廣電新聞播報品質電腦化評估系統之研發

國立政治大學

資訊科學系

指導教授：廖文宏

專題學生：蘇以暄

廣電新聞播報品質電腦化評估系統之研發

指導教授：廖文宏老師

學生：蘇以暄

國立政治大學資訊科學系

摘要

在目前的日常生活中，要掌握最新最快的新聞已經不是一件難事。但在要求新聞內容要新要快的同時，對於觀眾或聽眾而言，主播的播報品質也極為重要，也就是說，在追求內容的即時性之外，聽者的接受程度也應該要加以注意的。

因此我們期望發展出一套客觀的系統來對於播報品質加以評估，利用目前的語音處理技術來達到這樣的目標。本專題中主要是將焦點集中在顯性特徵上，最後以所選定的參數來達到評估的目的。

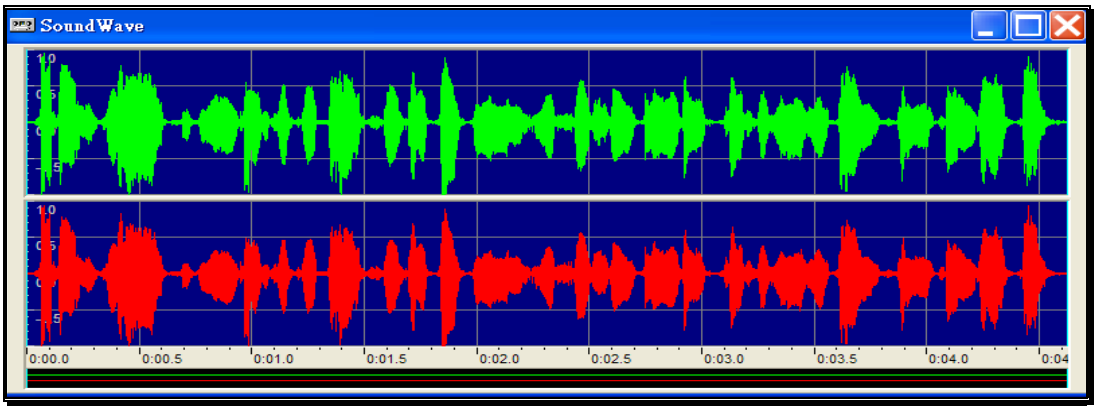
目次

1、導論

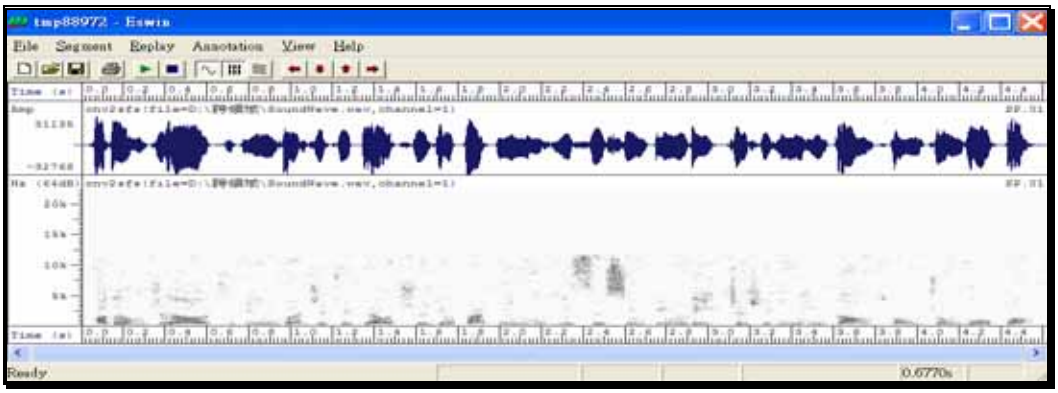
廣播或新聞播報的專業養成課程中，常有所謂的「新聞播報技巧」、「國語正音訓練」等實習科目，其主要目的無非希望聽眾或觀眾能很清楚的接受到新聞內容所要傳遞的訊息；針對實習成效的評量，傳統上大多以主觀的認知，輔以若干描述式(descriptive)的標準，例如：播報的速度、咬字的清晰度、流暢度、〝吃螺絲〞的頻率等，來判斷電視與電台主播的播音品質。類似的評估方式除了需要付出「專家」的昂貴人力代價，有時亦不免流於主觀，而對於播報者的建議改進事項亦難以精準的加以表達。

衡諸上述狀況，本計畫擬運用數位語音處理(digital speech processing)技術，將聲波數位化後進行處理與分析(如圖一)，從中擷取有意義的信號特徵(feature)，計算與播音品質相關之參數(如 speaking rate, pitch, timbre, breathiness, amplitude 等，如圖二)，並發掘聲波中的物理特性(physical characteristics)與知覺特徵(perceptual features)的關聯性，以建立相關參數權重對應關係，進而將評判標準加以量化，發展出一套客觀的電腦化自動評估系統。

本計畫範圍雖僅限定於處理廣播或電視新聞播報品質相關課題，但我們期望研究所得之若干結果能應用於聲學語音學(acoustic phonetics)領域，如語音的多重差異性研究(聲音中的情緒差異、表情差異等)、語言學習(腔調的偵測與分類)。



圖一、聲波經過數位化的結果



圖二、語音信號之聲譜儀(spectrogram)

2、步驟

如前所述，藉由數位語音處理的技術，我們可以將聲音數位化後進行處理分析，針對物理特性(physical characteristics)，例如：基頻(fundamental frequency)、能量、越零率(zero-crossing rate，ZCR)、調變率(modulation rate)；以及知覺特徵(perceptual features)，如：音調(pitch)、響度(loudness)、音色(timbre)、節奏(rhythm)…等參數進行估算。

針對本項研究課題，我們訂出了以下幾個關鍵步驟：

1.從聲音檔中找出顯性、隱性之特徵，並針對這些特徵訂定其權重。
2.定義受測者所要朗讀的文本內容與長度。
3.將固定文本延伸到任意內容。

首先，在收音之後，我們會從聲波之中找出顯性特徵，例如是聲音的頻率、說話節奏的快慢…等等，針對這些特徵，依據影響播報品質的程度，將其比重訂定出來，獲得一個較精確的對應關係，以供評估。

至於在隱性特徵方面，很多時候我們可以輕易的用耳朵聽出，哪些語調聽起來「抑揚頓挫」，哪些語調讓我們覺得「悅耳動聽」或「鏗鏘有力」，但這些描述式的辭語，究竟對應至聲波信號的哪些特質或特質的組合，都無法很明確的利用上面的方式來得知，因此在這一階段，我們將側錄電視新聞主播播報的片段，以圖型識別中 clustering 的技術、資料探勘(data mining)或類神經網路的相關演算法，從所蒐集大量的資料中，尋找其共通的特性。而我們預計蒐集與分析的資料有以下組合：

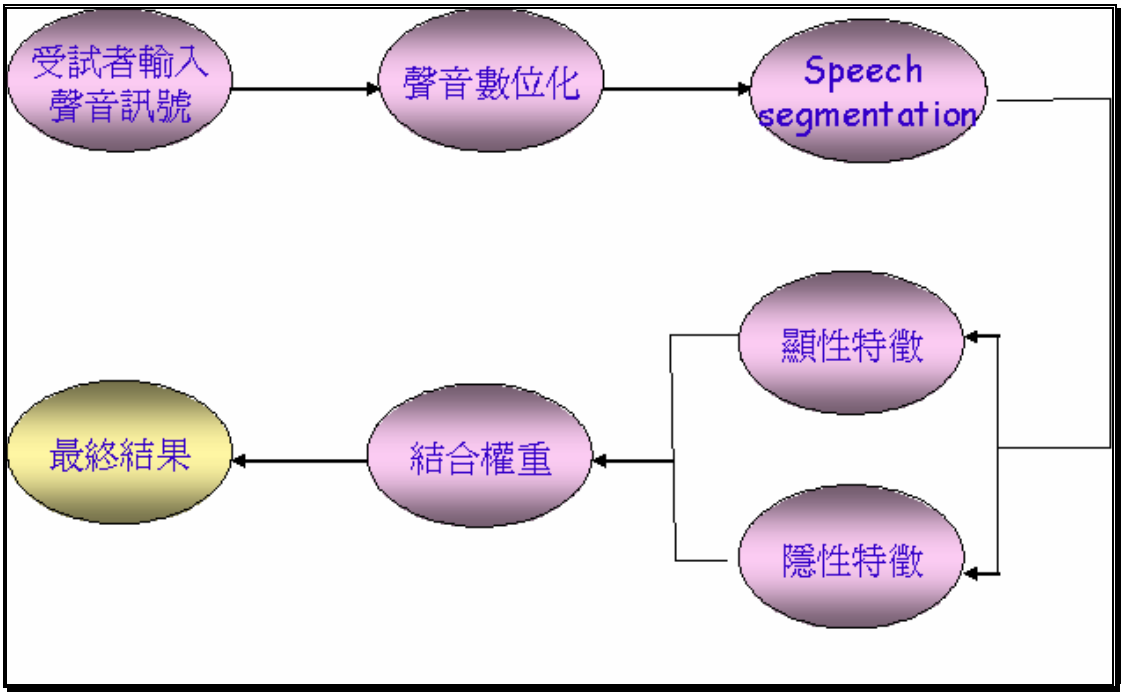
- 同一主播播報不同新聞內容
- 不同主播播報相同新聞內容
- 不同性別主播播報新聞內容

有關文本長度的選定，一般而言，讀入的文本必須有一定的長度，方能提供可信賴的分析結果，因此我們必須決定聲音樣本的最小長度，以作為文本樣本設計之依據。

至於文本內容的選擇，有鑒於此計畫是針對廣電新聞播報品質的評估，因此文本的選定十分重要，什麼樣的文本才能讓我們清楚的得到評估的結果？我們會希望文本能夠讓受試者表現出他們的播報能力，例如：饒舌的程度，文意的深淺…等。並且，指定文本內容，將可使讓分析大為簡化，如果我們有特定的文本，在

分析的過程之中，將會帶給我們在語音內容之辨識與比對上的方便。

有關核心的語音處理技術的設計與開發部分，我們將充分應用 MATLAB 內建之信號處理函式庫，以協助演算法的開發與快速測試，一旦驗證成功後便可轉為 BCB 之程式碼，建立 Windows 平台下可執行的應用程式。

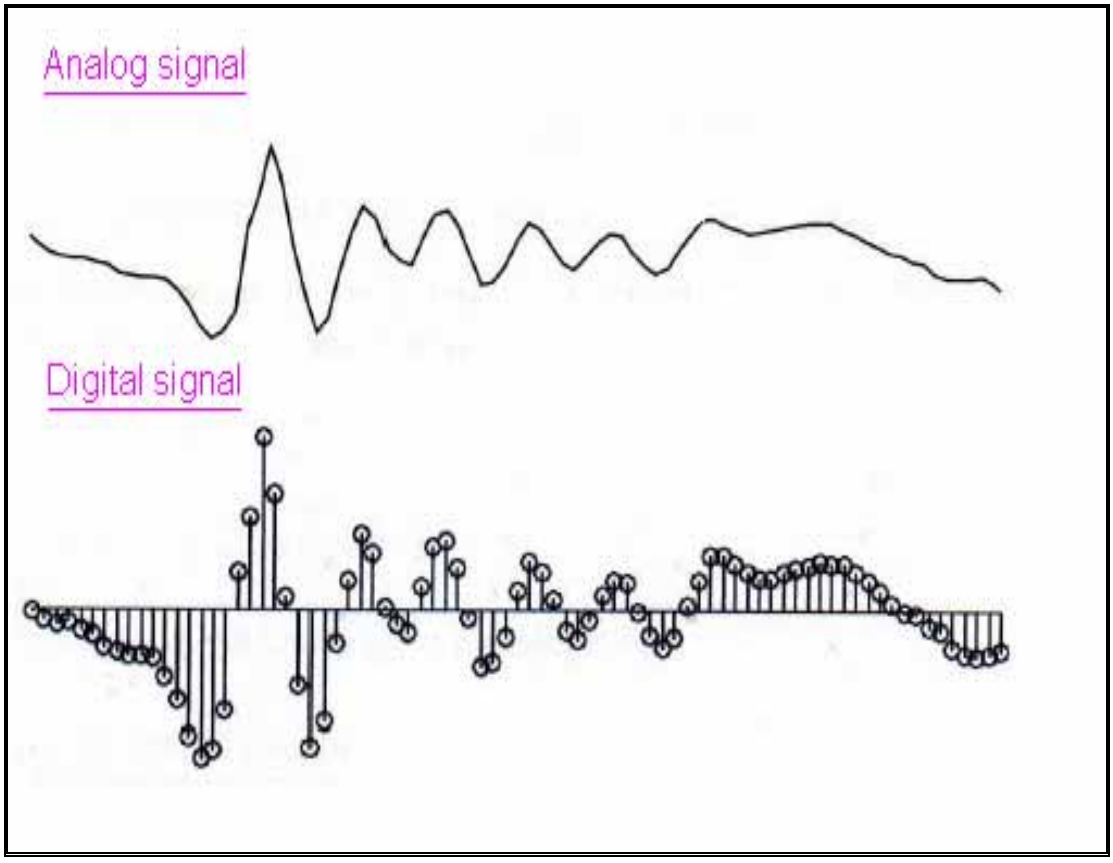


圖三、系統流程圖

3、原理

3.1. Digital speech processing

受試者所輸入的聲音轉換必須成電腦可接受的型態，也就是所謂的聲音數位化，Fs 為 44100Hz(sampling rate)，每份取樣訊號的量化階度為 16 bit。



圖四、聲音訊號由類比轉數位，所有的值存在 x[n]之中

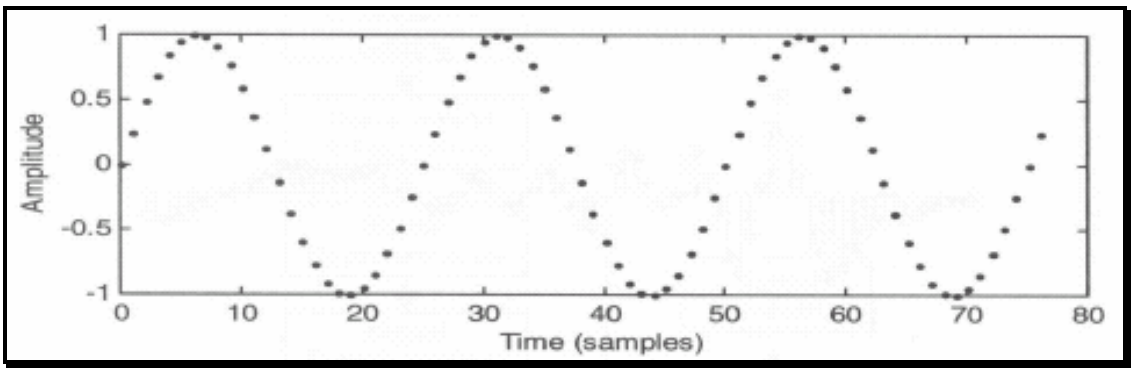
此後，我們必須讓 x[n]中的值調整至介於-1 到 1 之間，所以我們會將 x[n]寫成下面的式子：

$x[n]=A\cos(\omega n+\varphi)$

其中 A:振幅

ω:角頻率(即為$2\pi f$)

φ:相角



圖五、經過轉換之後，所有的值皆介於-1到1之間

3.2. 字元切割(speech segmentation)

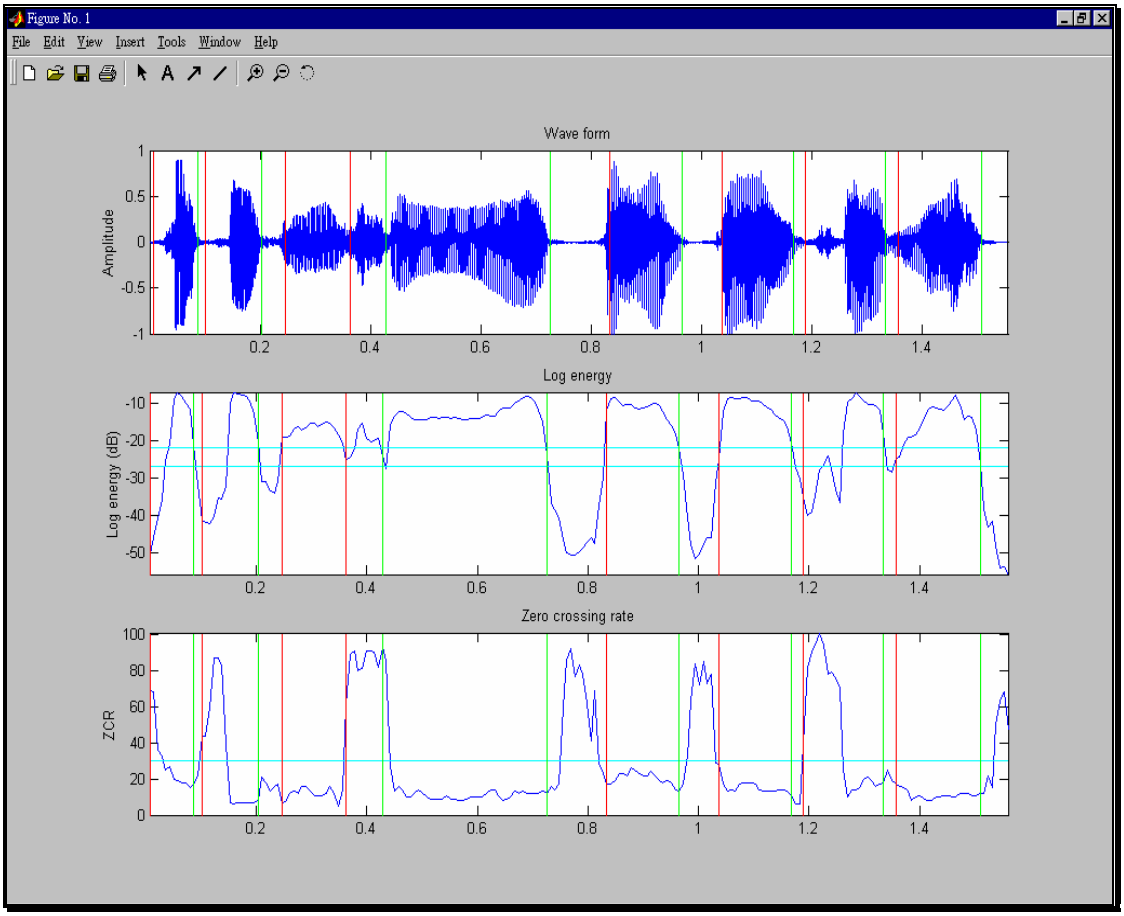
此部分主要是讓我們可以切割出一段話中的每一個字，主要用到的特徵有 energy 及 ZCR(zero-crossing rate)，我們分兩階段來執行。

首先，在第一階段中，必須要分割出一個個的 frame，且允許重疊，所以我們設定 end point 的 frame size 為 384，overlap 為 64，ZCR 定為 0.3，而能量(energy)部分是取對數值(log)來計算。

此外，必須找出一開始的 end points，主要是根據 log energy 為-15(upper level)這個值，這個值是經過不斷的實驗所調整出來的最佳值，此時很基本的分出每一個字的結尾是下一個字的開頭。再將 end points 以 log energy 為-20(lower level)這個值去展開(expand)。雖然我們預設了 upper level 及 lower level 的值，但是我們也開放給使用者自行調整，因為在不同的環境之下所需要的值可能會隨之不同。

接下來，我們將前面所得的 end points 展開(expand)到 ZCR 較高的區域，經過這兩的步驟的調整，每個字的分野被切割的更精確，不再只是一味的定義這個字的結尾必定是下一個字的開頭，因為中間可能會有停頓的地方。然後我們把重複的 sound segments 去除之後，轉換成 sample-point-based index，配合剛剛所做的結果以圖形的方式呈現出來。

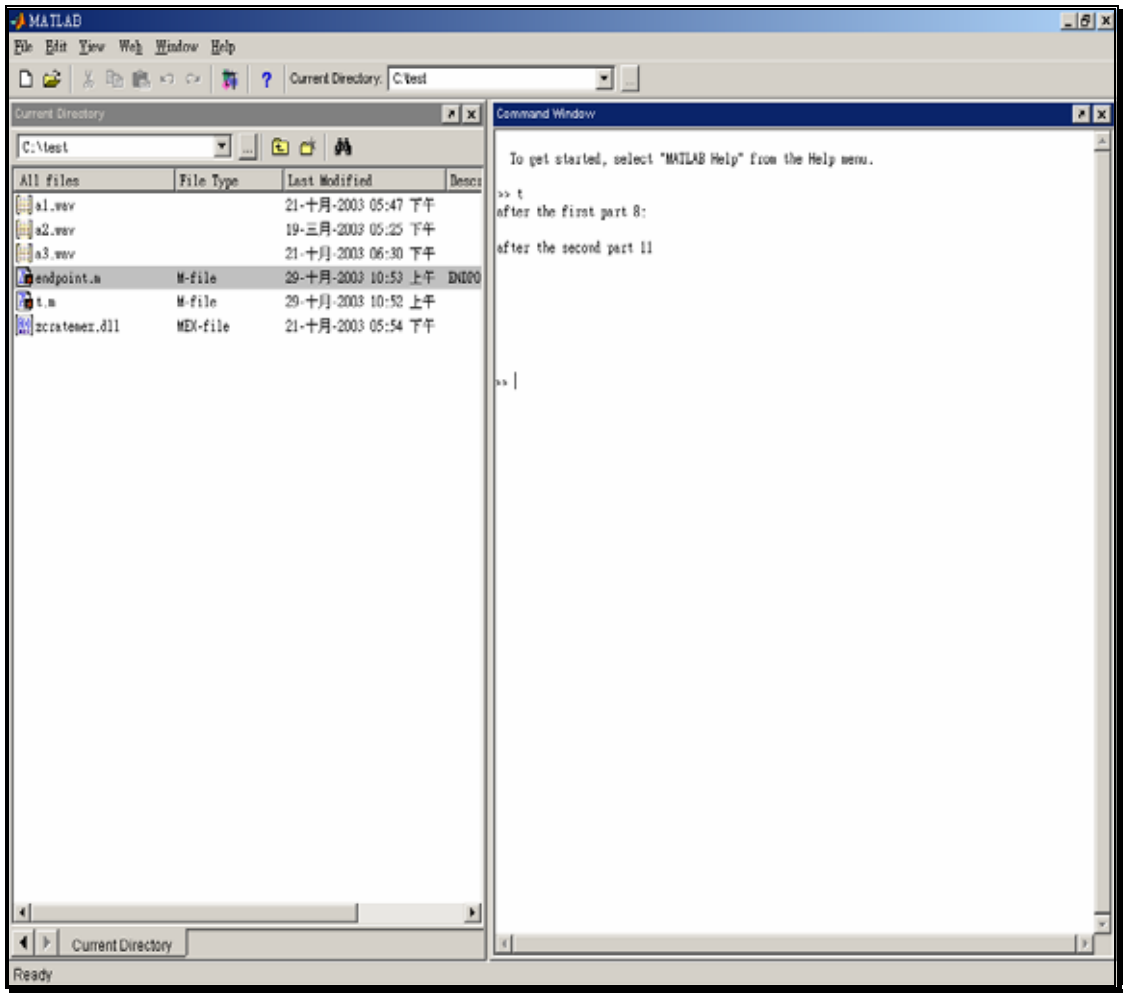
下圖即為切割的結果，上方的圖是原來的聲音，而中間的圖是用 log energy 去分割的結果。下方的圖是由 ZCR 所得的結果，兩者互相配合我們就可以得到上方的圖的切割，其中紅線代表字的開始，綠線代表字的結尾。



圖六、字元切割(speech segmentation)的第一個階段

如上圖所示，我們先取較短的資料來做測試，其內容為：「他最近是因為發高燒還有…」，共計十一個字，經過第一階段的切割之後，得到的結果是八個字，與實際有所出入，其原因在於口語中常會有連音的情況發生，例如〝就醫〞之類的詞，往往會在我們口語中不知不覺被連成了一個字。也就是在 fluent speech 中所會有的現象，若是在 read speech 中則不會有此種情況發生。就以上所測試的內容之中，〝是因為〞這三個字的部分，〝是〞和〝因〞就發生了這樣的問題，同樣的，〝還有〞亦是如此，所以會被當成了一個字，有鑑於此，我們必須採取第二階段的修正。

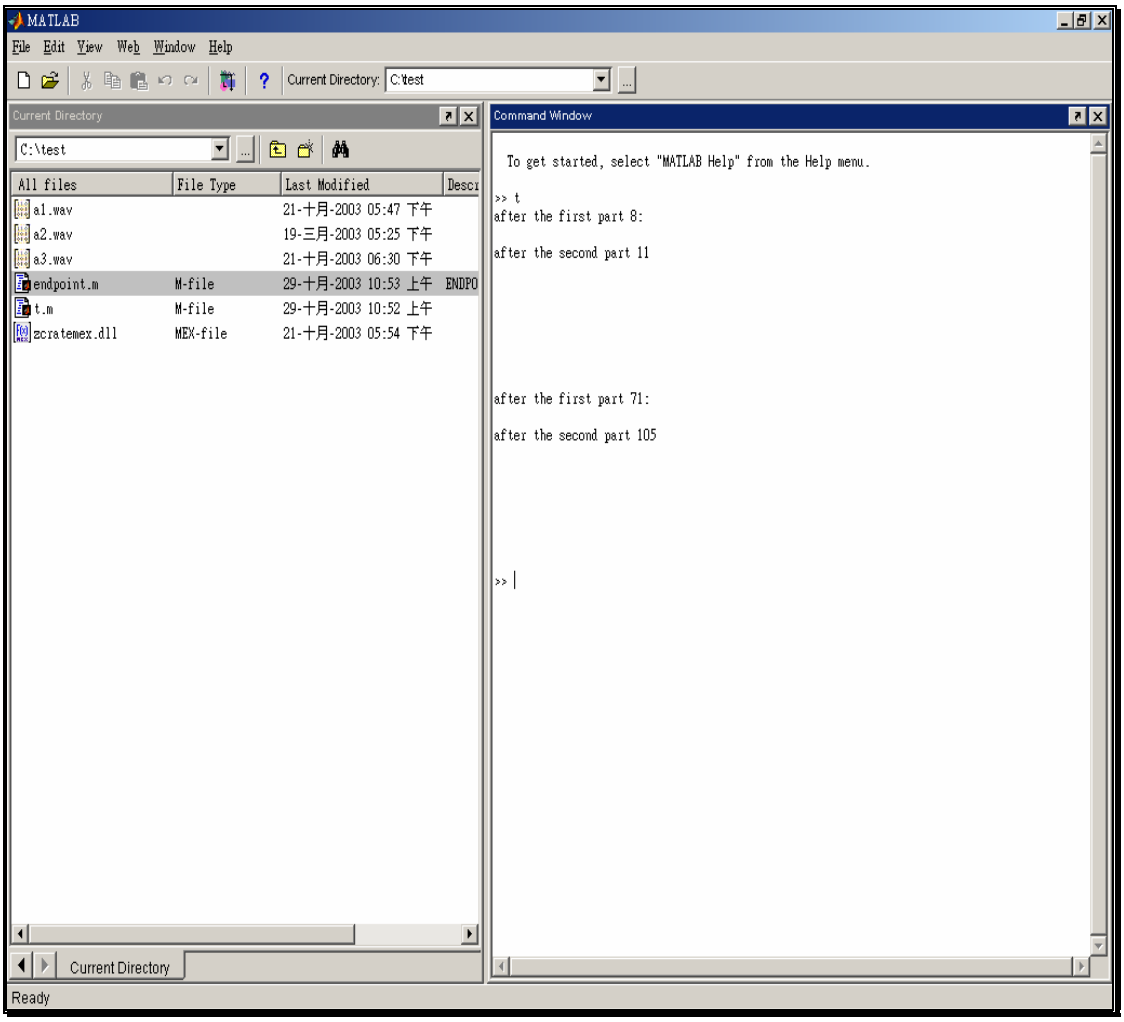
在第二階段之中，我們從第一階段可以獲得所切出的每個部分的長度，將其排序之後，剔除極值，也就是特別長的字，取得每個字的長度之平均值，在以總長度去除以這個平均值，四捨五入得到我們計算出的此段受試資料的字數，經過此階段的修正之後，我們計算出上述的例子的字數是十一個字，與受試資料的字數符合。



圖七、經過第二階段的修正結果

經過上面的實驗，我們將所側錄的整段聲音放入測試，以加長測試資料的長度來做實驗，以下部分是我們的測試資料的文稿：

「好的我們知道疾病管制局表示，又發現一名疑似新病例，這是一名剛從中國廣東旅遊回來的嘉義王姓男子，他最近是因為發高燒還有咳嗽等等的症狀，到嘉義基督教醫院就醫，醫院方面就懷疑是罹患嚴重急性呼吸道症候群，立刻將他隔離治療。」共計一百個字，以下是我們的測試結果：



圖八、以上為第二個實驗的測試結果，實際資料為一百個字，經過第一階段的分割得到七十一個字，經過第二階段的結果為一百零五個字。

3.3. 字數統計(速度)

有鑒於要評鑑一個主播的說話速度是否得宜，所以我們必須要知道每分鐘大致上該播報幾個字，此時我們就必須用到前面所提到的 speech segmentation，來得知受試者的播報速度，而這個準則，是來自於和廣電系老師訪談的結果，再配合我們側錄電視上的主播來統計而得到的數值，大約介於每秒鐘 4.5 到 6.5 個字之間，而在廣播新聞方面，因為缺乏畫面的輔助，所以一般來說速度上為每分鐘三百個字，也就是說每秒鐘五個字左右為宜。

3.4. 速度快慢的一致性

在播報的過程中，有時候播報員可能會受到某些因素，可能是個人情緒或者是文稿內容的饒舌程度，因而使得播報的速度呈現不一致的情況，也就是一般我

們所說的〝忽快忽慢〞。因此在參數之中我們將其列入考慮，其原理主要是取用每五個字為一個單位，以三個字為 overlap，來取的這個區段中的速度平均快慢，與下一個區段相比較，若差異過大，即可發現其有速度快慢不一致的傾向產生。

3.5. 聲音大小穩定性

在播報過程中，只要在音量呈現較為一致的狀況之下，聲音的大小其實是可以經過音控調整。但是若音量呈現忽大忽小的狀態之時，就不容易有音控去針對幾個字去動態調整，因此我們將聲音大小的穩定性列如考量，而非只是聲音的大小。

在一段資料之中，經過前面的切割之後，我們紀錄下每一個字的資料，所以我們可以藉由所紀錄下來的資料，去評斷受試者的聲音是否具穩定性，其中判斷的依據主要是在於我們所切出來的聲波當中，找出每個字的最高點，用以來判斷受試者的聲音是否忽大忽小，其原理同 3.4 的方法，取出區段並加以重疊來計算，以上是第一個部分。

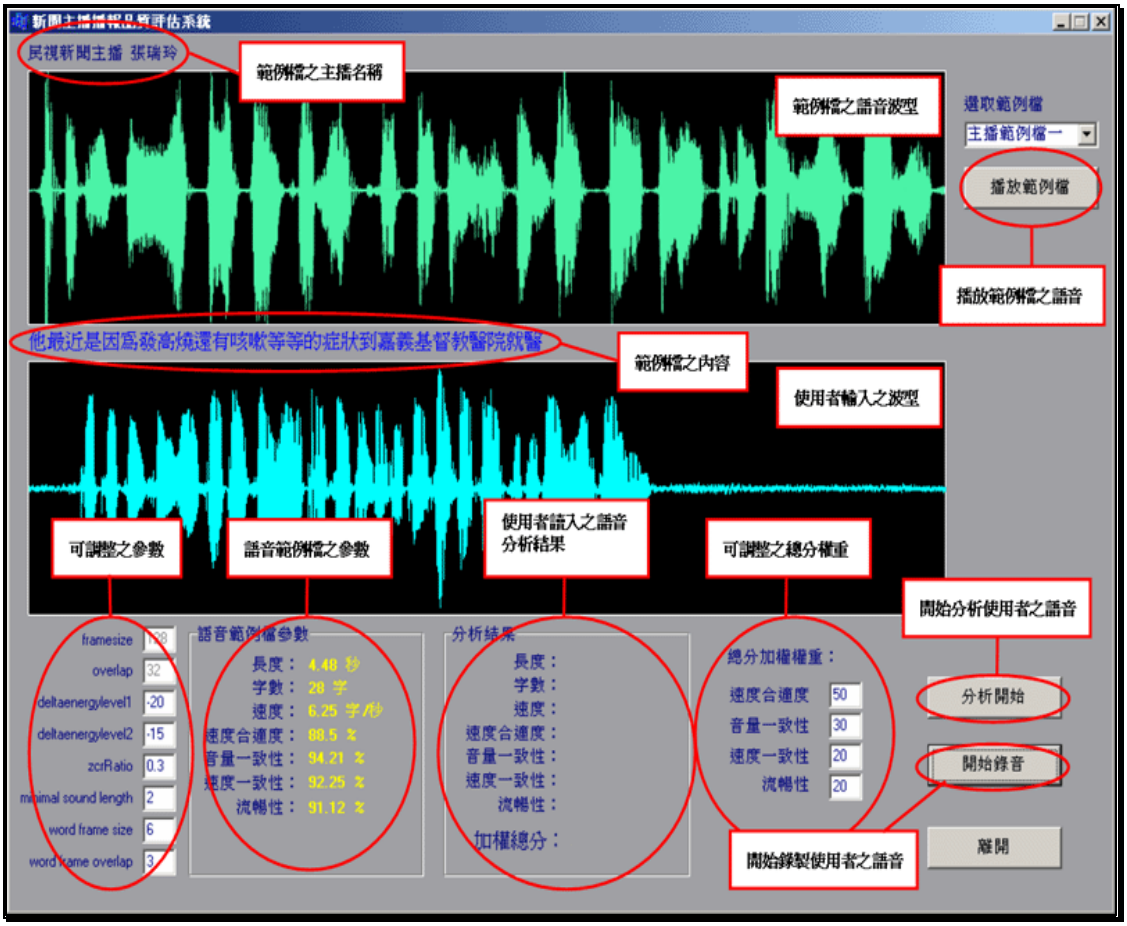
3.6. 隱性特徵

此部分的參數無法從聲波之中直接求得，因為何謂〝悅耳動聽〞並不是由物理特性可以清楚界定的，此部分牽涉到人的聽覺感受，所以我們需要側錄大量聲音來做 data mining 來執行，而目前此部分尚未完成，有待進一步閱讀參考 hearing 方面的知識，配合心理學上的認知心理學來加以輔助，以便做更精確的判斷。

3.7. 權重的制定

在權重的制定上，我們提供預設值，但每位使用者可以依其需求做調整，其原因在於，隨著不同的使用者，參數的權重會隨之不同，每個使用者所著重的方向並不一致，因此我們採用讓使用者自訂的方式來制定權重比例。

4、結果



圖九、介面上各組成成分的說明

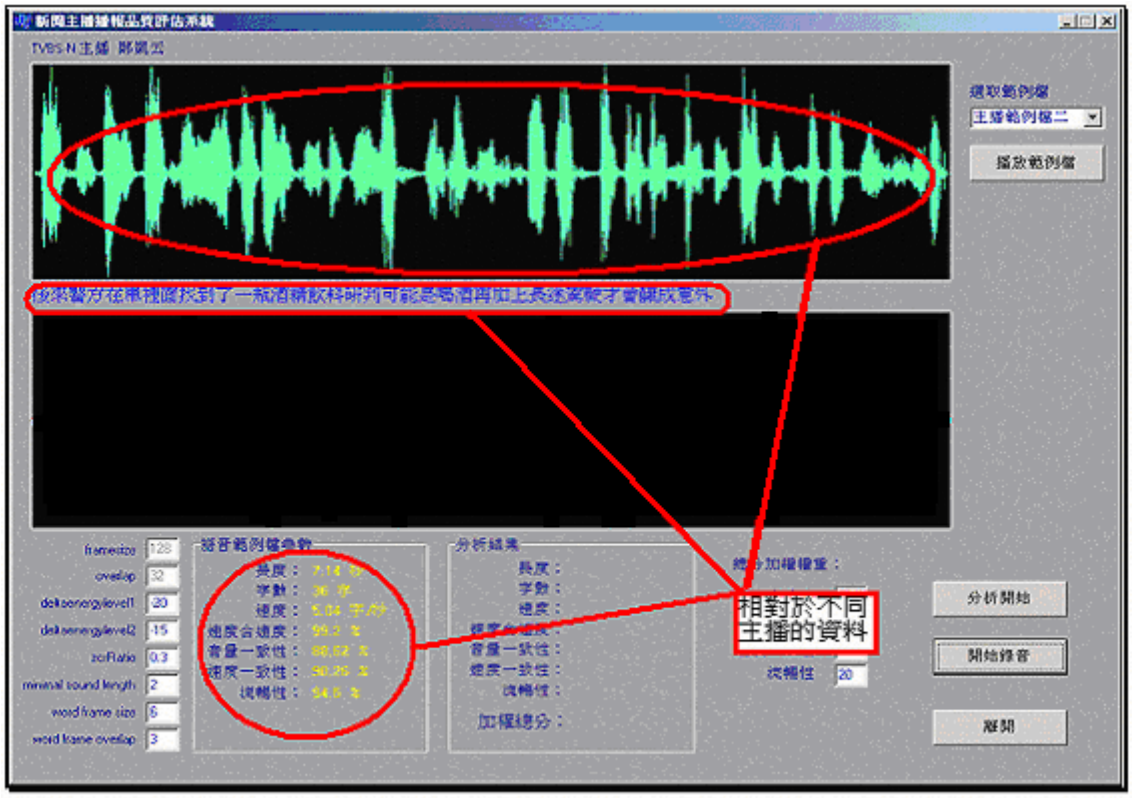
接下來為整個介面使用的流程，首先我們必須選定一個主播來做為我們評比的依據。在我們的範例中主要有八個主播的聲音，包含了民視新聞台、三立新聞台、中天新聞台、東森新聞台，以及 TVBS-N 等男女主播。

在選定之後隨即會跳出所相對應的文稿及聲音波形以供參考，其相對的速度、音量等等也會一併在下面的部分展出，讓使用者可以清楚看到。

接下來是播放主播聲音的部分，藉以讓使用者可以有參考的範例，對於所選定的主播的播報內容、速度有大致的了解。


新聞主播播報品質評估系統
TVBS-N主播 郭佩芸
主播範例檔二
播放範例檔
選定主播範例
後來警方在車裡座找到了一瓶酒精飲料研判可能是喝酒再加上長途駕駛才會釀成意外
frameize
overlap
deltaenergylevel1
deltaenergylevel2
zcrRatio
minimal sound length
word frame size
word frame overlap
語音範例檔參數
分析結果
總分加權權重：
速度合適度
音量一致性
速度一致性
流暢性
分析開始
開始錄音
離開

圖十、選定主播範例之處

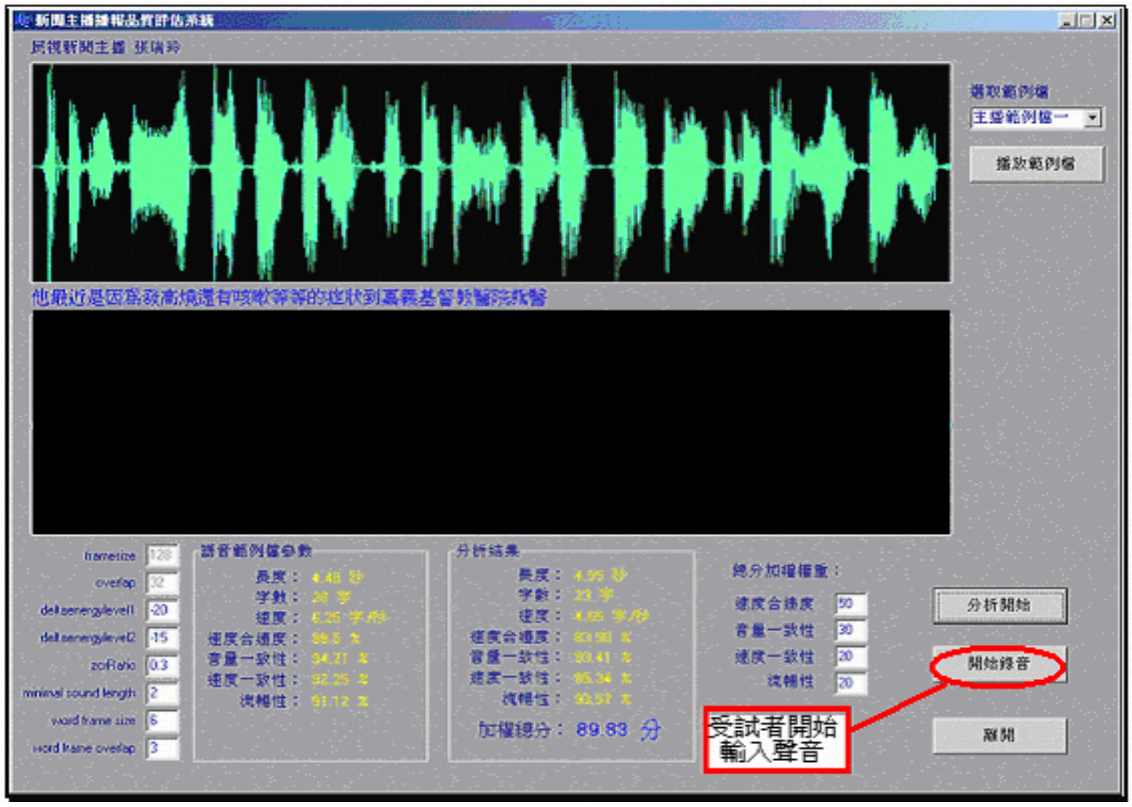

新聞主播播報品質評估系統
TVBS-N主播 郭佩芸
主播範例檔二
播放範例檔
後來警方在車裡座找到了一瓶酒精飲料研判可能是喝酒再加上長途駕駛才會釀成意外
相對於不同
主播的資料
分析開始
開始錄音
離開

圖十一、不同主播的畫面呈現

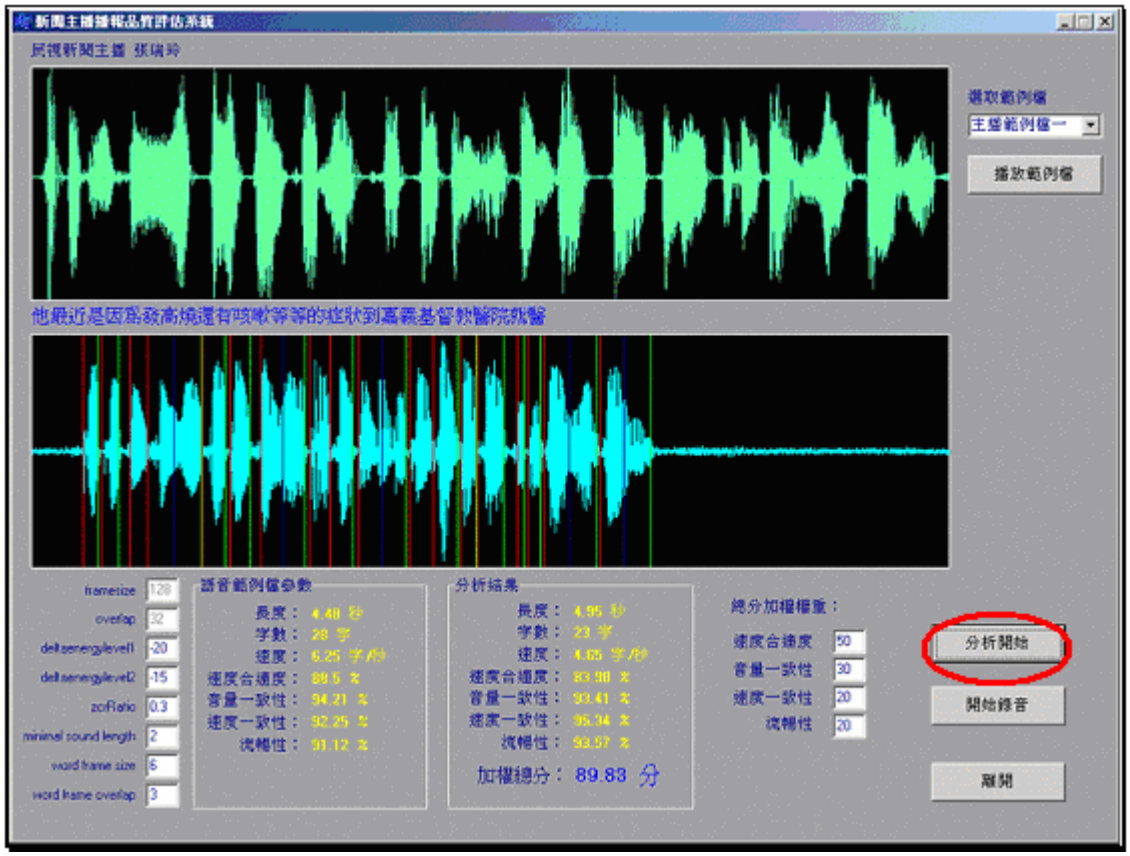
圖十二、播放選定主播的聲音

在聽過範例之後，接下來為使用者輸入自己聲音的步驟，按下〝開始錄音〞之後，使用者便可以照著主播的文稿內容來朗讀，在完畢之後，同樣再按一次剛才的按鍵即可停止錄音。

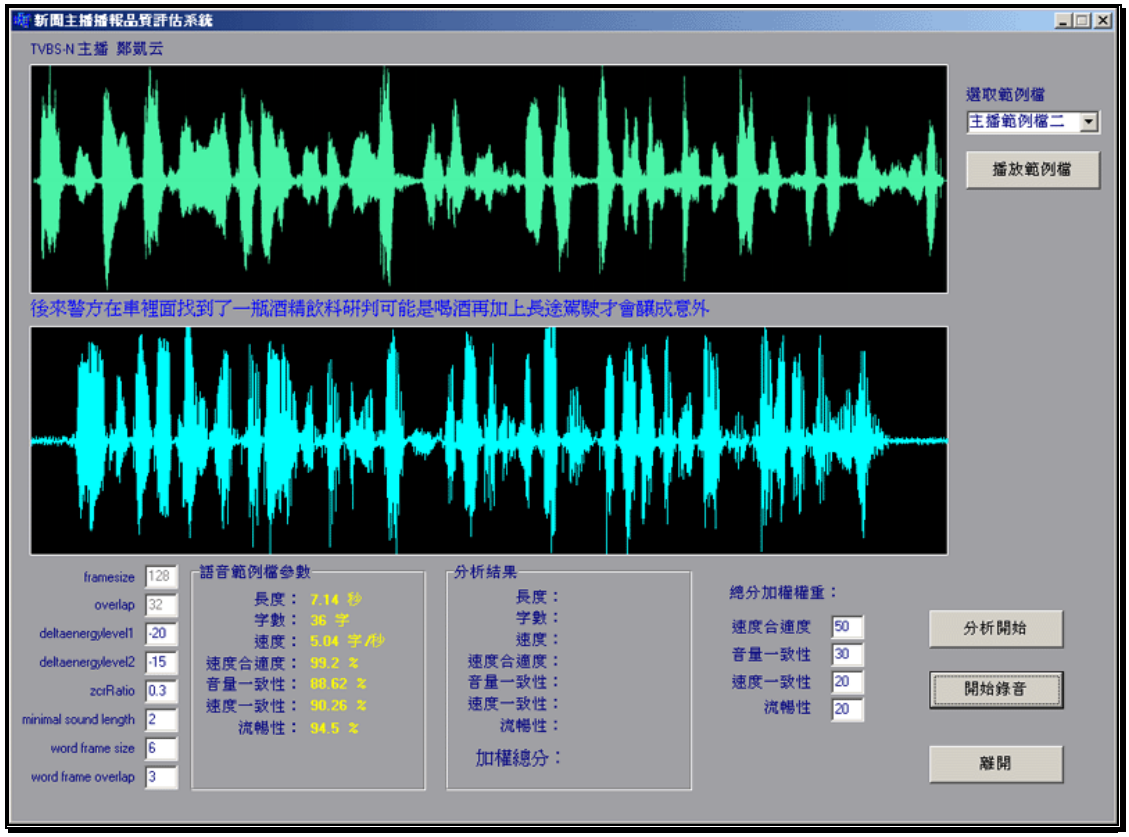
使用者經由麥克風讀入一段語音後，按下〝開始分析〞，程式即會自動完成 segmentation 並根據使用者選擇的語音範例檔完成各個參數的分析，包括有速度、速度的一致性、音量、音量的一致性…等等。



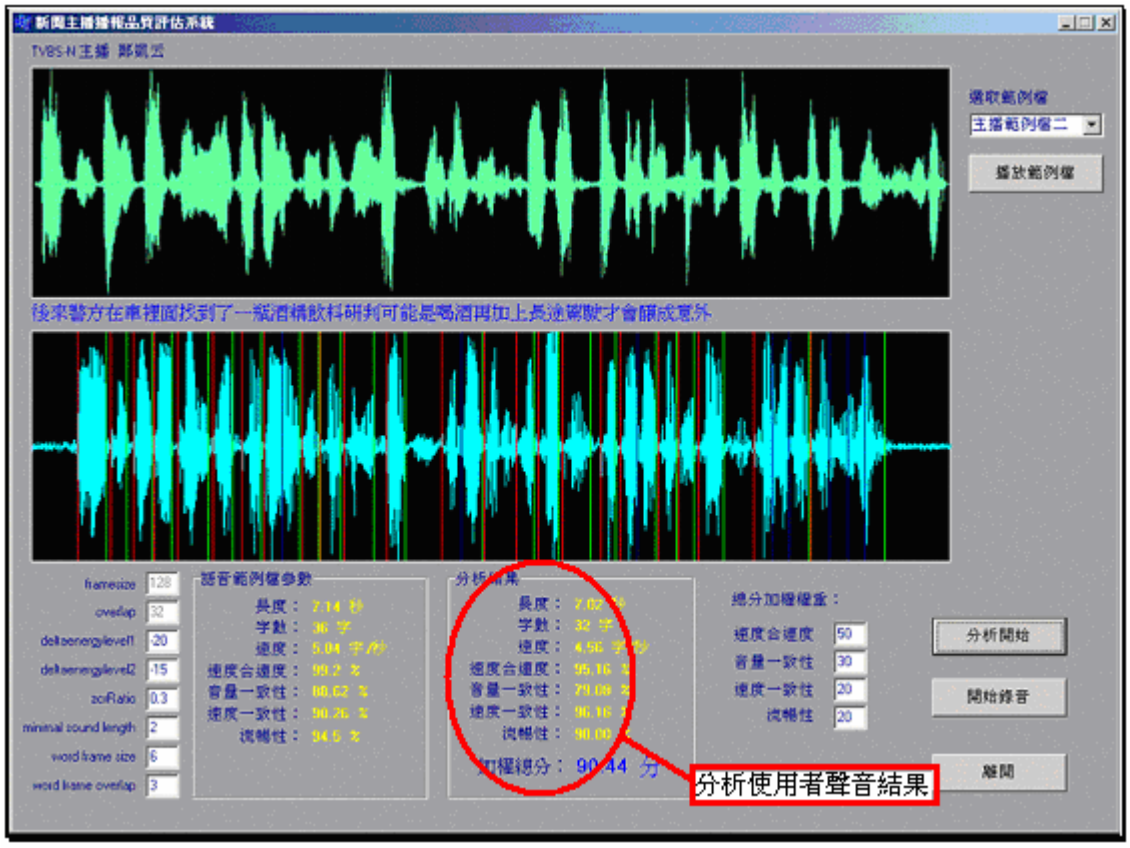
圖十三、使用者輸入聲音，再按一次即可停止



圖十四、按下分析開始

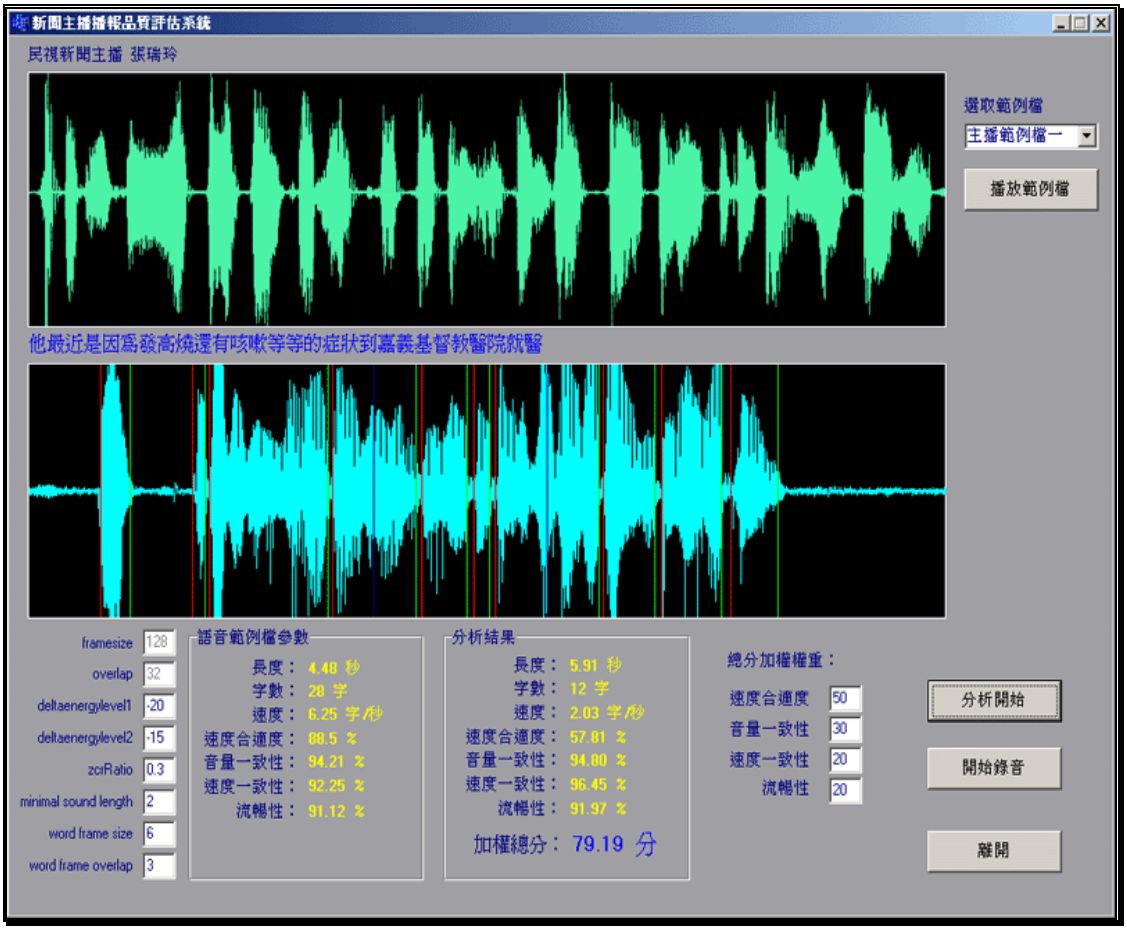

新聞主播播報品質評估系統
TVBS-N主播 鄭凱云
選取範例檔
主播範例檔二
播放範例檔
後來警方在車裡面找到了一瓶酒精飲料研判可能是喝酒再加上長途駕駛才會釀成意外
framesize 128
overlap 32
deltaenergylevel1 -20
deltaenergylevel2 -15
zcrRatio 0.3
minimal sound length 2
word frame size 6
word frame overlap 3
語音範例檔參數
長度： 7.14 秒
字數： 36 字
速度： 5.04 字/秒
速度合適度： 99.2 %
音量一致性： 88.62 %
速度一致性： 90.26 %
流暢性： 94.5 %
分析結果
長度：
字數：
速度：
速度合適度：
音量一致性：
速度一致性：
流暢性：
加權總分：
總分加權權重：
速度合適度 50
音量一致性 30
速度一致性 20
流暢性 20
分析開始
開始錄音
離開

圖十五、另外一則範例


新聞主播播報品質評估系統
TVBS-N主播 鄭凱云
選取範例檔
主播範例檔二
播放範例檔
後來警方在車裡面找到了一瓶酒精飲料研判可能是喝酒再加上長途駕駛才會釀成意外
framesize 128
overlap 32
deltaenergylevel1 -20
deltaenergylevel2 -15
zcrRatio 0.3
minimal sound length 2
word frame size 6
word frame overlap 3
語音範例檔參數
長度： 7.14 秒
字數： 36 字
速度： 5.04 字/秒
速度合適度： 99.2 %
音量一致性： 88.62 %
速度一致性： 90.26 %
流暢性： 94.5 %
分析結果
長度： 7.02 秒
字數： 32 字
速度： 4.56 字/秒
速度合適度： 95.16 %
音量一致性： 79.08 %
速度一致性： 96.16 %
流暢性： 90.00
加權總分： 90.44 分
總分加權權重：
速度合適度 50
音量一致性 30
速度一致性 20
流暢性 20
分析開始
開始錄音
離開
分析使用者聲音結果

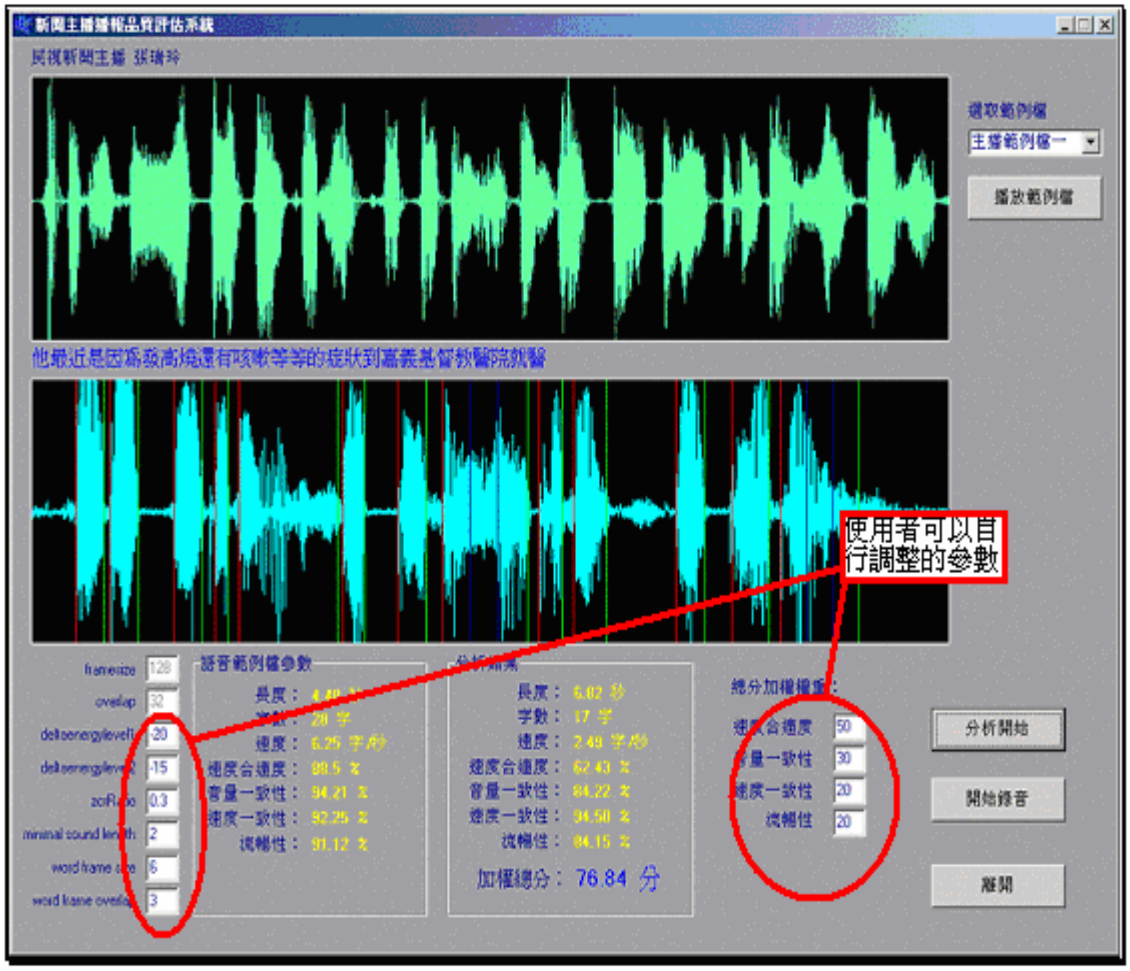
圖十六、分析之後的結果呈現

當使用者的聲音含糊不清如下，在 segmentation 部分就會切的不好，導致分數降低：



圖十七、上圖為咬字不清之時所生之波形

附帶一提的是，在介面中左下角及右下角的地方，權重等參數是使用者可以依當時情況而手動調整的。



圖十八、使用者可自訂調整部分

5、結語

在這樣的一個過程之中，我學習應用系統開發、整合與維護等技巧，這對我個人來說是一個極為特別的經歷。也讓我深深的知道，很多時候在人認為很直覺、主觀的東西，用電腦去實作起來有多麼的不容易，也藉著這個機會讓我對廣電方面的知識稍有提升，增廣了我的見聞。在這個計畫之中只考慮到了聲音的顯性特徵部分，另外還有隱性特徵需要開發，什麼樣的聲音是悅耳動聽，什麼樣的聲音可以讓人覺得精神抖擻，這都是屬於隱性特徵的部分，無法直接由聲波中觀察得知，因此需要與 data mining 的技術相結合，加以深入研究。此外有些時候主播的表情也是很重要的，我想這是我所沒有做到的部分，或許以後有機會可以繼續深入研究。

6、未來目標

考量程式之相容性，或許可以設計一透過 Web 介面傳送測試之音檔，透過後端處理後再將結果送回，以網頁方式呈現之應用系統，但由於音檔檔案較純文字檔大很多，將嚴重限制服務對象之數目與及系統反應時間，所以將會考慮採取以非即時(non real-time)的服務方式，以解決上述問題。

7、參考文獻

[1] Gerhard, David B. (2000) “Audio Signal Classification: An Overview. “ Canadian Artificial Intelligence. Winter 2000. 4-6.

[2] Jonathan Foote.”An overview of audio information retrieval” Institute of Systems Science, National University of Singapore, Heng Mui Keng Terrace, Singapore 119597
http://www.ee.columbia.edu/~mjr59/reviews/foote_overview.pdf

[3] Kay Berkling, Marc Zissman Julie Vonwiller, Chris Cleirigh “Improving accent identification through knowledge of English syllable structure.” M.I.T., Lincoln Laboratory 244 Wood Street Lexington, MA 02420-9108, USA. University of Sydney,Dept. of Electrical Engineering, Sydney, Australia.
http://www.clsp.jhu.edu/ws99/projects/asr/biblio/SL980394.PDF

[4] L. Arslan, J.H.L. Hansen, "Language Accent Classification in American English." Speech Communications, vol. 18(4), pp. 353-367, July 1996.
http://cslr.colorado.edu/rspl/PUBLICATIONS/PDFs/15-SpchComm-Accent.Jul96.PDF

[5] P. Angkititrakul, J.H.L. Hansen, "Stochastic Trajectory Model Analysis for Accent Classification”, ICSLP-2002:Inter. Conf. on Spoken Language Processing, vol. 1, pp. 493-496, Denver, CO USA, Sept. 2002
http://cslr.colorado.edu/beginweb/ICSLP2002/p1807.pdf

[6] Lie Lu, Hao Jiang and Hong-Jiang Zhang, ” A Robust Audio Classification and Segmentation Method. ” Microsoft research, China 5F,Beijing Sigma Center.
http://www.acm.org/sigs/sigmm/MM2001/ep/lielu/

[7]Jonathan Foote, ”A similarity measure for automatic audio classification.” Institute of Systems Science National University of Singapore. Singapore 119597.

[8]Jonathan T. Foote, “Content-based retrieval of music and audio.” Institute of Systems Science National University of Singapore. Heng Mui Keng Terrace, Kent Ridge. Singapore 119597.
http://www.cs.princeton.edu/courses/archive/spr99/cs598b/foote.pdf

[9] Speech Enhancement: quality assessment methods(lecture 9)
http://cslr.colorado.edu/classes/ECEN5022/sched2.html

[10] S.Umesh, L.Cohen, N.Marinovic, and D.Nelson, "Frequency-Warping in Speech," in Proc. International Conference on Spoken Language Processing , (Philadelphia, USA), THP2L2.2, 1996.
http://www.asel.udel.edu/icslp/cdrom/vol1/530/a530.pdf

[11]S.Umesh, L.Cohen, N.Marinovic, and D.Nelson, "Scale Transform In Speech Analysis," IEEE Trans. on Speech and Audio Processing , Jan. 1999.

http://home.iitk.ac.in/~sumesh/publications.html

[12]H. Kim, K. Obermayer, M. Bode, and D. Ruwisch.”Real-time noise cancelling based on spectral minimum detection and diffusive gain factors.” In Proceedings of the 8th Australian International Conference on Speech Science & Technology, pages 256-261, 2000.
http://citeseer.nj.nec.com/correct/410500

[13]Dong Wang, Lie Lui, Hong-Jiang Zhang “Speech segmentation without speech recognition.” Department of Electronic Engineering, Tsinghua University, Beijing.
http://research.microsoft.com/users/llu/Publications/ICASSP03_SenSeg.pdf