

在自適應性多重位元率語音編碼系統下之

語音端點檢測演算法研究

A Study on AMR Speech Codec System Based on Voice Activity Detection Algorithm

陳世豪

Shih-Hao Chen

空軍航空技術學院資訊電子組

義守大學資訊工程研究所博士生

Dept. of Information and Electronic, Air Force Institute of Technology

Graduated Student at Dept. of CSIE, I-Shou University

摘要

本論文提出一個基於小波(Wavelet)及支持向量機(Support Vector Machine, 簡稱 SVM)的新型強健性語音活動偵測(Voice Activity Detection, 簡稱 VAD)演算法, 該演算法能在噪訊環境中有效改善語音編碼系統的效能。文中使用小波去噪(Wavelets De-Noising)處理技術來消除語音訊背景中之雜訊源, 並藉由 SVM 之統計分析處理技術來訓練適應性多重位元率的寬頻編解碼器 (Adaptive Multi-Rate Wideband Codec, 簡稱 AMR-WB) 及窄頻編解碼器 AMR-NB (Adaptive Multi-Rate Narrowband Codec, 簡稱 AMR-NB) 之各頻帶音量特徵中的子頻帶分解(Sub-Band Decomposition)。

實驗音檔以奧羅拉(Aurora)語音資料庫為測試樣本, 多次的實驗結果顯示出, 使用經訓練過的 SVM, 在高雜訊背景下, 本文所提演算法的實驗結果精確度接近 94%, 可有助於提升語音編解碼系統效能方面之研究。

關鍵詞：支持向量機、語音活動偵測、小波去噪、適應性多重位元率、半徑式基底函數

Abstract

In this paper, a new voice activity detection (VAD) algorithm using wavelets and support vector machines (SVMs) is developed for improving the performance of speech coding systems working in noisy environment. The first step of the proposed algorithm is to apply wavelet-based de-noising method to reduce the background noise in the input speech. Then one can make use of SVM to train an adaptive multi-rate (AMR) involving the sub-band decomposition of various sub-band feature of level.

Various experimental results carried out from the Aurora speech database show that the proposed VAD algorithm could achieve 94% VAD accuracy under noisy environment. The proposed VAD algorithm can be applied to improve the performance of speech coding systems.

Keywords : SVM, VAD, Wavelets De-noising, AMR, RBF

一、前言

語音(Speech)是人類最古老的日常生活溝通方式,隨著科技日新月異的進步,語音的研究已成為科技中重要的應用範圍。目前,常見的語音之應用研究大致可分為四個方面,壓縮、合成、編碼及辨識等。就以語音編碼領域為例,語音訊息的傳輸、處理、存儲等各種資訊量的增加,大量資訊的壓縮處理已成為最主要的需求,隨著光纖通訊的來臨,無論是從減少傳輸頻帶資源,還是維持訊息傳輸的高效率等方面來看,研究可變速率語音編碼技術的系統有重要意義。VAD 是用於檢測語音信號中之有聲段與無聲段(Voice/Unvoice)程序,為了使頻寬作最有效的發揮,使其能夠充份控制語音實際訊息地傳遞,以有效的提昇並檢知訊號發送及接收的可靠度,因此,發展一套良好的 VAD 技術是有其必要性,而於實際環境之應用過程裡,尚必須考慮背景雜訊環境,當語音訊號於較大之背景雜訊音源處傳送、接收時,對於訊號實際內容之活動感測,勢必會影響音訊內容的判斷。

目前 VAD 技術已被應用於許多的語音通信系統上,例如,適應性多重位元率(AMR) 3G 行動電話系統,基於語音頻寬之考慮,AMR 可進一步區分為 AMR-NB (AMR Narrowband) 和 AMR-WB (AMR Wideband)。VAD 模組可降低可攜式設備[1]的同頻干擾 (Co-channel interference) 及功率消耗量(Power Consumption),並且能減少語音編碼時所需之計算量。此外,VAD 能夠有效率地使用頻寬以增加傳輸效率。

在許多的應用場合也可使用 VAD 技術

,如語音辨識技術(Speech Recognition) [2] 及語音增強技術(Speech Enhancement) [3]。在現今的手機、PDA 及數位型產品裝置裡也經常使用到語音方面的相關技術,而由於現實環境變化之多樣性如地下鐵(Subway)、嘈雜聲(Babble)、汽車(Car)、展覽會場 (Exhibition Hall)、餐廳(Restaurant)、火車站(Train Station)、機場(Airport)、街道(Street)等雜訊背景,語音辨識已經常地被應用於許多的日常生活中,其中 VAD 演算法即需具備良好的抗噪能力,而 Voice/Unvoice 臨界值的判斷將扮演重要的角色,以有效地分析出語音活動段與無語音活動段之變化,進一步使語音辨識技術能夠於高雜訊背景的雜訊中提升應用技術。

大部份傳統的 VAD 演算法都是藉由數個語音特徵參數與預設門檻值作比較來達成。如受測參數超過預設門檻值,則將制訂出有聲活動區域的規則。這些 VAD 演算法由於使用來自於在無噪訊條件下所導出的決策參數以及門檻值,因此,於噪訊環境下,這些參數和門檻值皆無法達到較好的效果。此外,因為大多數 VAD 演算法是依據固定點的門檻標準作為比較,而無法達到前後一致性的準確度,導致 Voice/Unvoice 判斷效果不佳,降低語音訊號處理的效能。

一個強健性的 VAD 應該避免背景噪訊的影響。為克服上述問題,本論文提出一種新的 VAD 演算法,透過使用小波和支持向量機,結合適應性多重位元率(AMR)來改善於噪訊環境下語音編碼系統的工作表現。在語音處理方面,本論文使用小波轉換(Wavelet Transform)技術應用於語音訊號之拆解,在傳統信號分析上,較常使用傅立

葉轉換(Fourier Transform)於頻域之分析，藉由正弦函數及餘弦函數作為基底，將整個訊號頻率變化分佈頻率軸上，但傅立葉轉換於頻域上無法提供時間訊息之變化量，而小波轉換具備縮張(scaling)與平移(translating)的函數特性，在信號分析上能呈現出在高頻率下之較短的持續時間以及在低頻率下較長的持續時間，因此，在時域及頻域上皆能獲得較好的解析度之表現。

在分類方法(Classification)之判別分析上，傳統較常使用之人工智慧的機器學習演算法有 k-最近鄰近決定法則 K-NN (K-th Nearest Neighbor)、Boosting、隱藏式馬可夫模型(Hidden Markov Model, 簡稱 HMM)、貝式網路(Bayesian Network)、高斯混合模型(Gaussian Mixture Model, 簡稱 GMM)、類神經網路(Artificial Neural Network, 簡稱 ANN)等等，其中以 ANN 而言，小樣本的分類問題上，ANN 容易發生過度學習(Overfitting)及陷入局部極小值(Local Minimum)的問題，而 SVM 是利用根據結構風險最小化(Minimize Structural Risk, 簡稱 SRM)原理，在泛化能力上有較好的表現，適合用於多類別的分類問題(Multi-Class Problems)上。

在早期的 VAD 演算法中[4]，語音偵測所使用的參數是依據自相關係數(Autocorrelation Coefficients)、線性預估編碼距離度量(LPC Distance Measure)、短時間能階(Short-time Energy Levels)、越零率(Zero Crossing Rate)及音高週期(Pitch Period)等等。而濾波器組(Filter Bank)和複雜信號分析(Complex Signal Analysis) [1] 是 VAD 近期的一些發展技術。對於受到雜訊破壞下的語音信號，小波轉換一直被視為是一種強而有力的非線性去噪工具。

本文在運用基於小波去噪方法[5]，藉由 Stein's unbiased risk estimate (SURE) [6]減低語音背景噪訊，而從中擷取出參數以完成 VAD。本文在 AMR-WB 編碼器下使用 1 至 12 個子頻帶分解(Sub-band

Decomposition)，在 AMR-NB 編碼器下使用 1 至 9 個子頻帶分解以作為 VAD 的語音特徵參數。此外，支持向量機(SVM)是由 Vapnik [7]所提出，在許多的方面一直被視為是一種新的學習演算法，例如語音分類(Audio Classification) [8]的應用。

由於 VAD 將有聲段及無聲段聲音視為一組類別，所以 SVM 能成為解決 VAD 問題的一種方法。本文選用半徑式基底函數(Radial Basis Function)作為 SVM 的核心決策。藉由上述音訊下語音特徵參數的萃取，透過訓練過後的 SVM，於噪訊環境下仍可獲得準確 VAD 結果。本文使用受實際噪音破壞的音源，實驗結果精確度能達不錯效果。因此，藉由本論文實驗證實，使用 Wavelets 及 SVM 的結合，可以提升 VAD 在克服環境方面之強健性(Robust)。

二、研究方法

2.1 小波去噪原理

近幾年，小波在音訊的研究領域裡，已有相當多之發展及貢獻，其中 Donoho 和 Johnstone [5]所提出以小波為基礎的去噪方式為常見的方法之一。此方法根據受雜訊破壞之音源下的小波係數作為篩選之後，能夠很輕易地應用於語音去噪處理。研究顯示[9]，小波臨界值篩選方法比傳統的語音去噪處理能有較好的表現。假設

$$x(n) = s(n) + \sigma g(n) \quad (1)$$

為在噪訊下的語音信號，其中 $s(n)$ 代表乾淨的音源、 $g(n)$ 表示零均值、變異數為 1、 σ 為噪音音壓級常數的一個高斯隨機變數。

從噪訊語音信號 $x(n)$ 中，找尋近似乾淨的發聲音源信號 $s(n)$ 中的估計量 $\hat{s} = \{\hat{s}(1), \hat{s}(2), \dots, \hat{s}(N)\}$ ，使具有如下定義之最小化均方差值：

$$\|s - \hat{s}\|^2 \triangleq \frac{1}{N} \sum_{n=1}^N |\hat{s}(n) - s(n)|^2 \quad (2)$$

以小波為基礎的語音去噪(圖 1)演算法可歸納成下列三步驟：

- 拆解含噪的語音信號成小波係數。
- 針對(a)式中的小波係數作臨界值篩選。
- 合成(b)式中篩選後的小波係數，以達到增強語音信號。

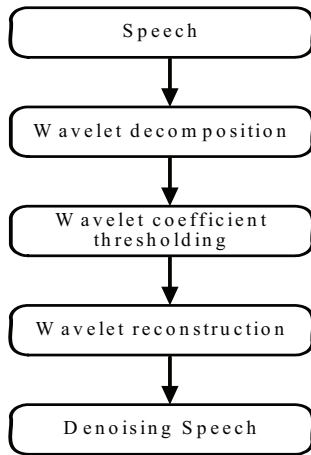


圖 1 以小波為基礎的語音去噪流程圖

2.2 濾波器組架構

本文藉由濾波器組架構來完成小波轉換。快速離散演算法由 Mallat [10]所提出。假設 $h(n)$ 和 $g(n)$ 是分析低通及高通的濾波器組。符號 $\downarrow 2$ (圖 2)表示濾波器的輸出減半。同時令 $\{a_3(n)\}_{n \in \mathbb{Z}}$ 是分析濾波器組的輸入，則分析濾波器組的輸出由下列式子可得

$$a_i(k) = \sum_n h(n - 2k) a_{i+1}(n) \quad (3)$$

$$d_i(k) = \sum_n g(n - 2k) a_{i+1}(n) \quad (4)$$

$a_i(k)$ 、 $d_i(k)$ 為分別對於 $a_{i+1}(n)$ 進行低頻及高頻小波之係數。

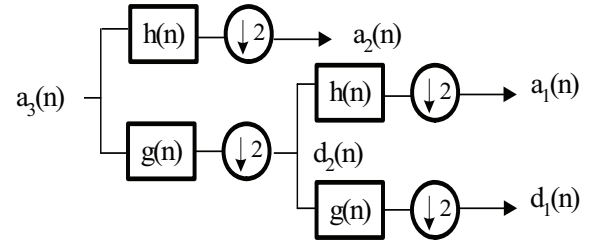


圖 2 二階小波拆解信號流程圖

2.3 支持向量機技術

支持向量機(SVM: Support Vector Machines)是由 Vapnik[11]和他的合作者在對統計學習理論 (Statistical learning theory, 簡稱 SLT) 多年的研究基礎上所共同提出的一套機器學習演算法，它使用統計學習理論中風險最小化及分類邊界最大化作為主要原則，同時並以結構風險最小化 (Minimize Structural Risk, 簡稱 SRM) 為目標。因此，支持向量機的訓練過程等價於求解一個線性且具限制式的二式規劃問題 (Quadratic Programming, 簡稱 QP)，以避免產生過度學習 (Over-fitting) 的問題。

在一般常見的統計分類方法中，大多數方法需要有足夠的音訊樣本數作為依據，且必須擁有明顯的音訊特徵作區隔，因而在結果的分析上才會有較好的表現。SVM 最大的優勢，即能在有限的音訊樣本條件下，產生較好的辨識效果，目前 SVM 廣泛地被應用於如影像辨識[12]、文字分類[13]、生物科技[14]等技術領域，在實驗效果上，都有不錯之結果。

文中 Pd (Probability of voice detection, 簡稱 Pd)、Nd (Probability of non-voice detection, 簡稱 Nd) 分別代表語音有聲段偵測機率及語音無聲段偵測機率。類別問題透過上述技術方法作成標準分類，以解決非線性、高維度識別及小樣本等實際問題，以提昇預測之確準性。因此，在有限訓練樣本下所得到的決策規則對於獨立的測試集能有較好的表現，大大地提升監督式學習的應用分析能力。

圖 3 為 SVM 之訓練向量 X_i 透過函數 ϕ

映射到較高維度之空間，接者，SVM 在此較高維度之空間下找到一個使邊界最大化的線性可分超平面(Linear Separating Hyper-plane) (圖 4)。

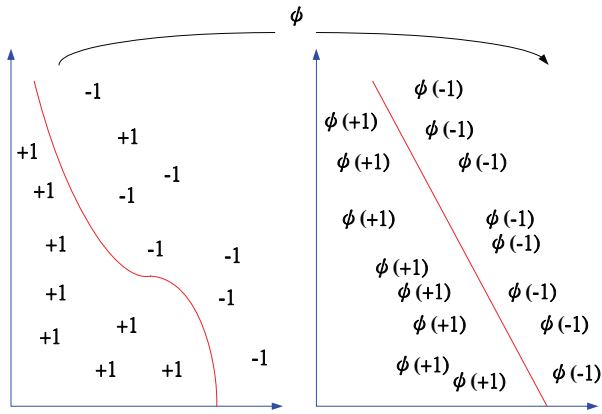


圖 3 簡化分類任務中之特徵映射

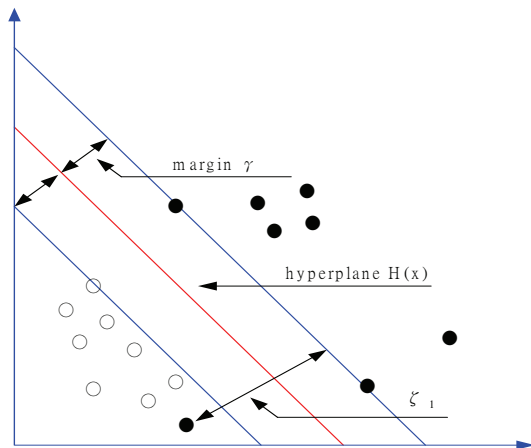


圖 4 分類問題中之鬆弛變數及邊界

2.4 系統設計架構圖

本文系統設計架構圖分三階段進行:第一階段進行語音端點偵測(圖 5)以作為進行語音端點偵測之校正。第二階段本文使用以多人為背景之語音除噪程序(圖 6)。第三階段使用 RBF 作為 SVM 核心參數，以產生 SVM 預測(Predict)分析結果(圖 7)。

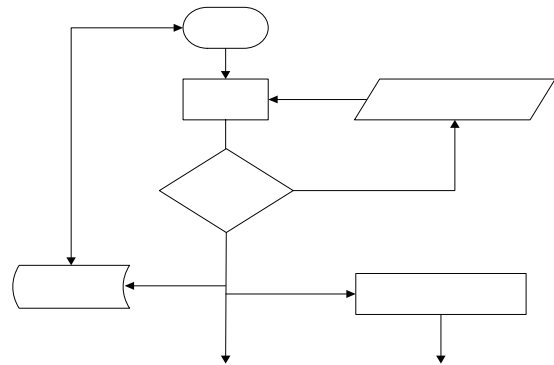


圖 5 第一階段 VAD 調整設計架構

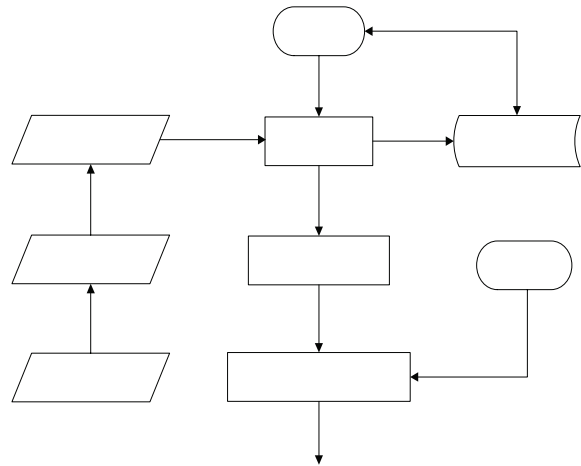


圖 6 第二階段系統設計架構圖

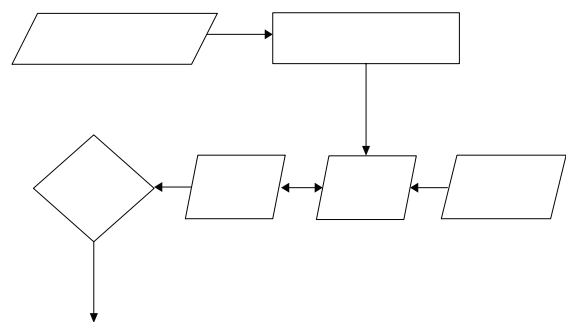


圖 7 第三階段系統設計架構圖

2.5 實驗結果

2.5.1 分析 AMR-NB 的聲訊環境

AMR-NB (Adaptive Multi-Rate Narrowband Codec) 有效頻率在 0 至 4000 Hz(圖 8)[15]分成 1 至 9 個頻帶(表 1)[15]

分佈範圍。對於一般語音而言，取樣頻率為 8kHz、每個取樣以 16 位元表示，音框大小為 160 取樣點的 AMR-NB 語音通道帶寬限制為 3.7M Hz。

AMR-NB 是根據實驗通信線路狀況來改變語音編碼器中的編碼速度，使用編碼速度如表 2[16]範圍內，AMR-NB 可從中挑選出最適合的編碼速度。

由圖 9 所示，分析在 0~20 dB 訊號雜訊比值下 VAD 之表現能力。配合乘法臨界值調整 sln 效果下，臨界值方法 Soft Threshold 表現優於臨界值方法 Hard Threshold(圖 10)。

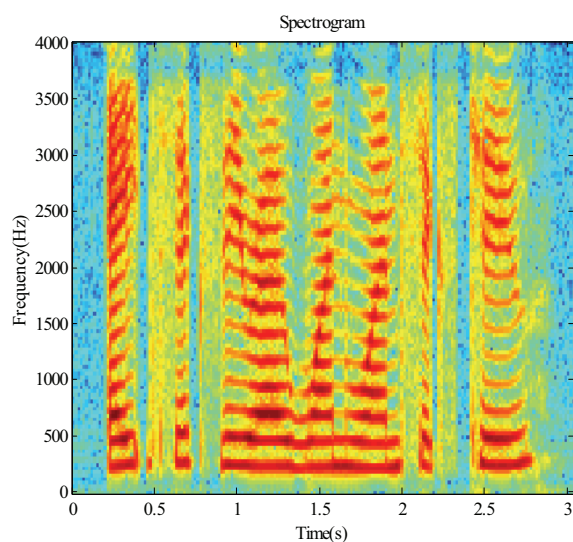


圖 8 音檔 FAK_86Z1162A 有效頻率
0~4000Hz

表 1 AMR-NB 編碼器 9 種頻帶範圍

頻帶數	頻率寬度	頻帶數	頻率寬度
1	0 – 250 Hz	6	1500 – 2000 Hz
2	250 – 500 Hz	7	2000 – 2500 Hz
3	500 – 750 Hz	8	2500 – 3000 Hz
4	750 – 1000 Hz	9	3000 – 4000 Hz
5	1000 – 1500 Hz		

表 2 AMR-NB 編碼位元率

編碼模式	位元率	編碼模式	位元率
AMR_12.20	12.20 kbit/s	AMR_5.90	5.90 kbit/s
AMR_10.20	10.20 kbit/s	AMR_5.15	5.15 kbit/s
AMR_7.95	7.95 kbit/s	AMR_4.75	4.75 kbit/s
AMR_7.40	7.40 kbit/s	AMR_SID	1.80 kbit/s
AMR_6.70	6.70 kbit/s		

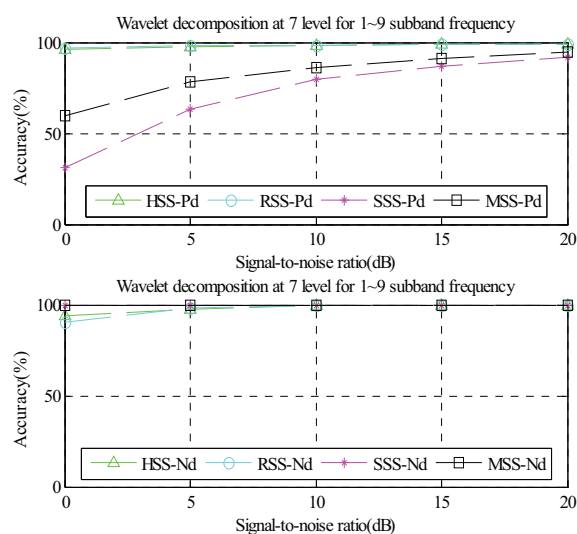


圖 3 小波去除噪訊後(sln、Soft Threshold)

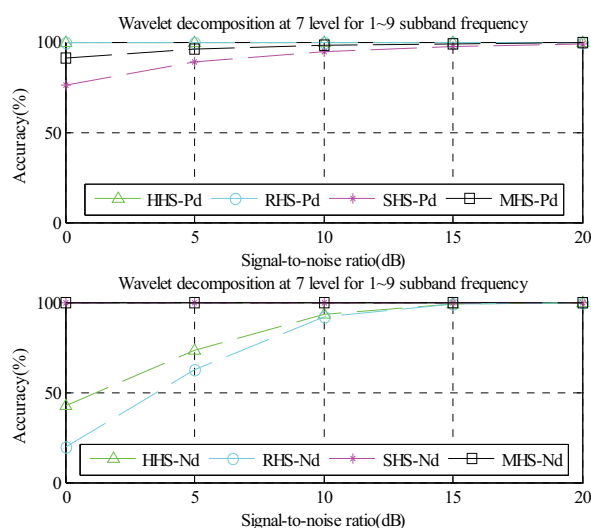


圖 4 小波去除噪訊後(sln、Hard Threshold)

2.5.2 分析 AMR-WB 的聲訊環境

AMR-WB (Adaptive Multi-Rate Wideband Codec) 有效頻率在 0 至 8000 Hz (圖 11)[17]分成 1 至 12 個頻帶如表頻帶(表 3)[17]分佈範圍。所使用之取樣頻率為 16kHz、每個取樣以 16 位元表示，音框大小為 320 取樣點，AMR-WB 語音通道帶寬為 7M Hz，AMR-WB 是根據實際通信線路狀況來改變語音編解碼器中的編碼速度，使用編碼速度如表 4[18]範圍內。AMR-WB 可從中挑選出最適合的編碼速度。

由圖 12 所示，分析在 0~20 dB 訊號雜

訊比值下 VAD 之表現能力。配合在乘法臨界值調整 sln 效果下，臨界值方法 Soft Threshold 表現優於臨界值方法 Hard Threshold(圖 13)。

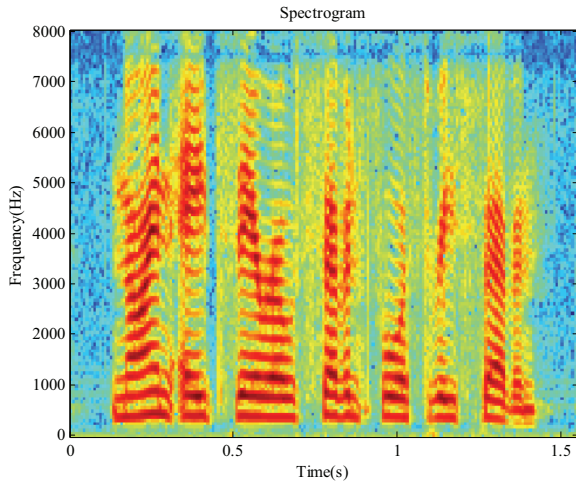


圖5 音檔 FBA_98Z7437A 有效頻率 0~8000Hz

表 3 AMR-WB 編碼器 12 種頻帶範圍

頻帶數	頻率寬度	頻帶數	頻率寬度
1	0 – 200 Hz	7	1600 – 2000 Hz
2	200 – 400 Hz	8	2000 – 2400 Hz
3	400 – 600 Hz	9	2400 – 3200 Hz
4	600 – 800 Hz	10	3200 – 4000 Hz
5	800 – 1200 Hz	11	4000 – 4800 Hz
6	1200 – 1600 Hz	12	4800 – 6400 Hz

表 4 AMR-WB 編碼位元率

編碼模式	位元率	編碼模式	位元率
AMR-WB_23.85	23.85 kbit/s	AMR-WB_14.25	14.25 kbit/s
AMR-WB_23.05	23.05 kbit/s	AMR-WB_12.65	12.65 kbit/s
AMR-WB_19.85	19.85 kbit/s	AMR-WB_8.85	8.85 kbit/s
AMR-WB_18.25	18.25 kbit/s	AMR-WB_6.60	6.60 kbit/s
AMR-WB_15.85	15.85 kbit/s	AMR-WB_SID	1.75 kbit/s

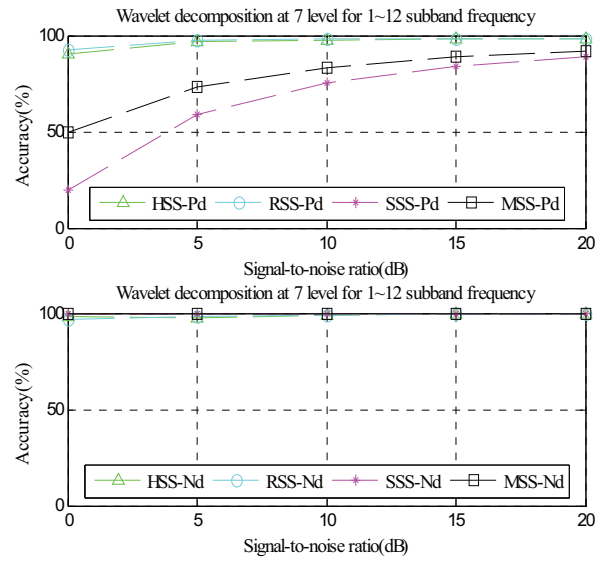


圖 6 小波去除噪訊後(sln、Soft Threshold)

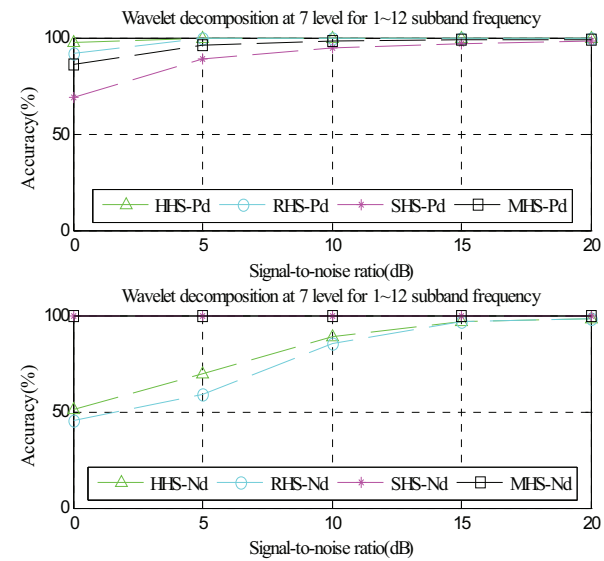


圖 7 小波去除噪訊後(sln、Hard Threshold)

圖文縮寫註解：

Using Wavelets De-noising option

RSS : rigrsure、soft and sln.

RHS : rigrsure、hard and sln.

HSS : heursure、soft and sln.

HHS : heursure、hard and sln.

SSS : sqtwolog、soft and sln.

SHS : sqtwolog、hard and sln.

MSS : minimaxi、soft and sln.

MHS : minimaxi、hard and sln.

三、結論

本文提出了一種新的 VAD 技術，該技術結合了小波去除雜訊效果和支持向量機預測分析能力。小波去噪的方法是將語音訊號中之雜訊給予濾除，使原始之語音信號能解析出來，於實驗過程中，比較各種小波參數除噪後之表現水準，本文提出解析子頻帶分解語音參數的配合，其表現能力高頻帶優於低頻帶語音參數的表現，並透過支持向量機之統計分析技術，找尋出原始音訊之語音活動區域，以致進一步作語音活動區域之分析。並於執行支持向量機時，有效地分析語音特徵的資料分佈特性，將有助於提升 SVM 的分類估測效果，並能將語音訊號於高雜訊干擾下，給予作正確地語音活動分類及預測分析。本文所提演算法的實驗結果精確度接近 94%，除可提升語音編解碼系統效能方面之研究外，未來進一步使語音辨識技術能夠於高雜訊背景的環境中提升應用技術。

四、參考文獻

1. ETSI EN 301 708 V7.1.1, Voice Activity Detector (VAD) for Adaptive Multi-Rate, 1999.
2. Javier Ramírez, José C. Segura, Carmen Benítez, Ángel de la Torre, and Antonio J. Rubio, "A New Kullback - Leibler VAD for Speech Recognition in Noise", IEEE Signal Processing letters, Vol. 11, no. 2, pp. 266-269, Feb. 2004.
3. N. R. Garner, P. A. Barrett, D. M. Howard, and A. M. Tyrrell, "Robust noise detection for speech detection and enhancement," Electron. Lett., vol. 33, no. 4, pp. 270 - 271, Feb. 1997.
4. ITU-T Rec. G.729 Annex B, A silence compression scheme for G.729 optimized for terminals conforming to ITU-T V.70. 1996.
5. D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," Journal of the American Statistical Association, 90:1200 - 1224, 1995.
6. C. Stein. Estimation of the mean of a multivariate normal distribution. Annals of Statistics, 9(6):1135 - 1151, 1981
7. V. N. Vapnik, Statistical Learning Theory, Wiley, 1998
8. Chien-Chang Lin, Shi-Huang Chen, T. K. Truong, and Yukon Chang, "Audio Classification and Categorization Based on Wavelets and Support Vector Machine," IEEE Trans. on Speech and Audio Processing, Vol. 13, No. 5, pp. 644-651, Sept. 2005.
9. Yi Hu and P.C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," IEEE Trans. on Speech and Audio Processing, Vol. 12, No. 1, pp. 59-67, Jan. 2004.
10. S. Mallat, "Multifrequency channel decomposition of images and wavelet model," IEEE Trans. Acoustic, Speech and Signal Processing, vol. 68, pp. 2091-2110, 1980.
11. Vapnik, V.N., An overview of statistical learning theory. 1999.
12. Pontil, M. and Verri, A., "Object recognition with support vector machines", IEEE Trans. On PAMI, 20, pp.637-646, 1998.
13. Joachims, T., "Text categorization with support vector machines", In Proceedings of European Conference on Machine Learning (ECML), 1998.
14. Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M., and Haussler, D. "Knowledge-base analysis of microarray gene expression data using

- support vector machines” , Technical report, University of California in Santa Cruz, 1999.
- 15.3G TS 26.094 : “AMR Speech Codec; Voice Activity Detection (VAD)” , Oct. 1999.
- 16.3GPP TS 26.071: 'AMR speech Codec; General description'[S], Aug. 1999.
- 17.3G TS 26.194: "AMR wideband speech codec; Voice Activity Detection (VAD)" , Mar. 2001.
- 18.3GPP TS 26.171: “AMR wideband speech codec; General description” , Mar. 2001.