The 2015 Conference on Computational Linguistics and Speech Processing
ROCLING 2015, pp. 306-315

結合非線性動態特徵之語音情緒辨識

Speech Emotion Recognition via Nonlinear Dynamical Features

林竹萱　Chu-hsuan Lin
美律實業股份有限公司
Merry Electronics Co.,Ltd.
tracy.lin@merry.com.tw

陳炎生　Yen-Sheng Chen
美律實業股份有限公司
Merry Electronics Co.,Ltd.
daryl.chen@merry.com.tw

摘要

本研究採用機器學習法對語音情緒辨識進行探討。除一般常被採用之語音特徵，如音高、共振峰、能量以及梅爾倒頻譜係數之外，研究中加入了夏農熵和曲率指標(curvature index)[9]兩項非線性特徵，再利用費雪鑑別比與基因演算法搭配的方式進行特徵挑選。最後使用支持向量機分類器，對柏林語音情緒資料庫進行情緒分類分析。在加入非線性特徵後，男性及女性之情緒辨識率分別為 88.89%及86.21%。

Abstract

This study is focus on speech emotion recognition through machine learning method. We add two nonlinear dynamical features: Shannon entropy and curvature index, of each frame other than the traditional features such as pitch, formant, energy, MFCCs. After feature extraction, Fisher discriminant ratio and Genetic algorithm were applied in order to reduce the number of features. We use SVM classifier and cross validation method to discriminate seven emotions in Berlin emotion database. The analyzed results after adding of the nonlinear features show that the emotion recognition rates were 88.89% and 86.21% for male and female, respectively.

關鍵詞：情緒辨識、非線性特徵、支持向量機

Keywords: Speech emotion recognition, non-linear features, support vector machine

一、緒論

在人工智慧、機器學習與網路資訊的快速發展下，在不同領域都已經有許多事情可以由機器取代，如會議安排、語言學習、語音服務、新聞播報、汽車駕駛等等，但如果僅僅只是由機器單方面提供制式化的回應服務，或許不是那麼適當，因此讓機器偵測得人類所要表達的情緒訊息，接著給予最適當的回應是一項重要的機制。這不僅僅可以增進人機互動的樂趣，也可在一般客服機器提供客觀資訊外，給予適切地問候話語；在智慧家庭與照護系統方面，若可得知使用者當下情緒而做出反應，如切換音樂、燈光控制等等，可以提升人機互動的成效；其他像是娛樂產品的介面也是可以應用的主題。目前在機器與人的互動上，基本上可利用視覺與聽覺兩種人類感官，本研究著重於聽覺之語音情緒辨識系統，期望藉由語音訊號來分辨使用者目前的情緒，進而提升溝通效果。

對於情緒的描述方式大致可分為離散與維度兩種形式，前者即為日常生活所使用之詞彙，如開心、生氣、悲傷等，在如此大量之情感詞彙中，一般認為能夠為人類與具有社會性之哺乳動物所共有情感稱為基本情感，不同學者對於基本情感的定義也不相同，其中以 Ekman 提出之六大基本情感較為廣泛被使用，當然亦有許多依此發展或其他理論而形成的基本情緒，如下表一[1]；後者則將情感狀態描述於激活度-效價情感空間(arousal-valence emotional space)或是激勵-效價-控制空間(activation - valence -dominance space)中，其中每一個維度對應著心理學的屬性[2、3]。基本上，透過聲音來傳遞情緒上大致可分為兩個方向，一為透過語意，即由字面上的意思；另外是藉由語調來傳遞情緒。而在本研究中則採用了離散情緒分類及透過語調來擷取特徵，進而作情緒分類判斷。

過去文獻中，Moataz El Ayadi 等人[4]提供不同語料庫收集方式之資訊及許多語音訊號特徵之計算方式與分類方法；Siqing Wu 等[5]利用調變頻譜特徵(MSFs)與不同特徵組合進行情緒分類，其最佳準確率 91.6%為 MSFs 與聲韻(prosodic)特徵的組合法；Patricia Henríquez[6]等利用非線性動態特徵進行語音情緒辨識研究，準確率最高可達 80.75%；Ali Shahzadi 等[7]以聲韻特徵、頻譜特徵與非線性動態特徵依不同組合進行研究，其準確率最高為男性 85.9%，女性為 82.72%。本研究的目標是透過分析語音來辨識情緒，以過去學者之研究為基礎，利用語音訊號擷取特徵量，再以挑選後的特徵量作為支持向量機(support vector machine, SVM)中的訓練資料，藉此訓練出分類模型，結果證明在一般常見語音特徵如音高(pitch)、能量(energy)、共振峰(formant)、梅爾倒頻譜係數(Mel-scale Frequency Cepstral Coefficients, MFCC)，額外加入了夏農熵(Shannon entropy)和曲率指標兩項非線性特徵有提升語音情緒辨識之效用。

表 一、基本情感之定義

學者	基本情感
Arnold	Anger, aversion, courage, dejection, desire, despair, dear, hate, hope, love, sadness
Ekman, Friesen, Ellsworth	Anger, disgust, fear, joy, sadness, surprise
Fridja	Desire, happiness, interest, surprise, wonder, sorrow
Gray	Desire, happiness, interest, surprise, wonder, sorrow
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
James	Fear, grief, love, rage
McDougall	Fear, disgust, elation, fear, subjection, tender-emotion, wonder
Mower	Pain, pleasure
Oatley, Johnson-Laird	Anger, disgust, anxiety, happiness, sadness Panksepp
Panksepp	Anger, disgust, anxiety, happiness, sadness
Plutchik	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise Tomkins
Tomkins	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Watson	Fear, love rage
Weiner, Graham	Happiness, sadness

二、研究方法

(一) 實驗資料庫

本研究的資料來自於德國柏林語音情緒資料庫(Berlin emotion database)[8]，其中包含了生氣(anger)、無聊(boredom)、厭惡(disgust)、害怕(fear)、開心(joy)、中性(neutral)和傷心(sadness)共七種情緒，由十位專業演員(五男、五女)各別演示上述七種情緒對應的句子所組成，共有 535 句語音訊號。

(二) 特徵擷取

將語音訊號進行音框(frame)的切割，通常視窗長度為 20~40ms，用來計算特徵參數，而為了讓特徵變化有延續性，會將部分視窗重疊(overlap)，本研究所使用之視窗長度為 32ms，重疊部分為 16ms。擷取的特徵分為兩部分，一為傳統使用之聲韻和頻譜特徵，另一部分則是非線性動態特徵 Shannon entropy 和 curvature index。

1. 聲韻特徵

在聲韻特徵中，收集了音高、能量、過零率(zero crossing rate,ZCR)、TEO(Teager energy operator)等常見語音分析特徵。音高擷取方式是使用 ACF(auto-correlation function)，但為了避免 ACF 的值介於一個不定的區間，將其正規化至 1 與-1 之間後，再搭配音量閾值判斷音高，即得$\mathrm{NACF}(\tau)=\frac{2\sum s(i)s(i+\tau)}{\sum s^2(i)+\sum s^2(i+\tau)}$。過零率即為訊號過零點的次數，一般而言其值在有語音的時候會比安靜或環境雜訊較大時低，因此本研究採用此方法搭配音量來判斷 voice activity ratio，voice activity ratio 即為一段訊號內有語音與無語音的比例(如下圖一)。TEO 則是在還原聲音經過氣管及人的腔體作用後所產生的語音訊號，$\mathrm{TEO}(s_i)=s_i^2-s_{i-1}s_{i+1}$，上述公式內之 s 即為一個音框內的原始訊號，i 表示第 i 點訊號。

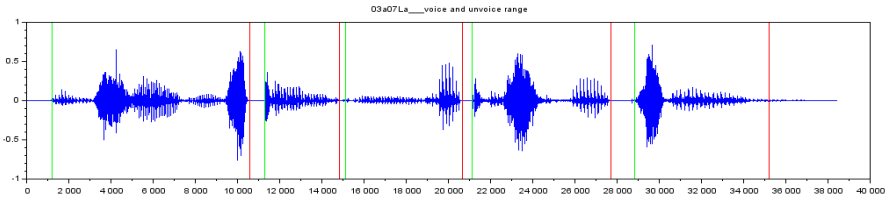


圖 一、 Voice activity detection，綠色線為起始位置，紅色線為結束

2. 頻譜特徵

頻域所使用之特徵，第一項為梅爾倒頻譜係數，配合人耳聽覺對不同頻率有不同的敏感度的特性，提出了這項係數；本研究所使用之 pre-emphasis 之高通濾波器參數為 0.9，共取 13 個梅爾倒頻譜係數。共振峰是將時域訊號轉為頻域後，取其包絡線(envelope)後可得到一條較為平滑的頻譜曲線，其中有若干個高點，這些高點表示能量集中的位置，也就是共振峰，可描述人類聲道中的共振情形(如下圖二)。本研究利用快速傅立葉轉換(FFT)及 linear predictive coding(LPC)方式取得第 1 到第 3 個共振峰(F1~F3)的頻率值及其頻寬。

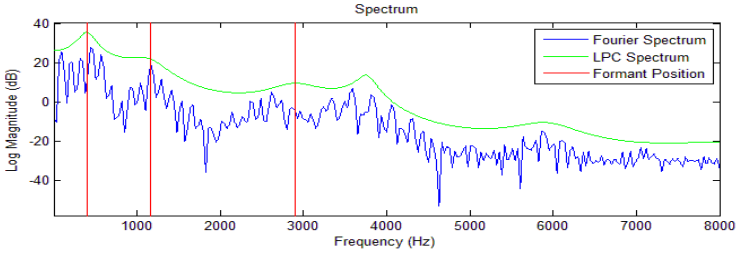


圖 二、Formant 結果

3. 非線性動態特徵

夏農熵在資訊理論中扮演了很重要的角色，除了可用來作為資訊量的量測外，同時也是對某個系統之不確定性或混亂程度的度量方法，若熵值越高則系統的不確定性(uncertainty)越高，反之亦然。隨機變數 的夏農熵可定義為

$$H(\ ) = -\sum_{\in} (\ ) \log (\ ) \ ,$$

其中 $(\ ) =$ $\{\ =\ \}$, $\in$ 。使用不同基底會有一轉換常數的差異。

曲率指標[9]是一動態系統的指標，曲率指標之定義如下，對於 維空間曲線 $(t) \in \mathbb{R}$ 可得 -1個高維度曲率$\kappa_i, 1 \le i \le -1$，則曲率指標為

$$K = \lim_{\to\infty} \frac{\int_0 (\ )dt}{\qquad}, 1 \le \ \le \ -1 \text{。}$$

由上式可知，曲率指標是藉由動態平均的方式來描述，其功用在於系統出現結構變化時，可以在指標上出現相應變化，是以吾人預期，當不同情緒變化表現在語音訊號時，其對應的曲率指標也會有所不同。計算曲率指標前，需要運用相空間重構的技術將語音訊號重構到高維度空間上，本研究中重構維度 $= 3$，且只有 K_1在特徵挑選過程中被選中。

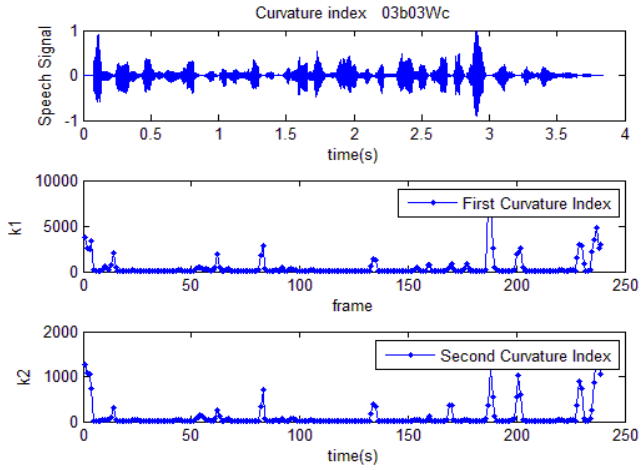


圖 三、 Curvature index 計算結果

4. 統計值

計算語音訊號每個音框的上述特徵值後進行統計，其統計量包含最小值(min)、最大值(max)、最大與最小值的差(range)、平均(mean)、中位數(median)、切尾均值(trimmed mean)之 10%與 25%、第 1、5、10、25、75、90、95、99 的百分位數(percentile)、四分差(interquartile range)、平均差(average deviation)、標準差(standard deviation)、偏態(skewness)和峰度(kurtosis)共 20 項。另外也計算相鄰兩音框之一階與二階倒數之統計量，以表示兩音框間的變化程度，最後將所有統計量當作語音訊號之特徵進行挑選與分類。

(三) 特徵挑選

特徵選取的目標是要從原有的特徵集合中挑選出鑑別能力較好的特徵，使其辨識率能夠達到最高值，不但能夠簡化分類器的計算，並可藉此了解分類問題關係。特徵挑選時使用了 10 折交叉驗證(10-fold cross validation)，避免對單一資料形成 over-fitting。

本研究利用了費雪鑑別比(Fisher discriminate ratio, FDR)與基因演算法(genetic algorithm, GA)進行特徵挑選。依據費雪判別分析的概念，分屬二個類別的特徵其組內差距越小，組間差距越大，可獲得越好的分類效果。多組類別之 FDR 計算方式如下[10]

$$\mathrm{FDR}(u)=\frac{2}{C(C-1)}\sum_{c_1}\sum_{c_2}\frac{\left(\mu_{c_1,\mu}-\mu_{c_2,\mu}\right)^2}{\sigma_{c_1}^2+\sigma_{c_2}^2},1\le c_1<c_2\le C,$$

利用 FDR 將不適用之特徵排除後，再經由 GA 挑出最後辨別所使用的特徵，GA 是人類依照生物學中「適者生存，不適者淘汰」的觀念所發展出來的一種演算法，利用選擇(selection)、複製(reinsertion)、交配(cross-over)、突變(mutation)等步驟

去尋找最適合環境的基因[11]。在本研究中即是將所有特徵的集合視為染色體，各個特徵即為基因，利用 GA 搭配 SVM 分類器，最後會得到一串 0 與 1 的序列，若為 1 則代表此特徵被選中[12]，反之亦然。其中 GA 挑選方式及使用的參數如下表二[7]。

表 二、 GA 參數設定

Selection technique	Roulette wheel
Crossover type	Single point crossover
Population size	50
Crossover rate	0.9
Mutation rate	0.001
Iteration number	200

(四) 分類方式

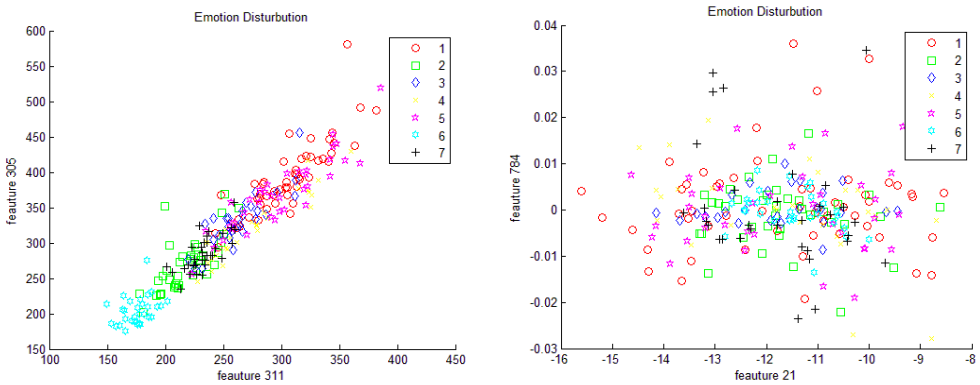
在特徵擷取前，已將資料以 80%與 20%的比例分為訓練資料集(training data set)與驗證資料集(validation data set)，驗證資料集內所有資料皆不會經過挑選與分類，而是作為訓練模型好壞的判斷依據，本研究所使用之分類器 SVM，採用的 toolbox 為 LibSVM [13]。

SVM 是一種機器學習的演算法，目的是為了建立一個模型以辨別不同資料的類別，利用 SVM 搭配核方法(kernel method)可以有效率地將原始資料轉換到高維度的空間，並在訓練資料集中找出餘裕(margin)最大的超平面(hyper-plane)，此 hyper-plane 將會是測試資料的分類依據，透過此方法我們可得到一個準確率高且具有高抗雜訊功能的分類模型，另外相較於其他機器學習而言，對於數量較少的資料其錯誤率及複雜性可被最小化[14]。

三、實驗結果

(一) FDR 結果

下圖四為 FDR 不同特徵之分布圖，其標籤 1 至 7 代表不同的七種情緒，圖四(a)為 FDR 分析後，將其最大的兩個值代表的特徵所畫的分布圖，圖四(b)則為最小兩個值的結果，可看出 FDR 值越大代表此特徵有較明顯的區分效果。



(a) (b)

圖 四、(a) FDR 最大兩值特徵分布圖，(b)FDR 最小兩值特徵分布圖

(二) GA 挑選結果

經 GA 挑選後，在男性方面共有 259 個特徵，其中 prosodic 與頻域特徵有 237 個，以 MFCC、formant 與 pitch 為主，而非線性特徵 Shannon entopy 之平均、中位數等共 5 個，curvature index 以百分位數為主的統計量共有 17 個。在女性方面，所得特徵共有 247 個，其中 prosodic 與頻域特徵共有 230 個，以 MFCC、formant 與 pitch 為主，非線性特徵 Shannon entopy 之切尾均值與百分位數共 5 個，curvature index 則以百分位數為主的統計量共 12 個。

(三) SVM 分類結果

比較不同性別使用傳統 prosodic 與頻譜特徵和加入非線性特徵後的混淆矩陣(confusion matrix)，下圖五為女性，使用傳統特徵準確率為 84.48%，加入非線性特徵後提升至 86.21%；圖六為男性，使用傳統特徵準確率為 84.44%，加入非線性特徵後提升至 88.89%。

Traditional features

Predict / Fact	Anger	Boredom	Disgust	Fear	Joy	Sadness	Neutral	Rate
Anger	13	0	0	0	0	0	0	100.00%
Boredom	0	7	0	0	0	0	2	77.78%
Disgust	0	0	7	0	0	0	0	100.00%
Fear	2	0	0	4	0	0	0	66.67%
Joy	1	0	0	2	5	0	0	62.50%
Sadness	0	0	0	0	0	7	0	100.00%
Neutral	0	2	0	0	0	0	6	75.00%
					Total recognition rate = 84.48%			

Traditional features + Nonlinear features

Predict / Fact	Anger	Boredom	Disgust	Fear	Joy	Sadness	Neutral	Rate
Anger	13	0	0	0	0	0	0	100.00%
Boredom	0	7	0	0	0	0	2	77.78%
Disgust	0	0	7	0	0	0	0	100.00%
Fear	2	0	0	4	0	0	0	66.67%
Joy	1	0	0	1	6	0	0	75.00%
Sadness	0	0	0	0	0	7	0	100.00%
Neutral	0	1	0	0	1	0	6	75.00%
					Total recognition rate = 86.21%			

圖 五、 女性分類結果
(上圖為傳統特徵，下圖為新增非線性特徵)

Traditional features

Predict / Fact	Anger	Boredom	Disgust	Fear	Joy	Sadness	Neutral	Rate
Anger	12	0	0	0	0	0	0	100.00%
Boredom	0	6	0	0	0	0	1	85.71%
Disgust	0	0	1	1	0	0	0	50.00%
Fear	0	0	0	6	1	0	0	85.71%
Joy	2	0	0	0	3	0	0	60.00%
Sadness	0	0	0	0	0	4	1	80.00%
Neutral	0	1	0	0	0	0	6	85.71%
					Total recognition rate = 84.44%			

Traditional features + Nonlinear features

Predict / Fact	Anger	Boredom	Disgust	Fear	Joy	Sadness	Neutral	Rate
Anger	12	0	0	0	0	0	0	100.00%
Boredom	0	6	0	0	0	0	1	85.71%
Disgust	0	0	2	0	0	0	0	100.00%
Fear	0	0	0	6	0	0	1	85.71%
Joy	2	0	0	0	3	0	0	60.00%
Sadness	0	0	0	0	0	4	1	80.00%
Neutral	0	0	0	0	0	0	7	100.00%
					Total recognition rate = 88.89%			

圖 六、男性分類結果
(上圖為傳統特徵，下圖為新增非線性特徵)

四、 結論

本研究以一般常用之語音特徵音高、共振峰、能量以及梅爾倒頻譜係數為基礎，加入了非線性特徵 Shannon entropy 和 curvature index，經由特徵擷取、特徵挑選到最後分類的方式建立語音情緒辨識模型。以柏林語音情緒資料庫做為分析對象，未加入非線性特徵量，所得男性及女性之情緒辨識率分別為 84.44%及 84.48%；加入非線性特徵量之後，男性辨識率提高至 88.89%，女性則提高至 86.21%。

針對各別情緒辨識改進的細部結果方面，可由Confusion matrix(圖五、圖六)得知，在加入非線性特徵量後，女性方面則因為誤判為害怕之開心情緒有部分被改正，使準確率由 62.5%提升為 75%；男性方面由於原本被誤判為無聊的中性情緒已判斷正確，使得中性準確率由 85.71%提升為 100%；而厭惡將原本誤判為害怕的情況改正，致使其準確率由 50%升至 100%。

另外，因目前所使用之資料為德文，對於不同語言及文化的在語音情緒影響的差異並未在研究中探討，因此有計畫建立中文語音情緒資料庫，藉以驗證本研究方法對於中文語音情緒辨識之可行性。

參考文獻

[1] 韩文静, et al. "语音情感识别研究进展综述." 软件学报 25.1 (2014): 37-50.

[2] Xie B. Research on key issues of Mandarin speech emotion recognition [Ph.D. Thesis]. Hangzhou: Zhejiang University, 2006 (in Chinese with English abstract).

[3] Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG. Emotion recognition in human-computer interaction. In: Proc. of the IEEE Signal Processing Magazine. 2001. 32−80. http://www.signalprocessingsociety.org/

[4]Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases."Pattern Recognition 44.3 (2011): 572-587.

[5] Siqing Wu, Tiago H. Falk, and Wai-Yip Chan. "Automatic speech emotion recognition using modulation spectral features." Speech communication 53.5 (2011): 768-785.

[6] Patricia Henríquez, et al. "Nonlinear dynamics characterization of emotional speech." Neurocomputing 132 (2014): 126-135.

[7] Ali Shahzadi, et al. "Speech emotion recognition using non-linear dynamics features." Turkish Journal of Electrical Engineering & Computer Sciences. doi10 (2013).

[8] Burkhardt, Felix, et al. "A database of German emotional speech." Interspeech. Vol. 5. 2005.

[9] Yen-Sheng Chen and Chien-Cheng Chang, 2012, “The Curvature Index and Synchronization of Dynamical Systems”, CHAOS 22, 023131.

[10] Suge Wang, et al. "A feature selection method based on fisher’s discriminant ratio for text sentiment classification." Web Information Systems and Mining. Springer Berlin Heidelberg, 2009. 88-97.

[11] Melanie Mitchell. An introduction to genetic algorithms. MIT press, 1996.

[12] Cheng-Lung Huang and Chieh-Jen Wang. "A GA-based feature selection and parameters optimizationfor support vector machines." Expert Systems with applications 31.2 (2006): 231-240.

[13] C-C Chang and C-J Lin. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[14]Christopher JC Burges. "A tutorial on support vector machines for pattern recognition." Data mining and knowledge discovery 2.2 (1998): 121-167