語音情緒辨識技術與應用之研究
Speech Emotion Recognition and its Applications

李俊昇*　黃珠娟**　許馨仁***　林明慧****
Jiun-Sheng Li*　Chu-Chuan Huang**　Shin-Tzen Sheu***　Ming-Wheng Lin****

*工業技術研究院南分院人機互動科技中心　副工程師
**工業技術研究院南分院人機互動科技中心　副工程師
***工業技術研究院南分院人機互動科技中心　副工程師
****工業技術研究院南分院人機互動科技中心　正工程師

摘　要

本研究針對語音在情緒上的辨識之技術與其相關應用進行討論。在語音情緒辨識技術上，主要計算音高(pitch)，共振峰(formant)，音框能量(frame energy)以及梅爾倒頻譜係數(Mel-scale Frequency Cepstral Coefficients, MFCC)等與語音情緒相關的特徵參數，利用支持向量機(Support Vector Machine, SVM)分類器，依特徵參數對情緒進行分類。根據實驗結果本研究在語者獨立的情形下之辨識率爲 63.8%；語者相關的辨識率爲 86.8%。

關鍵詞：語音辨識、情緒辨識、語音情緒

一、 導論

在人與機器互動的過程裡加入情感的因子，除了可以增進人機互動的樂趣[1]，亦可以提升人機互動的成效。藉由情感因子的加入人對機器的感受可由較爲淺層的心理愉悅進而延伸至深層的心靈慰藉，讓機器產生擬人化之效果。而且因爲機器懂得人的情感，它可以提供給人們更適合的服務與回應，而不是一直給予制式的回應，可以提升人們使用之意願。一般主要應用可分爲幾個部分：

1. 智慧型電腦：情緒偵測的賦予使得智慧型電腦在和人類進行溝通時，擁有近似人類的表現。若電腦可以像人一樣具有互動行爲能力，那在日常生活脈絡下，智慧型電腦的應用將能夠積極地彌補人們生活的缺陷，提升生活的品質。
2. 智慧型人機互動產品：將人機互動產品賦予情緒辨識的功能時，不但可以增加使用產品的樂趣，更可以幫助使用者從事特定活動。

本論文針對語音情緒辨識技術來進行研究，期望藉由使用者的語音，來識別使用者目前之情緒，以增進人機互動過程中的豐富度，以往在語音的情緒辨識上，主要是選擇帶有情緒資訊如聲韻(Prosody)以及能量相關的特徵，最常見的情緒特徵是音高(Pitch)和能量(Energy)[2-5]，有些文獻則採用共振峰(Formant)[4]，在辨識上主要是根據這些特徵的統計值，如平均值、標準差、最大值、最小值、梯度變化等特徵當作不同情緒分類的特徵。Schuller et al.[5]根據線性分析對不同統計方式的特徵進行排名，X.H. Le et al. [6]以及 T.L. Pao et al.[7] 採用梅爾倒頻譜係數(Mel-scale Frequency Cepstral Coefficients, MFCC)以及線性預估參數(Linear predictor coefficient, LPC)，並透過分類方法辨識出情緒。在情緒辨識上，音高相關的特徵明顯比能量相關的特徵更能分辨出不同的情緒種類。而在辨識分類方法上，大部分則採用包括類神經網路[5]、隱藏式馬可夫模型(Hidden Markov Model, HMM)[2][3][7]、高斯混合模型(Gaussian Mixture Model, GMM)[5][6]、線性鑑別分析[7]等。

二、 特徵擷取

當我們去判斷他人情緒時，我們通常會經由臉部表情以及說話的方式、語調以及說話內容來判定情緒，本研究內容以語音爲情緒辨識之訊號來源，在語音訊號中，情緒通常包含兩部分，一個是語調，另一個則是語意，語意通常需加入語音辨識，再從文字去解讀情緒反應；本研究採取另一種辨識方式，希望藉由語調變化來分析講話者的情緒反應，語調變化包含幾個重要的特徵，以下將就幾個我們使用的語音特徵與語音計算方法加以說明。

1. 加視窗(window)與取音框(frame)

將一段語音離散時間訊號 x(n)，用固定長度

的視窗(window)套上去，只看視窗內的訊號，對此視窗內的訊號作運算，用以求出在此視窗內的語音特徵(speech features)。這樣的處理方式，就叫做加視窗(windowing)，而此段語音即稱爲音框(frame)。通常視窗的長度是取 15~30 毫秒(ms)，用以計算語音的特徵參數(feature parameters)。視窗與視窗之間的移動距離，大約會取 5~20 ms，讓前後的音框有部分重疊，這樣比較能看到語音特徵改變的延續性。例如對一個取樣頻率 16kHz 的語音訊號來說，其取樣間距爲 62.5 微秒(μs)，若我們的音框長度取 256 點，相當於 16 ms，我們讓前後音框重疊 1/2 音框長度，則每次移動視窗的距離就是 8 ms，也就是 128 個取樣點。圖一爲一段語音訊號及其中一個 frame，其取樣長度爲 256 點。

2. 音高(Pitch)

音高指的是人類心理對音符基頻 f（fundamental frequency）之感受。中央 C 上之 A 音符發出的頻率爲 440Hz，通常被當作「標準音高」。一般而言，計算基頻變化的研究有很多，本研究選用了 Auto-correction function (ACF)的方法來當作基頻的計算。其方法說明如下，首先，將我們所選取的 frame 平移 n 點，然後將兩音框重疊部分做內積，找尋除了第一點外的 local maximum，將此點的位置換算成對應的頻率，此頻率即爲音高。圖二即爲原始聲音取 256 點經過 ACF 計算之結果，下圖取第一點除外之最大值對應之頻率值即爲 Pitch。

3. 共振峰(formant)

在有基頻諧振峰值的頻譜上。取其包絡線而得出一條較爲緩和的頻譜曲線，可以看到若干個高點，這些高點顯示能量集中的頻率位置，這就是共振峰(formant)之所在。如圖三所示，頻譜曲線中的高點，就是共振峰之所在。我們習慣上從低頻算起，第一個共振峰標示爲 F1，第二個共振峰標示爲 F2，依次類推。如果以人發音的頻率範圍來看(20~4kHz)，一般會有大約四至五個共振峰值。

4. 音框能量(Frame energy)

在語音特徵中，聲音強度的變化是相當重要的訊息，聲音強度與波形振幅有關，振幅越大音強(intensity)越大，在固定音框長度情況下，計算音框能量可以表示爲：

$$E_x(m)=\sum_{n=m-N+1}^{m}\left|x(n)^2\right|$$

(2.4.1)

由於人耳對音強的感知並非線性，而是接近於對數的曲線，將能量以對數方式表示，其計算式如下所示：

$$EL_x(m)=\log\left[\sum_{n=m-N+1}^{m}\left|x(n)^2\right|\right]$$

(2.4.2)

其中，x 表示第幾個音框，從訊號起始端開始，第一個音框編號爲 1，以此類推，可以得到一音框能量變化的序列，寫成 Ex，圖四爲一語音訊號的能量曲線。

5. 梅爾倒頻譜係數(Mel-scale Frequency Cepstral Coefficients, MFCC)

人耳在頻域上的感知並非全頻域有相同的敏感度，在正常的情況下，對於低頻有較高的解析度，也就是在低頻可以分辨較小的頻率差異，此外還有臨界頻帶的現象，在 1kHz 頻率以下的臨界頻帶寬度約爲 100Hz，1kHz 頻率以上的臨界頻帶寬度成指數增加。因此，配合人耳聽覺特性，在頻域中以梅爾(mel-frequency)劃分頻帶，將屬於一個頻帶中的頻率成分，合在一起當作一個能量強度，然後將這些頻帶強度，以離散餘弦轉換(DCT)，轉換成倒頻譜，其轉換方法如下，首先，設計一組梅爾頻率的帶通濾波器，來得到通過帶通的音強，用以計算倒頻譜，圖五爲三角形濾波器所組成的梅爾濾波器組，梅爾頻率刻度是以 1kHz 以下爲等間距，1kHz 以上爲對數間隔，在 4kHz 範圍內設計成 20 個頻帶，其中心頻率設定成；100，200，300，400，500，600，700，800，900，1000，1148，1318，1514，1737，1995，2291，2630，3020，3467，4000Hz。以數學式表示，第 m 個濾波器的函數式如下：

$$B_m(k)=\begin{cases}0 & k<f_{m-1}\\ \dfrac{k-f_{m-1}}{f_m-f_{m-1}} & f_{m-1}\le k\le f_m\\ \dfrac{f_{m+1}-k}{f_{m+1}-f_m} & f_m\le k\le f_{m+1}\\ 0 & f_{m+1}<k\end{cases}\qquad 1\le m\le M$$

(2.5.1)

Bm(k)表示是第 m 個頻帶的三角形濾波器，fm 爲第 m 個頻帶的中心頻率，fm-1 與 fm+1 就是前後兩個頻帶的中心頻率，M 爲全部的頻帶數目。

將各頻率的能量，乘上三角形濾波器，然後累加起來，就是通過這個濾波器的能量，取對數值，得到

$$Y(m)=\log\left[\sum_{k=f_{m-1}}^{f_{m+1}}\left|X(k)^2\right|B_m(k)\right]$$

(2.5.2)

對全部 M 個濾波器輸出的對數能量，做離散餘弦轉換(discrete cosine transform, DCT)，得到梅爾頻率倒頻譜。

$$c_x''(n)=\frac{1}{M}\sum_{m=1}^{M}Y(m)\cos\left(\frac{\pi n\left(m-1/2\right)}{M}\right)$$

(2.5.3)

$c_x''(n)$就是訊號 x(n)的梅爾倒頻譜係數(Mel-frequency Cepstral coefficient, MFCC)，本研究用前 13 個係數，即 n = 1,2,3……,13 作爲語音情緒辨識的頻譜特徵。

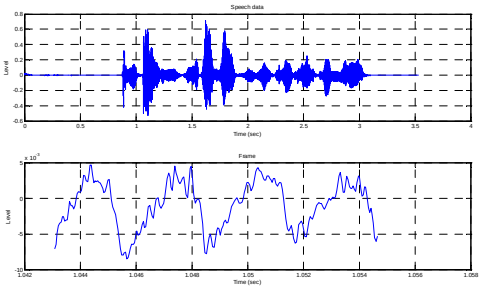

圖 1 語音訊號與單一 frame

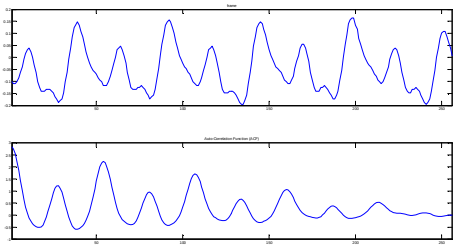

圖 2 ACF

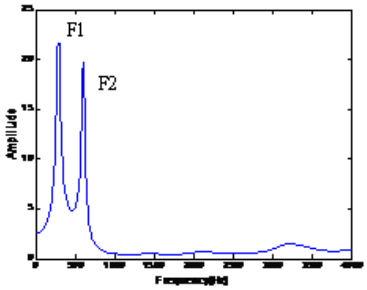

圖 3 Formant

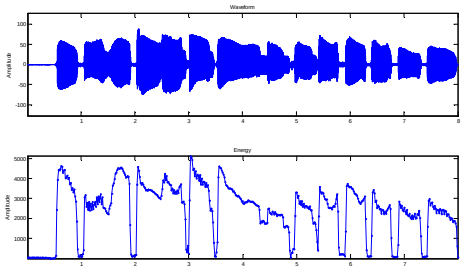

圖 4 波形與能量圖

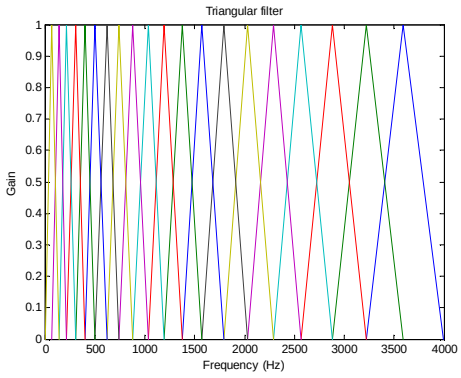

圖 5 三角形濾波器組成的梅爾頻率濾波器組

三、語料庫及實驗流程

在日常生活中，最容易表現出來也最常會流露出來的情緒不外乎是開心、生氣、悲傷及無情緒，在此無情緒指的是平靜不帶其它情感，故本研究選定此四種情感來做爲識別之情緒。

本研究辨識使用者之語音情緒之流程爲：在訓練階段時，首先我們把語料庫之所有語句取出，將語音訊號作前處理，將沒有語音或非語音訊號部分去除，然後進行特徵擷取，其特徵分別爲 pitch、energy、formant 的平均值及標準差以及 MFCC 的前 13 個係數的平均值共 23 個特徵來代表此語音訊號，並將這些特徵透過 SVM[8]訓練出一辨識模型，此一模型即爲之後辨識受測者語音情緒之核心；而在辨識階段時，同樣的將受測者沒有語音或非語音訊號部分去除，然後進行同樣的特徵擷取，取得 23 個特徵來代表此語音訊號，最後讓分類器根據先前所訓練出來的辨識模型來識別受測者之語音情緒。訓練及辨識階段之流程如圖 6、圖 7 所示。

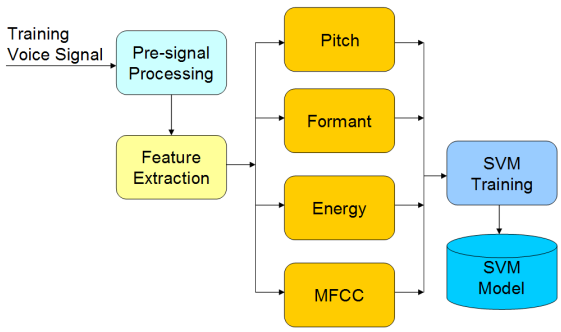

圖 6 語料庫訓練流程

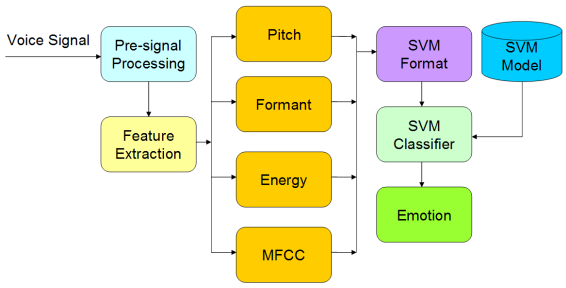

圖 7 語音情緒辨識流程

先前提到本研究之方法爲使用者之語音特徵透過分類器與資料庫進行比對，藉此來判斷使用者之情緒，由於目前尚未有公認的中文語音情緒資料庫，故本研究自行錄製語料庫，來做爲實驗測試與分析之資料。本研究自行錄製之語料庫的細項如表 1 所示。

表 1 語料庫說明

項目	時間
語者個數	共 20 位，P1~P10 爲女生、P11~P20 爲男生)
語者年紀	約 25 至 35 歲
情緒類別	共 4 種情緒分別爲中性(無情緒)、喜(開心)、怒(生氣)、哀(悲傷)
語句個數	每個語者各個情緒皆錄製 30 句，故共 20×4×30=2400 句

四、實驗結果

在實驗結果部份，本研究分成二個部份討論，分別爲語者獨立與語者相關，所謂的語者獨立簡單來說就是資料庫裡並沒有受測者的資料，通常只要使用者在使用語音系統前不需事先錄製本身之語料即可馬上立即使用，則此語音系統屬於語者獨立之系統。語者相關則是資料庫裡含有受測者部份的語音資料，語者相關之系統通常在使用前會先要求使用者講幾句系統所設定之語句，之後系統將會對使用者之語料進行處理，並將其納入資料庫中，語者相關之系統可減少資料不匹配的效應，以期系統有較高的辨識率。

在語者獨立實驗的部份，我們將受測者的資料從 20 人的情緒資料庫取出，並以剩餘 19 人的資料來做爲訓練的資料，而受測者資料爲測試資料，如此重複 20 次，每個人的資料皆會當一次測試資料，用以測試受測人員資料不在訓練資料內時其辨識率結果。實驗結果爲每個語者的辨識率範圍在 35%至 86%之間，而平均辨識率爲 63.8%。詳細的辨識結果如表 2 所示。

在語者相關實驗的部份，我們將受測者的一半的資料做爲訓練資料，另一半的資料來做爲測試資料，用以測試當受測人員在訓練資料裡時其辨識結果。實驗結果爲每個語者的辨識率範圍在 70%至 94%之間，而平均辨識率爲 86.75%，詳細的辨識結果如表 3 所示。由實驗的結果可以得知，當訓練資料包含使用者時，其辨識率有顯著的提升。

根據語者獨立及語者相關的混沌矩陣(Confustion Matrix)如表 4、表 5 所示，可以得知當語料庫裡沒有受測者的資料時，喜容易辨識爲怒，而哀跟中彼此容易誤辨，會造成這樣的原因可能是喜跟怒在語音上的表達都是屬於音量上較爲大聲的，而哀跟中是屬於音量上較爲小聲的，而本研究所使用的特徵 Energy、MFCC 的對數能量係數與音量(能量)有關，才會造成誤辨的情形，但若資料庫裡若有部份自己的資料時，也就是語者相關之情形下，則可藉由自身之資料，去修正辨識核心模型，提升原本易混淆情緒之鑑別度。

表 2 語者獨立辨識率

P1	P2	P3	P4	P5
74.17%	50%	64.16%	78.33%	70%
P6	P7	P8	P9	P10
78%	62.5%	51.67%	85.83%	70%
P11	P12	P13	P14	P15
85%	82.5%	55.83%	72.5%	57.5%
P16	P17	P18	P19	P20
50.83%	35%	62.5%	51.67%	37.5%
平均辨識率		63.79%		

表 3 語者相關辨識率

P1	P2	P3	P4	P5
78.33%	93.33%	78.33%	91.67%	88.33%
P6	P7	P8	P9	P10
90%	78.33%	86.67%	93.33%	88.33%
P11	P12	P13	P14	P15
91.67%	90%	91.67%	85%	88%
P16	P17	P18	P19	P20
85%	95%	81.67%	90%	70%
平均辨識率		86.75%		

表 4 語者獨立之混沌矩陣

	喜	怒	哀	中
喜	401	130	20	49
怒	32	379	14	22
哀	32	18	445	105
中	30	18	246	306

表 5 語者相關之混沌矩陣

	喜	怒	哀	中
喜	264	34	1	1
怒	39	253	5	3
哀	2	2	260	36
中	2	1	33	264

表 6 語者獨立與語者相關實驗的訓練及測試句數

	全部的訓練句數	每個語者的測試句數
語者獨立實驗	19(語者)×4(情緒)×30(語句數)=2280 句	1(語者)×4(情緒)×30(語句數)=120 句
語者相關	20(語者)×4(情緒)×15(語句數)=1140 句	1(語者)×4(情緒)×15(語句數)=60 句

五、結論

隨著電腦系統與資訊技術的發展，使用者介面也不斷推陳出新。而其中重要的趨勢之一，是越來越多的自動化服務與回應機制取代了原本人工應答服務，而這類服務應用最廣泛的莫過於電話語音回覆。自動化回應服務可提供精確、客觀的資訊，並節省大量的人工成本，提升系統效率。然而，精確客觀的資訊未必是對使用者最有效的資訊。在客服專員以人工方式與人應答時，除了確切掌握對方言語中的需求外，同時要感受對方當下的情感資訊，才夠在提供客觀資訊的同時，適切地給予問候、道歉甚至是鼓勵等話語。因此未來在自動化的應答服務中，如能掌握對方情感資訊，進而在應答內容中適度加入具同理心的語調和詞句，必能增進此服務在感性訴求上的提升。其他應用領域如大型機台的情人告白甜言蜜語製造機、獨居老人情緒分享應用、兒童音樂學習情緒辨識…等，除了可延伸應用至娛樂玩具與互動遊戲機上，增加玩具或遊戲的娛樂性。在智慧居家生活科技與智慧照護系統上，也有相當大的應用性。

依據本研究的實驗結果，使用者語音情緒的辨識率可達到 63%~86%，尤其語者相關的辨識率爲 70%~94%，有機會達到商品實用水準，從而佐證採用音高、共振峰、音框能量和梅爾倒頻譜係數等特徵，搭配 SVM 的向量支持分類器，是具潛力的語音辨識解決方案。因此，未來自動應答系統可先從蒐集或核對使用者基本資料的過程中，先進行個人化參數的調整與訓練，掌握使用者的個人化語音特徵，讓語者獨立的狀況轉變爲語者相關的狀態。而實際服務應答時則可加入語音情緒的辨識機制，辨認出使用者的情緒參數，並依據此參數調整合成語調或語句，給予較適當的回應。

但根據本研究所使用的語料訓練方式，以及國內外相關研究所採用的語料訓練方法，不容易動態地加入單筆或少數語料，並快速地完成分類器訓練達到可辨識的狀態。因此，本研究的下一階段工作，將著重於研究個人化參數的動態調整與訓練機制，期望爲智慧化語音應答技術累積更多的能量。

六、參考文獻

1. P.W. Jordan(1998).”Human factors for pleasure in product use”, *Elsevier Science Ltd*, vol. 29, pp. 25-33.
2. D.N. Jiang and L.H. Cai(2004).“Speech Emotion Classification with the Combination of Statistic Features and Temporal Features”, *IEEE International Conference on Multimedia and Expo* , Taipei, Taiwan, vol. 3, pp. 1967-1970.
3. B. Schuller, G. Rigoll and M. Lang(2003). “Hidden Markov Model-based Speech Emotion Recognition”, *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, China, vol. 2, pp. 1-4.
4. D. Ververidis, C. Kotropoulos and I. Pitas(2004).“Automatic Emotional Speech Classification,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Quebec, Canada, vol. 1, pp. 593-596.
5. B. Schuller, G. Rigoll and M. Lang(2004). “Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture”, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Quebec, Canada, vol. 1, pp. 577-580.
6. X.H. Le, G. Quénot and E. Castelli(2004). “Recognizing Emotions for Audio-Visual Document Indexing," *Proceedings of 9th Symposium on Computers and Communications*, Alexandria, Egypt, vol. 2, pp. 580-584.
7. T.L. Pao, Y.T. Chen and J.H. Yeh(2004). “Emotion Recognition from Madarin Speech Signals,” *Proceedings of IEEE International Symposium on Chinese Spoken Language Processing*, Hong Kong, pp. 301-304.
8. C.C. Chang and C.K. Lin, LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Speech Emotion Recognition and its Applications

Jiun-Sheng Li* Chu-Chuan Huang** Shin-Tzen Sheu*** Ming-Wheng Lin****

*Human Computer Interaction Technology, ITRI South, Industrial Technology Research Institute, lgs@itri.org.tw
** Human Computer Interaction Technology, ITRI South, Industrial Technology Research Institute, chuchuan@itri.org.tw
*** Human Computer Interaction Technology, ITRI South, Industrial Technology Research Institute, SerinaSheu@itri.org.tw
**** Human Computer Interaction Technology, ITRI South, Industrial Technology Research Institute, lmw@itri.org.tw

Abstract

This paper proposed a speech emotion recognition method and its applications. Several speech features such as pitch, formant, frame energy, and Mel-scale frequency cepstral coefficients (MFCC) are considered in the proposed system. Support vector machine (SVM) is used to classify speeches into four emotions. The experimental results showed the proposed system performed 63.8% in outside tests and 86.8% in inside tests.

keywords：*Speech Recognition*、*Emotion Recognition*、*Speech Emotion*