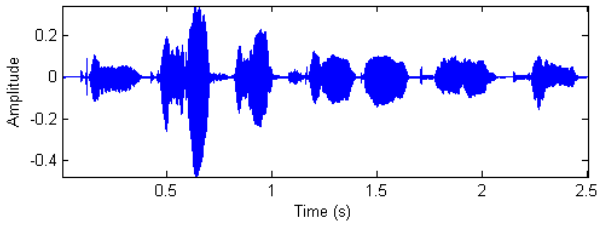
由於母音是週期性訊號，刪除其中部分的重複片段後並不影響其母音的特性，因此我們嘗試母音裁剪，將過長的歌唱母音直接切短，使其較像語音訊號的長度。裁剪方法同樣是先偵測歌聲中的母音位置，然後針對過長的母音直接剪去其後半部分一定比例的長度。圖六(d)為上述圖六(b)之歌唱聲音經由母音裁剪後的聲音訊號。



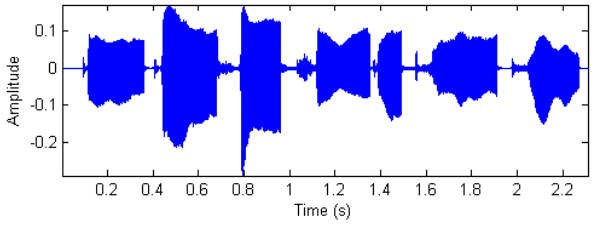
圖六(a)、正常語速唸歌詞「等到風景都看透」之語音訊號



圖六(b)、演唱歌詞「等到風景都看透」之聲音訊號



圖六(c)、將(b)之歌聲經由母音壓縮後的聲音訊號



圖六(d)、將(b)之歌聲經由母音裁剪後的聲音訊號

四、實驗

（一）、資料庫

因為並沒有先前研究探討唱詞確認問題，我們因此自行錄製歌唱聲音資料庫進行實驗。本研究邀請了五位女歌者與一位男歌者，每位歌者皆在同樣一安靜的房間內清唱十五首中文流行歌曲，包含約各半的快歌與慢歌。然後模擬在卡拉 OK 歌唱環境下可能發生的四種唱錯詞情況，分別請歌者錄製相同唱錯詞的歌聲，如表一所示，因此每位歌者共錄製七十五個歌唱音檔。

表一、歌唱情況列表

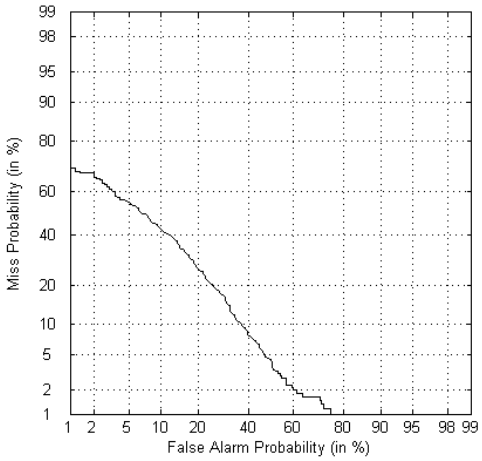
情況編號	演唱方式	例如
1	依歌詞正確演唱	歌詞為「等到風景都看透」， 唱詞為「等到風景都看透」。
2	模擬部分唱錯詞	歌詞為「等到風景都看透」， 唱詞為「等到人生都看透」。
3	模擬部分歌詞前後顛倒	歌詞為「等到風景都看透」， 唱詞為「都看透等到風景」。
4	模擬遺漏部分歌詞	歌詞為「等到風景都看透」， 唱詞為「等到風景都看　」。
5	未唱歌詞，僅哼出旋律	歌詞為「等到風景都看透」， 唱詞為「亨亨亨亨亨亨亨」。

接著，我們將七十五個歌唱音檔依照歌詞內容斷句切割為五百個歌唱片段音檔，這五百個歌唱片段音檔即是用來做為測試樣本的單位。因此，六位歌者總共會產生三千個測試樣本。音檔的取樣頻率皆為 16 kHz，解析度為 16 bits，單聲道；而每一個測試音檔的長度皆介於二至十三秒之間。錄音的過程中，每一首歌曲的伴奏音樂皆由耳機輸出，因此未被收錄至音檔之中。

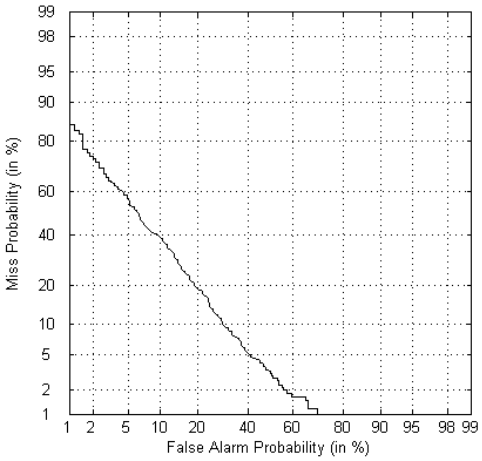
（二）、實驗結果

1、應用語句確認系統於唱詞確認之結果

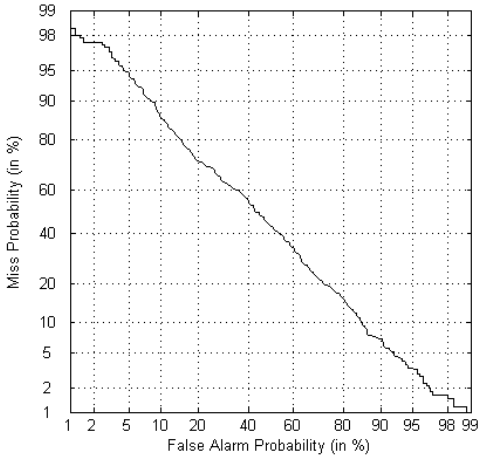
圖七為方程式(1)測試歌唱訊號所獲得之分數的 DET (Detection Error Tradeoff)曲線圖[9]，該曲線圖橫軸(False Alarm Probability)表示測試樣本為唱詞正確，但卻被判斷為有錯的機率；而縱軸(Miss Probability)表示測試樣本為唱詞有誤，但卻被判斷為無誤的機率。圖七包含(a)、(b)、(c)與(d)四張圖，分別為使用歌唱情況 1 與另外四種唱錯詞情況繪製而成。



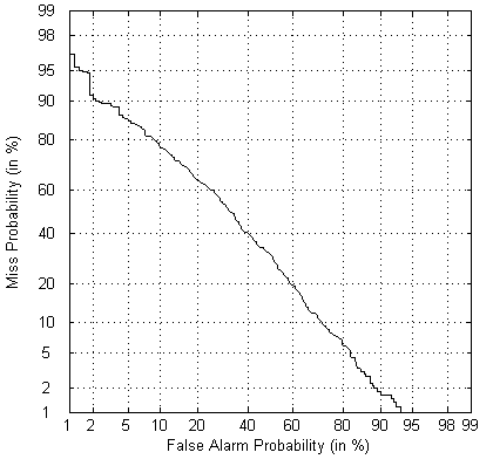
(a)、歌唱情況 1 與 2 之 DET 曲線圖



(b)、歌唱情況 1 與 3 之 DET 曲線圖



(c)、歌唱情況 1 與 4 之 DET 曲線圖



(d)、歌唱情況 1 與 5 之 DET 曲線圖

圖七、根據方程式(1)測試歌唱訊號所獲得之分數的 DET

另外，若我們將方程式(1)的臨界值 δ 設定為 0.22，可得接近等錯誤機率(Equal Error Probability)，即 False Alarm Probability = Miss Probability，其唱詞確認正確率如表二所示，其中確認正確是指：當歌唱情況為 1 時，系統判定其演唱的歌詞內容為無誤，或是當歌唱情況為 2、3、4 與 5 時，系統判定其演唱的歌詞內容為有錯。表二中的整體平均確認正確率為 63%。

表二、方程式(1)的臨界值 δ 設定為 0.22 的唱詞確認結果

情況編號	演唱方式	確認正確率
1	依歌詞正確演唱	76%
2	模擬部分唱錯詞	79%
3	模擬部分歌詞前後顛倒	84%
4	模擬遺漏部分歌詞	32%
5	未唱歌詞，僅哼出旋律	45%

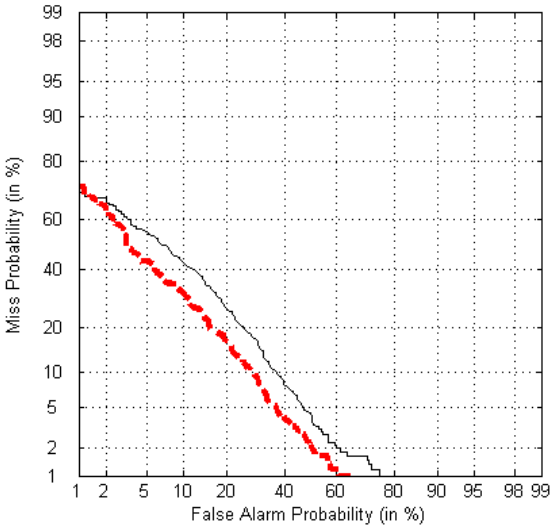
2、經母音壓縮後的唱詞確認結果

接著，我們將歌唱聲音訊號的母音部分進行壓縮後，再送入語句確認系統進行唱詞確認。確認結果如表三所示。經過不同壓縮比例的實驗與觀察，發現壓縮比例使用 1/2 的改善效果為最好，因此我們使用壓縮比例為 1/2 的改善方式。最後得到平均確認正確率為 72%，較表二中未經改善的確認結果提升了 9%。

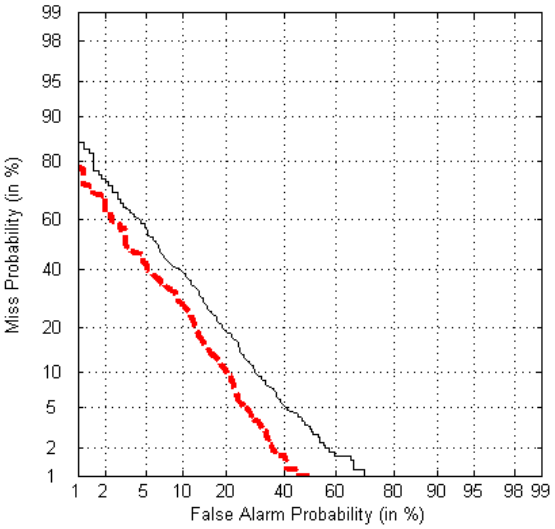
表三、經母音壓縮後的唱詞確認結果

情況編號	演唱方式	確認正確率
1	依歌詞正確演唱	75%
2	模擬部分唱錯詞	86%
3	模擬部分歌詞前後顛倒	89%
4	模擬遺漏部分歌詞	39%
5	未唱歌詞，僅哼出旋律	73%

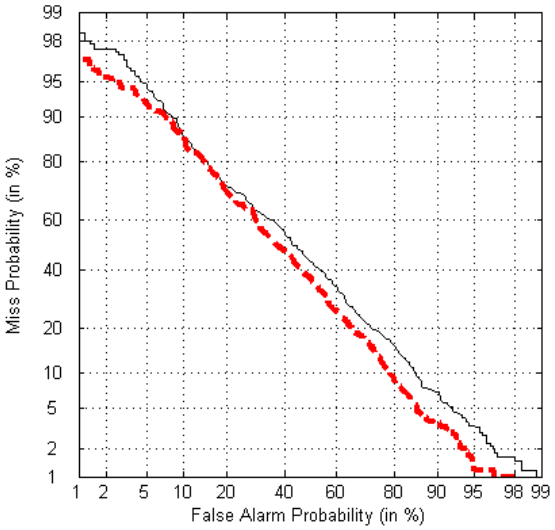
圖八比較母音壓縮前後之方程式(1)所獲得的分數 DET 曲線，其中的實線即圖七中的曲線，而虛線為經過母音壓縮後的結果。我們可以清楚看到母音壓縮後可讓確認系統的 False Alarm Probability 與 Miss Probability 皆下降。



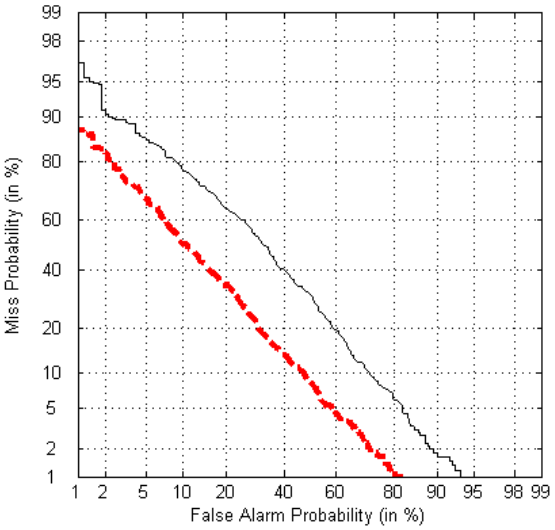
(a)、歌唱情況 1 與 2 之 DET 曲線圖



(b)、歌唱情況 1 與 3 之 DET 曲線圖



(c)、歌唱情況 1 與 4 之 DET 曲線圖



(d)、歌唱情況 1 與 5 之 DET 曲線圖

圖八、母音壓縮前後之方程式(1)所獲得的分數 DET 曲線比較

3、經母音裁剪的唱詞確認結果

最後，我們測試將歌唱訊號的母音進行裁剪後，再利用語句確認系統進行唱詞確認。確認結果如表四所示。經過不同裁剪比例的實驗與觀察，發現裁剪比例使用 1/2 的改善效果為最好，因此我們使用裁剪比例為 1/2 的改善方式。最後得到平均唱詞確認正確率為 75%，較未經改善的確認結果提升了 12%。圖九比較母音裁剪前後之方程式(1)所獲得的

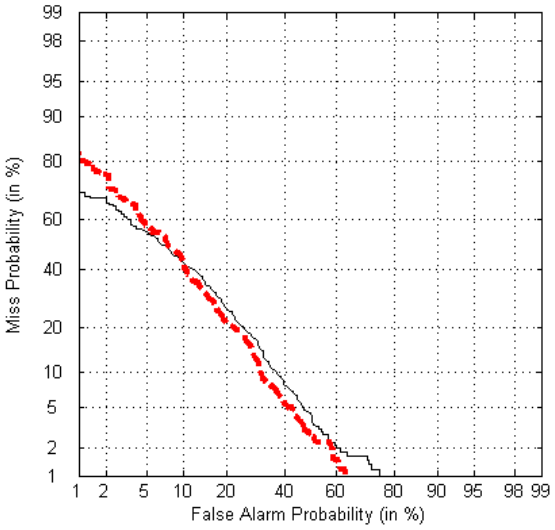
分數 DET 曲線，其中的實線為圖七中的曲線，而虛線為經過母音裁剪後的結果。我們可以看到母音裁剪後可更明顯讓確認系統的 False Alarm Probability 與 Miss Probability 皆下降。

表四、經母音裁剪後的唱詞確認結果

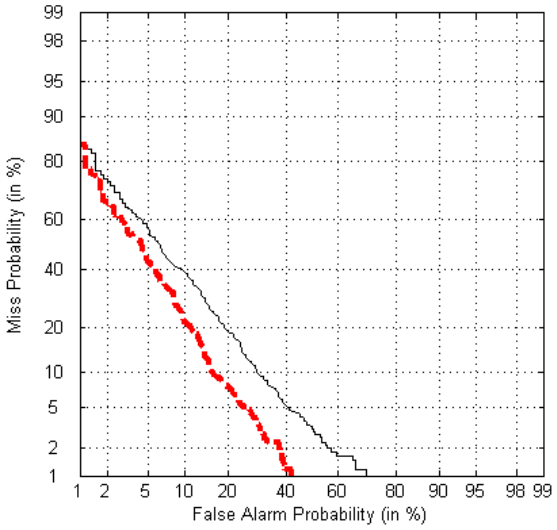
情況編號	演唱方式	確認正確率
1	依歌詞正確演唱	56%
2	模擬部分唱錯詞	90%
3	模擬部分歌詞前後顛倒	95%
4	模擬遺漏部分歌詞	57%
5	未唱歌詞，僅哼出旋律	79%

五、結論

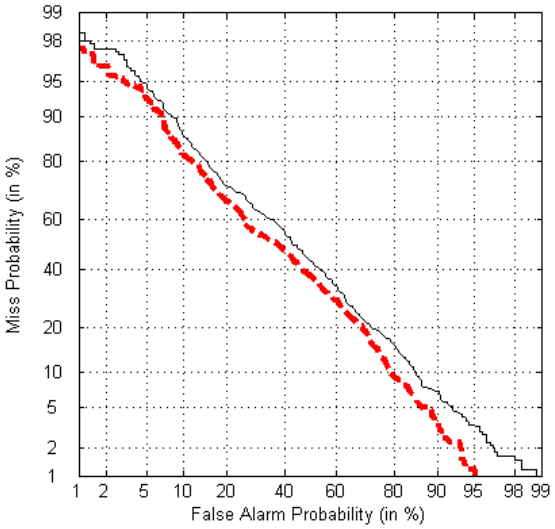
本研究發展出一種唱詞確認系統，可自動判斷演唱者是否唱錯歌詞。我們以語音辨認上的語句確認系統為基礎，並針對歌聲中的母音拉長特性進行處理，以改善唱詞確認的正確性。經實驗評估，透過母音長度壓縮或裁剪的前置處理，約可分別提升語句確認系統 9%與 12%在判斷唱詞方面的正確率。



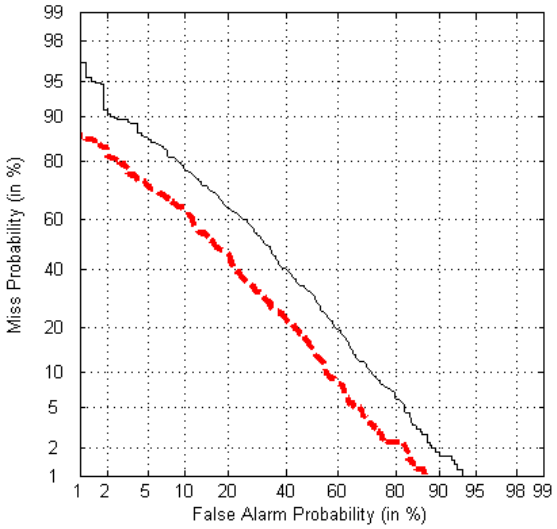
(a)、歌唱情況 1 與 2 之 DET 曲線圖



(b)、歌唱情況 1 與 3 之 DET 曲線圖



(c)、歌唱情況 1 與 4 之 DET 曲線圖



(d)、歌唱情況 1 與 5 之 DET 曲線圖

圖九、母音裁剪前後之方程式(1)所獲得的分數 DET 曲線比較

參考文獻

[1] W. H. Tsai and H. C. Lee, “Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features,” *IEEE Trans. on Audio, SpeechandLang. Processing*, vol. 20, no. 4,2012, pp. 1233-1243.

[2] W. H. Tsai and C. H. Ma, “Automatic speech and singing discrimination for audio data indexing,”*The 4th IEEE International Congress on Big Data*, Taipei Satellite Session, 2014, pp. 276-280.

[3] The Hidden Markov Model Toolkit (HTK) - http://htk.eng.cam.ac.uk/

[4] The Association for Computational Linguistics and Chinese Language Processing (ACLCLP) - http://www.aclclp.org.tw/use_mat_c.php

[5] H. Jiang and C. H. Lee, "A new approach to utterance verification based on neighborhood information in model space," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, 2003.

[6] S. A. Zahorian and H. Hu, “A spectral/temporal method for robust fundamental frequency tracking,” *The Journal of the Acoustical Society of America*, vol. 123, no. 6,

2008, pp. 4559-4571.

[7] J. L. Flanagan and R. M. Golden, “Phase Vocoder,” *Bell System Technical Journal*, vol. 45, no. 9, 1966, pp. 1493-1509.

[8] M.Dolson, “The phase vocoder: A tutorial,” *Computer Music Journal*, vol. 10, no. 4, 1986, pp. 14-27.

[9] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, “The DET curve in assessment of detection task performance,” in *Proceedings of Eurospeech,*Greece, 1997, pp. 1895-1898.