

用於 G.729E 語音編碼器之修正式全域脈衝搜尋法

陳福坤
南台科技大學
資訊工程系/所
副教授

fkchen@mail.stut.edu.tw

陳冠銘
南台科技大學
資訊工程系/所
碩士生

joec913137@gmail.com

摘要⁺

在目前語音壓縮技術領域中，以代數碼激式線性預測編碼(Algebraic code excited linear prediction, ACELP)為主要的編碼方式。為了提升語音品質，ITU-T 制定了品質高於 G.729 的 G.729E。本論文以 ITU-T G.729E 為實驗平台，針對單軌多脈衝之隨機碼簿，研究並修正全域脈衝替換之搜尋流程以降低搜尋計算複雜度。所提出的修正式全域脈衝搜尋法並具有複雜度可階化(Complexity scalability)的架構，可隨時依照運算核心的負載程度，動態設定之。實驗顯示，所提出之修正法，計算複雜度僅為原始 ITU-T G.729E 之 10.5% 至 30.3%，且不影响編碼語音的聽覺感官品質。

關鍵詞：語音編碼、ACELP、隨機碼簿、固定碼簿。

Abstract

The algebraic code excited linear prediction (ACELP) algorithm due to low complexity and high quality in its analysis-by-synthesis optimization has been adopted by many speech coding standards. To improve the quality of the coded speech, ITU-T G.729E is recommended rather than G.729. This paper proposed the modified global-pulse-replacement (GPR) search algorithm for stochastic codebook of G.729E speech coder which is with multi-pulses in a track. The proposed approach is with the structure of complexity scalability. Experimental results show that the proposed approach is able to reduce the computational complexity about 89.5%-69.7% as well as preserves the speech quality.

Keywords: Speech coding, ACELP, Stochastic codebook, Fixed codebook.

1. 前言

近年來，科技應用於網路研究方面正持續發展，數據網路通信也融入傳統的電話業務領域。相對的，語音編碼技術在多媒體壓縮領域中一直是各界研究的重點之一，隨著 VOIP 網路電話[5]的發展迅速，在網路頻寬有限的情況下，低位元率的編碼一直是我們追求的目標，為了兼具語音品質和低位元率，分析再合成(Analysis-by-synthesis, AbS)[1]的技術是語音編碼中最佳的解決途徑。其中以碼激式線性預測(Code excited linear prediction, CELP)為主要的設計基礎。

以 CELP 為架構的語音編碼器，其編碼端需要有：短時距分析(Short-term prediction, STP)、長時距分析(Long-term prediction, LTP)、激發訊號編碼等主要分析程序。其中短時距分析是以線性預測分析(LPC analysis)來做口腔模式(Vocal tract modeling)的分析，用來取得語音的特徵參數；而長時距分析此部分在語音編碼應用上稱之為基週預測(Pitch estimation)，主要為取得語音音框中具有週期性的特性；最後在激發訊號編碼部份，可比擬為喉嚨的激發訊號，理論上做到此階段，訊號已無法再預測，因此這些激發訊號就須直接將之編碼，最好能夠將這些激發訊號的位置和振幅大小完整編碼儲存或傳送，使得解碼端可以利用來合成原始語音訊號。碼激式線性預測模式編碼對合成語音品質的主要關鍵點在激發訊號編碼，所需編碼的資訊包括激發訊號脈衝的振幅和位置，而碼向量的尋找就要相當精準，因此碼向量尋找的準則就改以閉迴路搜尋(Close-loop search)方式為之。閉迴路搜尋方式為將候選的碼向量經由現行分析出的口腔模式合成濾波器合成出語音，再看合成的語音和原始語音間的誤差，希望其均方誤差愈小愈好。但基於聽覺感官效應，在做合成的語音和原始語音間的均方誤差時，會加上所謂感官加權(Perceptual weighting)，此時的均方誤差稱之為感官加權誤差(Perceptual weighted error)。

⁺ 本研究感謝國科會計畫編號 NSC 97-2221-E-218-022 經費支持。

在語音壓縮技術領域中，擁有專利的代數碼激式(Algebraic code excited LPC, ACELP)編碼架構可以獲得較佳的編碼語音品質，現在已經是 ITU [8-10]和 MPEG-4 [7]語音編碼器的核心架構了。ITU-T G.729 [8]可以說是 ACELP 編碼架構中最具代表性、規劃最完整者，而後 ITU 又針對各種不同的編碼需求，進而再對 G.729 延伸出其它標準，為了降低 G.729 複雜度，提出了 G.729A [9]，為了再降低位元率，規劃了 G.729D，為了提升音質，又規劃了 G.729E[10]，為增加通道使用率 G.729B 定義了靜音壓縮演算法，可使位元率能隨語音資料動態變化。在 ACELP 的架構中脈衝數量的多寡則直接的會影響到語音品質，脈衝數越多的碼簿結構，合成出來的語音訊號品質則較佳，反之脈衝數越少，合成語音品質相對較低。ITU-T G.729E (11.8 kbps)之隨機碼簿的脈衝位置便使用了單軌多脈衝的設計，以提升編碼語音品質為主。G.729E 特別增加了音樂信號及帶背景噪聲的語音信號的處理，使其更適用於電話會議、VOIP 等多媒體通信，唯其編碼位元率為 11.8kbit/s，略高於 G.729。對於 G.729E 語音編碼器中隨機碼簿搜尋而言，為了避免計算複雜度太大，因此使用深度優先樹狀搜尋(Depth first tree search)的方式。我們以全域脈衝替換法[3]為出發點深入研究並創新方法，使計算複雜度更為簡化。

本論文主要針對 ACELP 單軌多脈衝之隨機碼簿進行搜尋演算法設計，並實際使用浮點化版本之 G.729E 為實驗平台做性能分析。因此，論文第二節將先簡介 ITU-T G.729E 之隨機碼簿編碼程序。為了在一些較低效能運算處理器的行動裝置上能夠實現，論文第三節則討論了全域脈衝搜尋法(Global Pulse Replacement)，並提出了修正原則。論文第四節會對所提出之方法進行實驗數據，進行語音品質和複雜度比較。最後，第五節對所提的搜尋法做一個結論。

表 1 G.729E 隨機(固定)碼簿結構

Pulse	Sign	Positions
i_0, i_5	+1/-1	T ₀ : 0, 5, 10, 15, 20, 25, 30, 35
i_1, i_6	+1/-1	T ₁ : 1, 6, 11, 16, 21, 26, 31, 36
i_2, i_7	+1/-1	T ₂ : 2, 7, 12, 17, 22, 27, 32, 37
i_3, i_8	+1/-1	T ₃ : 3, 8, 13, 18, 23, 28, 33, 38
i_4, i_9	+1/-1	T ₄ : 4, 9, 14, 19, 24, 29, 34, 39

2.代數碼激式(ACELP)編碼

G.729E 是為了提升語音品質而設計，位元率為 11.8 k bit/s，其固定碼簿具有單軌多脈衝的特性，並使用了深度優先樹狀搜尋(Depth first tree search)的方式。其隨機(固定)碼簿結構如表 1 所示，包含了 10 根非零的脈衝 ($i_k, k=0,1,\dots,9$)，每軌配置二根脈衝，分別為五軌，且所有脈衝的振幅皆以+1 或-1 來表示。

在 ACELP 架構中，隨機碼簿的搜尋依據是讓加權原始語音和加權合成語音之間有著最小的均方誤差(Minimum Mean Square Error, MMSE)，定義如下：

$$\epsilon_k = \|\mathbf{x} - g\mathbf{H}\mathbf{c}_k\|^2 \quad (1)$$

其中， \mathbf{c}_k 為定義脈衝位置的碼向量， k 為其碼向量索引， \mathbf{x} 為目標訊號向量， g 為增益值， \mathbf{H} 為口腔模式合成濾波器的脈衝響應 $h(n)$ 組成之下三角 Toeplitz 迴旋積分矩陣。由於最小化(1)式的準則，需同時進行增益值與脈衝位置搜尋，因此可改進行合成向量與目標向量相似度大小比對，再藉以求出增益值大小；即搜尋準則可改為求出最大的正規化相似度：

$$Q_k = \frac{(\mathbf{H}'\mathbf{x}\mathbf{c}_k)^2}{\mathbf{c}_k'\mathbf{H}'\mathbf{H}\mathbf{c}_k} = \frac{(\mathbf{d}'\mathbf{c}_k)^2}{\mathbf{c}_k'\mathbf{\Phi}\mathbf{c}_k} \quad (2)$$

其中， $\mathbf{d} = \mathbf{H}'\mathbf{x}$ 為固定碼簿目標訊號與脈衝響應交相關係數，其表示式如下：

$$d(n) = \sum_{i=n}^{39} r(i)h(i-n) \quad n=0,\dots,39 \quad (3)$$

$\mathbf{\Phi} = \mathbf{H}'\mathbf{H}$ 為合成濾波器的脈衝響應 $h(n)$ 的自相關函數(Auto correlation)，其向量元素表示式如下：

$$\phi(i, j) = \sum_{n=j}^{39} h(n-i)h(n-j), \quad (j \geq i) \quad (4)$$

3.隨機碼簿搜尋法

原始標準的 G.729E 固定碼簿搜尋方式是使用深度優先樹狀搜尋(Depth first tree search)的方法，選出最能夠合成目標訊號的一組位置向量。此搜尋法首先求出每一軌 $|b(n)|$ 最大之脈衝位置，從中把 $|b(n)|$ 的全域最大值對應到的脈衝位置設定為 i_0 ，因此脈衝與位置的關係並不會固定不變按照表 1 的碼簿結構，接著步驟二，將 i_0 下一軌 $|b(n)|$ 最大值對應之脈衝位置設

定為 i_1 ，而剩下的脈衝對 $\{i_2, i_3\}$ 、 $\{i_4, i_5\}$ 、 $\{i_6, i_7\}$ 和 $\{i_8, i_9\}$ 則分別做巢狀搜尋，因為每一軌有 8 個位置，因此一個脈衝對會有 8×8 個組合，剩下的總共有 4 個脈衝對就有 $4 \times (8 \times 8) = 256$ 個迴圈數，當上述步驟完成後，將 i_0 所在位置的下一軌 $|b(n)|$ 最大值脈衝位置重新設定為 i_0 。例：假設找到 $|b(n)|$ 最大所在位置為 T_2 中的 27，則 i_0 設成 27， i_1 則為 i_0 所在軌的下一軌 T_3 的 $|b(n)|$ 最大所在位置，接下來 $\{i_2, i_3\}$ 則是 T_4 與 T_0 做組合， $\{i_4, i_5\}$ 為 T_1 與 T_2 做組合， $\{i_6, i_7\}$ 為 T_3 與 T_4 做組合， $\{i_8, i_9\}$ 為 T_0 與 T_1 做組合。接著 i_0 再設成 T_3 中 $|b(n)|$ 最大所在位置，重複以上動作以此類推，一直做到 i_0 連續三軌都輪替完。因為 i_0 設定位置會做三次改變，所以脈衝位置組合數共有 $3 \times 4 \times 8 \times 8 = 768$ 種組合，搜尋的計算量還是相當龐大。根據固定碼簿搜尋的定義，碼簿搜尋依據是在求出(2)式為最大的脈衝組合，由此可知可令(2)式中分子部份求得最大的交相關係數 $d(n)$ 來考量與脈衝出現的相對位置。本節我們針對 G.729E 隨機碼簿搜尋設計此方法，都是以脈衝換替法[6]為基本架構，以其能找出與原搜尋方法之語音品質和計算複雜度的差異，以下針對全域脈衝換替法與提出的方法做一詳細介紹。

3.1 全域脈衝換替搜尋法

以全域脈衝換替法來說，在進入固定碼簿搜尋脈衝位置之前，會先計算一個預測函數 $b(n)$ ，來預測脈衝最可能出現的位置。進而將初始向量(位置)定義在每一軌的 $|b(n)|$ 最大值所在的脈衝位置上。其預測函數 $b(n)$ 如下：

$$b(n) = \frac{res_{LTP}(n)}{\sqrt{\sum_{i=0}^{39} res_{LTP}(i)res_{LTP}(i)}} + \frac{d(n)}{\sqrt{\sum_{i=0}^{39} d(i)d(i)}} \quad (5)$$

$n = 0, \dots, 39$

其中 $res_{LTP}(n)$ 為長時距預估殘值訊號 (Long-term prediction residual)，是由線性預測殘值訊號(LP residual)扣掉乘上基週增益的適應性碼向量計算而得。

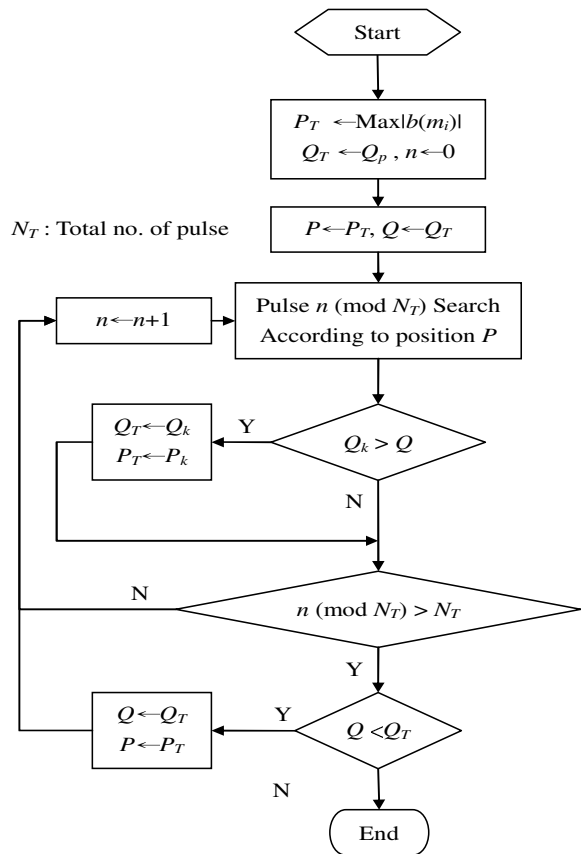


圖 1、全域脈衝替換法

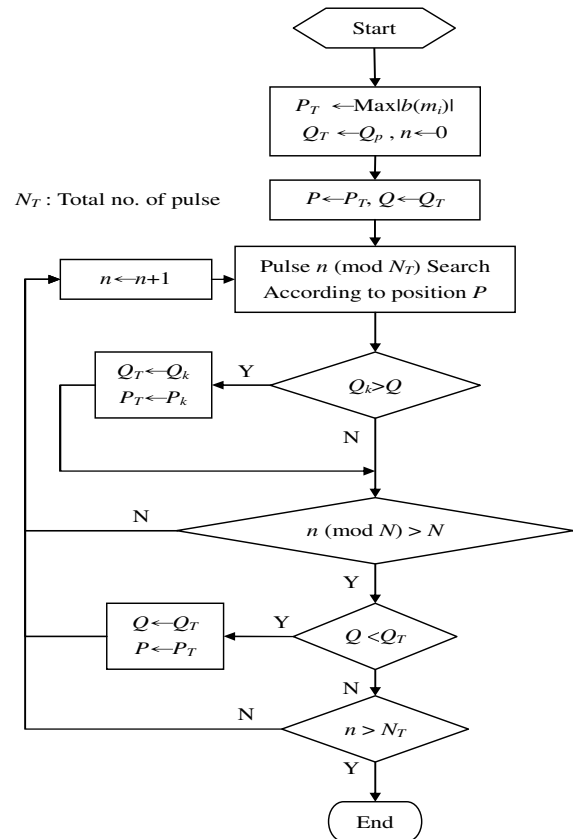


圖 2、修正式全域脈衝替換法

使用初始向量(位置)對每一軌的每一根脈衝位置進行替換,使得(2)式之正規化相似度 Q_k 值為最大。以 G.729E 為例,全域脈衝替換搜尋法如圖 1 所示,先求出初始位置向量 P 的 Q 值,接著對 10 根脈衝之每一脈衝位置都進行替換搜尋(位置向量 P_k),求出每一脈衝位置替換後之 Q_k 值,在找出最大者,以其位置向量 P_k 置換目前的初始位置向量 P 並紀錄目前最佳 Q 值為 Q_k 值。全域脈衝替換搜尋法之搜尋終止條件為:當替換脈衝之後 Q_k 值已無上升時,則將最後所替換之最佳脈衝組合為輸出碼向量。

全域脈衝替換法因為並無事先限制替換次數,因此無法判別替換程序會進行幾次,需由實際搜尋迴圈數記錄來決定。文獻上實際實驗結果均顯示其能夠有效的減少搜尋脈衝的組合數,以達到降低計算複雜度。用於單軌多脈衝隨機碼簿架構的 G.729E 時,也確實比原 G.729E 搜尋方法計算複雜度還要低。

3.2 修正式全域脈衝搜尋法

全域脈衝替換法之最佳位置更換法則為:待 G.729E 之 10 根脈衝位置均搜尋後,再進行最佳位置更換決定,也因此才稱作「全域脈衝替換法」。本論文提出「修正式全域脈衝搜尋法」乃增加設定一參數 N ,用以決定每 N 根脈衝位置均搜尋後,即進行最佳位置更換,亦即不待 10 根脈衝位置均搜尋後,再進行最佳位置更換決定。因此,所提出的修正式全域脈衝搜尋法並具有複雜度可階化(Complexity scalability)的架構。參數 N 值的設定,可隨時依照運算核心的負載程度,動態設定之。

我們亦可發現:「全域脈衝替換法」即為「修正式全域脈衝搜尋法」之 $N=10$ 之特例。但是,因為修正式全域脈衝搜尋法是「輪流」進行脈衝替換,並依照設定的 N 值,每 N 次脈衝替換再進行最佳位置更換程序。可以預見的是:全域脈衝替換法的搜尋迴圈數約等於修正式全域脈衝搜尋法之 $N=9$ 的情況;全域脈衝替換法的語音品質應等於修正式全域脈衝搜尋法之 $N=10$ 的情況。

N 值的大小左右了編碼後的語音品質;較大的 N 值會使搜尋範圍變大,相對的計算複雜度也提高一些。較小的 N 值會使搜尋範圍變小,卻有機會讓搜尋及早跳脫區域最大值區。需注意的是:當「修正式全域脈衝搜尋法」之 N 值設定較小時,搜尋終止條件若與原始「全

域脈衝替換法」相同,則極易發生立即終止搜尋的情況。因此所提出的修正式全域脈衝搜尋法搜尋終止條件,需再加設一基本限制;即需在所有脈衝都替換過的情況下,才能進行與原始全域脈衝替換法相同搜尋終止條件判斷。

4. 實驗結果

以 ITU-T G.729E 語音編碼器為實驗平台下,本節將對各項編碼後語音品質進行標準化評估。首先,我們進行了平均意見分數(Mean Opinion Score, 簡稱 MOS)的主觀評量。因為使用語音編碼器的即為一般電話使用者,所以以未受專業訓練的受測者來進行實驗亦屬合理。由於此種非官方 MOS 主觀評分並不具意義,因此本實驗改以預先告知受測者 G.729E 語音編碼品質 MOS 分數為 4,再由受測者評出 MOS 衰退值。所實驗用的語音分別有國語、英語;同時由男女發聲。如表 2 所示,測試者均表示無法判定 MOS 衰退值。這顯示出對於人類耳朵主觀的聽覺來說,絲毫感覺不出聲音品質的下降。

為了分辨出修正式全域脈衝搜尋法、全域脈衝搜尋法與原始深度優先樹狀搜尋法之語音品質差異,接著,我們改進行 A/B/A 測試。所設計的 A/B/A 測試為:隨機播放原深度優先樹狀搜尋法之語音與所測試搜尋法之語音,然後由受測者說出所喜歡的語音為何者。未避免先入為主的觀念,播放的順序為 A/B/A。如表 2 所示,測試者還是無法分辨出兩語音的差異。

表 2 主觀語音品質測試與搜尋迴圈數比較

Method	Informal MOS Degradation Tests	A/B/A Tests	Loops
DFS	-	-	768
GPR	~ 0*	=50%†	218.54
$N=1$	~ 0	=50%	80.48
$N=2$	~ 0	=50%	82.59
$N=3$	~ 0	=50%	104.49
$N=4$	~ 0	=50%	112.10
$N=5$	~ 0	=50%	108.66
$N=6$	~ 0	=50%	137.11
$N=7$	~ 0	=50%	164.22
$N=8$	~ 0	=50%	191.49
$N=9$	~ 0	=50%	217.23
$N=10$	~ 0	=50%	232.69

* ~ 0 表示測試者無法判定 MOS 衰退值

† =50% 表示測試者無法分辨差異

因此，我們再進行 ITU-T 所規範的客觀語音品質評定標準。MOS-LQO (MOS-listening quality objective) 為 ITU-T P.862.1 [11] 指定之以客觀 MOS 評估方式來評定編碼後的語音品質。ITU-T P.862.1 乃是將原先 ITU-T P.862 [12] 計算的語音品質感官評估分數 (Perceptual evaluation of speech quality, PESQ) 進行轉換，使其轉換後的評分值與主觀 MOS 評分有更線性的關係。在本論文實驗中，我們亦使用 ITU-T P.862 所提供的 23 個測試音檔，同時我們比較了各音檔 MOS-LQO 的變異百分比。在與原始深度優先樹狀搜尋法為比較基準下，表 3 顯示並比較了不同 N 值設定下 MOS-LQO 的平均值、各音檔 MOS-LQO 衰退百分比平均、各音檔 MOS-LQO 衰退百分比標準差、實際迴圈數紀錄。如表 3 所示；修正式全域脈衝搜尋法的搜尋迴圈數隨著參數 N 的增加而增加。但是，即使設定 $N=1$ 的情況下，搜尋迴圈數僅約 80.48，編碼語音品質僅下降 -0.43%，也因此人耳無從分辨其差異性。

觀察表 3 中 $N=9$ 的情況，MOS-LQO 衰退百分比平均出現了 +0.0908%。此數據並非為不合理現象，因為語音編碼中，各級編碼程序均會相互影響，逐級最佳化者，並非為整體最佳化之解。在以 ITU-T G.729E 所使用的深度優先樹狀搜尋法為比較基準下，不同的搜尋法，有可能找到以編碼後語音為比較依據之更佳解。

表 3. MOS-LQO 與搜尋迴圈數比較

Method	Average	Normalized Degradation	Standard deviation	Loops
DFS	4.21	-	-	768
GPR	4.20	-0.2772%	6.079×10^{-3}	218.54
$N=1$	4.19	-0.4363%	7.225×10^{-3}	80.48
$N=2$	4.18	-0.6262%	7.952×10^{-3}	82.59
$N=3$	4.20	-0.2632%	8.133×10^{-3}	104.49
$N=4$	4.19	-0.3654%	7.862×10^{-3}	112.10
$N=5$	4.19	-0.3502%	6.692×10^{-3}	108.66
$N=6$	4.19	-0.3585%	10.115×10^{-3}	137.11
$N=7$	4.19	-0.4019%	8.083×10^{-3}	164.22
$N=8$	4.19	-0.4084%	6.834×10^{-3}	191.49
$N=9$	4.21	+0.0908%	8.070×10^{-3}	217.23
$N=10$	4.20	-0.1969%	6.079×10^{-3}	232.69

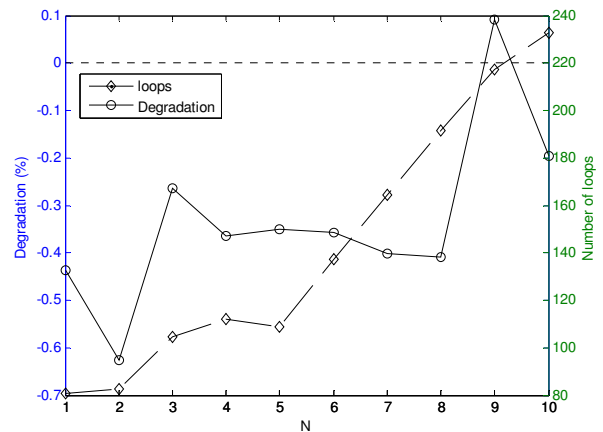


圖 2、搜尋迴圈數與客觀語音品質權衡圖

為了顯示出所提之修正式全域脈衝搜尋法不同的參數 N 、MOS-LQO 衰退百分比平均、實際搜尋迴圈數三者間的關係與趨勢；圖 2 顯示了這三者間的權衡關係。觀察圖 2 曲線可發現權衡點出現在 $N=6$ 的情況下。但是，進一步觀察可發現，最佳權衡條件應在 $N=5$ 時，因為其迴圈數更低、品質更好。這個現象相當合理！因為，在 10 根脈衝下，在每搜尋 $N=5$ 根脈衝時，即進行最佳位置更換，最可能最快（有一半機會）跳脫搜尋區域最佳解！

從圖 2 的觀察，我們決定觀察各關鍵 N 值的 23 個測試音檔個別 MOS-LQO 衰減情況。圖 3 描繪了， $N=1, 3, 5, 9$ 與全域脈衝替換法 (GPR) 之個別測試音檔 MOS-LQO 衰減百分比。與 ITU-T G.729E 所使用的深度優先樹狀搜尋法相比，可明顯看出 $N=9$ 時，多數音檔明顯的有語音品質增加。

5. 結論

本論文針對 ACELP 語音壓縮編碼器具有單軌多脈衝之隨機碼簿，設計並研究全域脈衝替換之搜尋流程以降低搜尋計算複雜度。在以 ITU-T G.729E 為實驗平台下，驗證了所提出的修正式全域脈衝搜尋法，在與全域脈衝搜尋法相同搜尋複雜度下，可提供更加的語音品質。同時，所提出的修正式全域脈衝搜尋法並具有複雜度可階化 (Complexity scalability) 的架構。在聽覺感官語音品質無損條件下，所提出的修正式全域脈衝搜尋法計算複雜度僅剩為原始 ITU-T G.729E 之 10.5% 至 30.3%，且不影响編碼語音的聽覺感官品質。

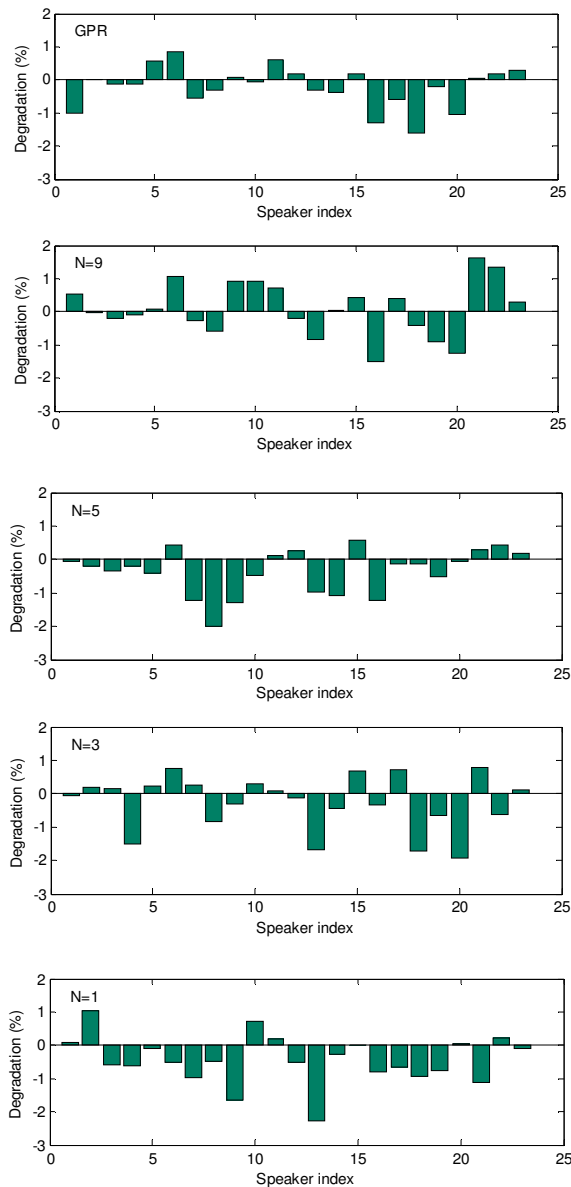


圖 3、23 個標準測試音檔之 MOS-LQO 衰退百分比

參考文獻

[1] A. M. Kondoz, “*Digital Speech-Coding for Low Bit Rate Communications Systems*”, John Wiley & Sons Ltd., 1994.

[2] E. D. Lee, D. Y. Kim, “Method for searching fixed codebook based upon global pulse replacement”, *United States Patent Application Publication*, No US20040193410A1, Sep. 2004.

[3] E. D. Lee, M. S. Lee, and Kim, D.Y. “Global pulse replacement method for fixed codebook search of ACELP speech codec”, *Proc. 2nd IASTED Int. Conf. on Communications*,

Internet and Information Technology, pp.372-375, November. 2003

[4] Fu-Kun CHEN and Yu-Ruei TSAI, “Switching Search Method for Pulse Assignment in ITU-T G.729D”, *IEICE Transactions on Information and Systems 2008*

[5] G. Thomsen and Y. Jani, “Internet telephony: going like crazy”, *IEEE Spectrum*, Vol. 37, No.5, pp. 52-58, May 2000.

[6] H. Park, Y. Choi and D. Lee, “Efficient Codebook Search Method for ACELP Speech Coders”, *IEEE Workshop Proceedings*, pp.6-9, Oct. 2002

[7] ISO/IEC 14496-3, “*Information technology -very low bitrate audio-visual coding*”, Part 3:Audio, Mar., 2001.

[8] *ITU-T Recommendation G.729*, “Coding of speech at 8 kbit/s using conjugate structure algebraic code excited linear prediction (CS-ACELP)”, March 1996.

[9] *ITU-T Recommendation G.729 Annex A*, “Reduced complexity 8 kbit/s CS-ACELP speech codec”, March 1996.

[10] *ITU-T Recommendation G.729 Annex E*, “11.8 kbit/s CS-ACELP speech coding algorithm”, Sep.1998.

[11] *ITU-T Recommendation P.862.1*, Telephone transmission quality methods for objective and subjective assessment of quality: Mapping function for transforming P.862 raw result scores to MOS-LQO, Nov. 2003.

[12] *ITU-T Recommendation P.862*, Telephone transmission quality methods for objective and subjective assessment of quality: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech coders, Feb. 2001.

[13] Kyung Jin Byun, et. al., “A fast ACELP codebook search method”, *Proc. ICASSP*, pp. 422-425, 2002.

[14] *3GPP TS 26.090*: “AMR Speech Codec; Speech Transcoding Functions”, 3GPP, Dec. 2004.