■ 特輯 · 技術解說

音聲認識

吳 永 煥
韓國科學技術大學 副教授 (工博)

1. 緒 論

사회가 점차 복잡해지고 情報化되어 감에 따라 機械와 사람간의 자유로운 意思傳達과 情報交換의 必要性이 漸增하고 있다. 特히 컴퓨터에의 入出力에 펀칭카드, 키보드나 CRT디스플레이등의 媒体를 거치지 않고 人間 相互間의 意思傳達手段인 "말(言語)"을 使用하려는 試圖는 오래 전부터 있었다.

1940년을 前後한 Dudley의 vocoder[1]와 Potter 등의 visible speech[2] 연구가 근대음성학의 基礎를 이룬 이래, 최근에는 수백단어 정도의 단어음성 인식기기의 실용화에까지 到達하였다. 그러나, 이와 같은 연구이전에도 소설 아라비안 나이트에 나오는 "열려라 참깨"에서 보는 음성인식의 응용이나, 18世紀 後半에 이루어진 Kratgenstein, von Kempelen[3]의 풀무를 이용한 기계적 음성합성장치에서 보는 바와 같이, 인간의 "말하는 기계"나 "듣는 기계"에의 희망은 古代로 거슬러 올라 갈 수 있다.

人間의 聽覺에서 이루어지는 廣義의 音聲認識(speech recognition)을 機械(컴퓨터)를 利用하여 實現하는 技術은 크게 둘로 나눌 수 있다. 첫째는 音聲이 傳達하는 情報중에서 言語的 內容을 機械的 手段으로 抽出하는 狹義의 音聲認識이며, 둘째는 音聲이 傳達하는 話者情報중에서 個人性을 利用하여 發聲者를 確認 또는 認識하는 話者認識(speaker recognition)이다.

本稿에서는 音聲認識시스템의 構成, 音聲認識의 障碍要因, 利點, 應用 및 代表的인 方法에 關해 記述하고, 끝으로 將來의 展望과 結論의 順으로 展開할 것이다.

II. 音聲認識 시스템의 構成

일반적인 음성인식시스템의 구성을 그림1에 보인다.

時系別 信號인 음성파형은 前處理 段階를 거쳐, 特徵抽出단계에서 特徵파라미터(feature parameter)로 變換되어 다음의 認識단계에서 特定한 pattern class로 分類된다. 以下, 各 段階에 対해 記述한다.

1. 前處理(preprocessing)

다음 段階에서의 適切한 音聲分析이 이루어지기 위해서, 先行해야 할 몇 가지의 處理가 있다. 以下, 前處理 段階에서 考慮할 事項을 보인다.

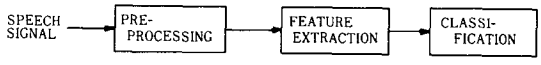


그림 1. 音聲認識시스템의 構成

① 音聲區間의 檢出

音聲信號의 始点과 終点을 決定하는 過程으로, 背景雜音下에서의 音聲區間의 檢出은 特히 無聲子音으로 시작되는 境遇等에는 相當한 어려움이 뒤따르나, 餘分의 計算時間과 記憶容量을 浪費하지 않고, 過誤率을 줄이기 위해서라도 適切히 이루어져야 한다.

② 分割(segmentation)

後段에서 必要한 境遇에, 認識單位로 入力音聲을 分割한다. 예를 들면, 連續音聲認識에서의 單語나 句節單位로의 分割 等이 이에 包含된다.

③ 雜音除去

背景雜音의 影響을 最小化시켜 入力音聲의 音質을 높인다.

④ 其他의 考慮事項

以上의 處理外에 A/D變換에 따른 振幅의 正規化,antialiasing filter에 의한 harmonics의 除去, 分析frame單位의 切斷에 의한 spectral distortion을 줄이기 위한 適切한 time window의 選定, 高周波成分을 強調하기 위한 pre-emphasis 等이 考慮되어야 한다.

2. 特徵抽出

音響分析을 통해 特徵타라미터를 抽出한다. 特徵파라미터는 되도록 變形에 対해 安定하며, 패턴class 내

의 分散이 最小이면서도 同時에 패턴class間의 分散을 最大로 하는 音響파라미터가 理想的이다.

勿論, 可能한 限 次元数가 낮은 物理量을 選擇하여 認識을 위한 情報量과 處理量을 輕減함과 同時에, 認識対象의 內容을 本質的으로 表現하는 特徵을 찾아 내는 것이 課題다.

3. 識 別 (classification)

未知의 패턴이 주어졌을 때, 어느 category에 속하는가를 決定하는 段階다.

識別手法은 特徵파라미터에 識別函数를 適用시켜 入力音聲이 屬하는 패턴클래스를 決定하는 決定理論的 手法과, 特徵의 存在有無나 出現形態 등을 表現하는 文法에 適合한지 與否를 決定하는 言語理論的 接近法으로 나눌 수 있다.

一般的으로 單語音聲認識시스템에서는 前者의 手法이 連續音聲認識의 境遇는 後者의 手法이 많이 쓰인다.

Ⅲ. 音聲認識의 特性

1. 音聲認識의 障碍要因

음성인식을 實現하는데에는 몇 가지의 難点을 克服하지 않으면 안된다. 即, 調音結合, 時間軸의 整合問題와 個人差의 正規化 問題를 들 수 있다.

(1) 調音結合 (coarticulation)

言語는 聽覺的으로 區別할 수 있는 基本單位인 音素(phoneme)를 가지고 있으나, 이를 發聲하는 調音器官의 運動은 連續的이기 때문에 前後의 音素의 影響을 받아 變形된 音聲學的 特性을 나타 내게 된다. 이와 같은 現象을 調音結合이라 하며, 음성인식을 어렵게 하는 本質的인 原因이다. 子音接變, 口蓋音化, 無聲子音의 有聲音化, 母音의 鼻音化現象등이 代表的인 境遇로 音素環境에 따른 音素의 特性變化에 対處할 必要가 있다.

(2) 時間軸의 正規化

同一한 單語를 同一人이 反復해서 發聲해도 特別한 訓練을 받지 않는 경우 音聲의 持續時間은 一定치 않다. 이 問題는 서로 다른 길이의 音聲패턴의 整合(matching)時 어려움을 주었으나, 近年에 들어 dynamic time warping 手法이 実用化되면서 解決되었다 볼 수 있다.

(3) 個人差

發聲하는 各 話者는 調音器官의 解部學的 構造가 다르며, 言語의 習得過程에서 얻어진 發聲習慣의 差異에서 音聲波의 音響學的 特性이 많이 달라지게 된다. 前者의 境遇가 先天的 要因이라면 後者는 後天的 要因이라 볼 수 있다.

音聲認識시스템의 性能을 向上시키기 위해서는 認識可能한 話者数를 增加시켜야 하며, 이를 実現하기 위해서는 話者에 対한 正規化 方法이 先行되거나, 효과적이고 간편한 學習法이 뒤따라야 한다.

2. 音聲認識

음성인식은 狹義의 음성인식과 話者認識으로 나눌 수 있다. 本節에서는 單語音聲認識과 連續音聲認識 및 理解시스템으로 나누어 說明하고 話者認識의 特性과 그 應用에 관해 記述한다.

1) 單語音聲認識

單語音聲인식시스템의 특징은, 연속음성인식의 原理的인 難点을 시스템的으로 廻避하여 音聲認識을 実用化하기 위한 工學的 解決策이라 볼 수 있다. 그 特徵은 크게 다음의 네 가지로 要約할 수 있다.

① 單語別로 區分發聲한 限定된 어휘를 使用하여 單語間의 連續發聲에 의한 調音結合의 影響을 排除할 수 있다.

② 單語패턴 自体를 標準패턴으로 使用하여 單語內에서의 調音結合의 영향을 줄일 수 있다. 어휘수가 늘면 VCV Chain(母音-子音-母音 連鎖)등을 認識單位로 하는 方法도 있다.

③ 話者別로 標準패턴을 設定하므로써 個人差에 対處하기 위한 話者正規化의 難題를 避할 수 있다.

④ 非線形 時間軸整合을 使用하여 單語의 지속시간의 差를 正規化할 수 있다. Dynamic programming 手法을 利用한 pattrn matching의 경우, 最小二乘誤差의 意味에서 最適整合經路를 求하게 된다.

위와 같은 手法을 利用하여 実用化 單語音聲 시스템이 市販되고 있으며, 區分發聲의 境遇에 最大 500單語程度를 数% 以內의 過誤率로 認識이 可能한 段階에 와 있다. 限定된 分野에서의 應用은 더욱 擴大될 것으로 豫測되고 있다.

2) 連續音聲認識

連續으로 發聲된 音聲信號로 부터 離散的인 言語情報를 抽出하는 過程으로, 單語音聲에 比해 ① 單語間의 境界가 不明確하고 ② 單語境界付近의 音이 先行 또는 後續單語의 影響을 받아 調音結合이 일어나며 ③ 單語別로 떼어 發音할 때 보다 速度가 빨라져서, 各音 또는 個個의 單語의 持續時間이 짧아져 애매한 發聲이 되기 쉽다.

위와 같은 特性이 연속음성인식을 飛躍的으로 어렵게 하는 要因이 되며, 單語認識시스템과는 相異한 情報를 利用하게 한다. 即, 연속음성에서는 特定한 task를 設定함으로써, 文章의 構造를 規定하는 構文情報,

task에서 使用되는 單語間의 槪念関係나 屬性関係를 表現하는 意味情報, 設定된 世界에서의 一般常識, 시스템과 利用者와의 対話를 通해서 얻어지는 文脈情報와 音聲의 抑揚등의 韻律情報등을 利用할 수 있다.

連續音聲認識의 경우, 音聲理解시스템(Speech Understanding System : SUS)도 包含해서 말하는 것이 普通이며, 前者에서는 入力音聲의 한 音씩 또는 한 單語씩 모두 正確히 認識해야 하는데 比해, 後者는 發聲者가 意圖한 言語的 內容을 把握하여, 그에 따른 行動이나 應答을 正確히 하면 音聲을 "理解"했다고 본다.

音聲理解시스템의 硏究는 1971年 부터 5年間 美國의 ARPA가 지원한 "Speech Understanding Project"에서 本格的으로 硏究가 推進되어 많은 成果를 거두었다. 그 중에서도 當初의 目標를 유일하게 달성한 Carnegie-Mellon大學의 Harpy시스템은 文献檢索用 task의 1011單語로 된 連續音聲을 過誤率 5%以下로 理解하는 性能을 가지고 있다.

그럼에도 不拘하고 연속 음성인식 시스템은 넘어야 할 障碍要因이 많아 実用化와는 먼 거리에 있어, 앞으로의 硏究에 거는 期待가 크다 하겠다.

3) 話者認聲

위에서 언급한 音聲認識의 境遇, 音聲이 지니는 言語情報를 利用하고 個人差를 最小로 하기 위한 努力을 하는 反面, 話者認識은 言語情報를 最小로 하면서 同時에 個人性 情報를 最大로 할 수 있는 識別方法 및 特徵파라미터의 抽出이 課題이다.

話者認識에는 使用目的에 따라 둘로 나눌 수 있다. 하나는 特定人이라고 自稱하는 認識対象이 本人인지 與否를 判定해 주는 話者確認 (speaker verification) 시스템이고, 다른 하나는 發話者가 登録된 話者中에서 누구인지를 判定하는 話者認識(speaker identification) 시스템이다.

前者의 경우 判定에 必要한 候補의 数는 한 개 이며, 決定은 Yes나 No중의 하나이고, 패턴의 比較는 1回이므로 過誤率은 確認対象패턴의 数에 無關하다. 한편, 後者의 경우는 候補가 複数인 N개이고, 決定은 N個中 어느 것인지 判定하고 N回의 패턴比較가 必要하므로, 過誤率도 N에 比例해서 增加하는 特徵이 있다.

話者認識은 또 使用하는 데이타의 種類에 따라 text independent 話者認識시스템과 text dependent 話者認識 시스템으로 分類할 수 있다. 前者는 發話內容에 関係없이 話者를 認識可能한 시스템으로서 發聲器官等의 生理的 差異에 起因하는 特徵파라미터를 使用하게 된다. text dependent한 시스템은 同一한 發話內容을 利用하여 話者를 認識하는 方法으로, 識別은 音韻性에 基礎를 두게 되어, 大体的으로 音聲認識에서와 거의 같은 알고리즘과 特徵파라미터를 使用하는 것이 普通이다.

一般的으로는 後者의 境遇가 시스템構成이 보다 容易하다. 한편, 話者認識시스템의 実用化 時에 考慮해야 할 問題로,

첫째 信賴性과 經濟性을 들 수 있는데, 特히 時間이 지남에 따른 特性變化등에 有効하게 対處해야 信賴性을 높일 수 있으며, 特히 多数話者를 対象으로하는 話者確認시스템 등에서는 그 性能이 더욱 重要視된다.

둘째는 發聲에 따른 背景雜音 問題와 電話回線 使用時의 傳送特性의 影響을 줄이거나 除去하는 問題이며,

셋째로는 詐稱者(imposter)를 어떻게 가려내서 棄却하느냐인데, 特히 쌍둥이나 肉親間의 흉내를 가려내는 硏究結果도 報告되고 있으며 이러한 問題들의 根本的인 解決없이는 話者認識시스템의 実用化는 어려울 것으로 보인다. 끝으로, 疾病 등의 原因에 의한 音聲質의 變化에 어떻게 対處할 것인가도 実用化를 위한 重要한 前提條件이 된다.

Ⅳ. 音聲의 分析方法

音聲波形으로부터 認識에 有効한 音響파라미터를 抽出하는 方法에 関해서는 훌륭한 文献들이 多数있으므로 (例를 들면(4)～(8)), 詳細한 原理等은 參考文献에 미루기도 하고, 本稿에서는 代表的인 音聲分析 方法에 대해서만 簡單히 言及하기로 한다.

1. 時間領域에서의 分析

① 에너지 : 短區間內에서의 에너지를 求해, 有聲·無聲音 分析등에 利用한다.

② 零交叉波分析(gero crossing wave analysis) : 音聲信號의 符號만 남겨 振幅을 1 bit로 量子化한 것으로, 数理的 解析이나 聽取test의 結果, 音聲情報를 잘 保存하고 있는 것이 確認되었다. Spectrum 중에서의 優勢한 周波数 成分속에 対應이 잘 되므로 적은 情報로 스펙트럼의 大体的인 特徵을 表現하는데 適合하다.

③ 短時間 自己 相関分析

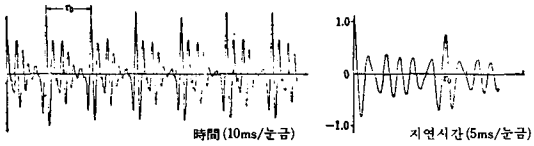
短時間 自己 相関係数는

$$\varphi(m)=\frac{1}{N}\sum_{n=0}^{N-1}x(m)\,x(n+m)$$

$$(0\leq m\leq Mo-1) \qquad (1)$$

으로 定義하는데 $\varphi(m)$의 값은 m=0일 때 最大를 나타내고 m이 增加함에 따라 急速히 減小하는 性質이 있다. 또, 原信號의 基本周期만큼 遲延된 경우 顯著한

peak를 보여, 基本周波数의 檢出에 널리 쓰여 왔다.



(a) 〔아〕의 波形 (b) 正規化된 短時間 自己相關系數

그림 2. 短時間 自己相關係數

2. 周波数領域에서의 分析

音聲의 言語로서의 特徵은 主로 聲道(vocal tract)의 共振系에 依存하는 바 크며, 共振의 表現은 音聲스펙트럼이나 聲道의 周波数 傳達函数가 보다 明確하고 理由하기 쉽다. 또한 人間의 聽覺에서의 音聲處理는 周波数分析에 바탕을 둔다는 事実을 감안하면 스펙트럼 分析의 重要性을 더 잘 알 수 있다.

① bandpass filter bank에 의한 分析

分析하는 周波数 範圍를 隣接하는 여러 帶域通過필터로 덮을 수 있도록 配置한 것으로, 実時間에 音聲의 短時間스펙트럼의 大体的인 形態를 얻을 수 있는 長點과 스펙트럼의 概形이 音聲의 音韻性을 잘 지니는 特性이 있어 現在도 音聲認識에서의 用途가 넓다.

② Fourier 變換

離散的 Fourier變換(Discrete Fourier transform; DFT)과 高速 Fourier變換(fast Fourier transform; FFT)은 時間領域의 音聲信號를 周波数領域의 스펙트럼으로 變換하는 가장 一般的인 手法이라 볼 수 있다. 詳細는 他 文献으로 미룬다.

3. Cepstrum分析

本來는 地震波의 解析에 쓰이던 手法으로, 音源의 基本周期檢出에 使用된 以來, 主要한 音聲分析法이 되어 音聲研究全般에 널리 쓰이고 있다. Cepstrum分析의 過程을 그림 3에 보인다. 한편, Cepstrum은 Spectrum의 前半部를 反轉하여 만든 用語이며 같은 方法으로 quefrency라는 變数는 freguency로 부터 新造한 말이다.

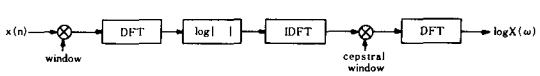


그림 3. Cepstrum 分析의 過程

4. 線形豫測分析(LPC分析)[6]

音聲生成 model의 parameter를 音聲波形으로부터, 直接 求하는 方法으로 提案된 以來, 音聲의 分析, 合成 및 認識등 거의 全 分野에 걸쳐 現代 音聲分析 手法中에서 가장 有力하고 廣範圍하게 利用되고 있다. 線形豫測係数, PARCOR分析에 의한 反射係数(部分自己相関係数), 最近의 線스펙트럼雙(line spectrum pair; LSP)[8]등의 파라미터가 모두 LPC分析에 基盤을 두고 있다.

本項에 関해서도 餘他의 文献에 詳細를 미루기로 한다.

V. 音聲認識의 現狀과 應用

여기에서는 代表的인 音聲認識시스템의 実用化 程度를 概括的으로 記述하고, 국어 音聲認識研究의 動向에 대해 살펴 본 다음, 実生活에서의 音聲認識시스템의 用途와 認識시스템 導入의 利点에 관해서 略述한다.

1. 音聲認識의 現狀

最初에 商用으로 市販된 音聲認識시스템은 美國의 Threshold Technology社(TTI)의 10数字音 専用認識機였으며, 70年代 後半에 들어 單語音聲機의 販賣가 美國과 日本의 메이커를 中心으로 開始되었다. 例를 들면 Interstate Electronics의 VDES-1000, TTI의 VI-P-600과 NEC의 DP-100등이 있다.

特定話者가 區分해서 發聲하는 單語音聲의 境遇에 最大 900單語(VDES-1000) 정도 認識可能하나, 大部分 50~500單語認識機가 主宗을 이루고 있다.

連續音聲의 認識이나 理解시스템은 実用化와는 아직 相當한 距離에 있어 앞으로의 研究成課에 거는 期待가 크다.

한편, 話者認識시스템에 있어서는 狹義의 話者認識에 関한 実用化 成果는 뚜렷한 것이 없는 反面, 話者確認은 電話回線을 使用한 Bell연구소의 시스템[9]과 Texas Instruments社의 電算室 出入管理시스템[10]이 代表的이다. 特히 後者의 境遇 200人 程度의 話者를 対象으로 text dependent한 시스템을 構成하여 数年째 試用하고 있으나 아직 商品化된 話者認識시스템은 紹介되지 않았다.

끝으로 韓國語의 音聲認識 研究를 살펴 보기로 한다.

基本이 되는 8母音에 対한 基礎的인 認識実験이 報告된 以來,[12] 주로 區分해서 發聲한 10数字音에 対한 実驗結果가 發表되었으나(예를 들면[13][14]), 多数音聲에 対한 実用化 研究가 不足한 実情이다. 認識分野以外의 全般的인 音聲研究도 包含하여 많은 基礎研究가 앞으로 이루어져야 할 것이다.

2. 音聲認識의 應用

以上에서 記述한 音聲認識시스템 使用時의 利点을 列

擧하면,

① 使用에 따른 特別한 訓練이 不必要하여 便利하며

② 發聲하면서 눈, 귀 등을 同時에 使用할 수 있어 並列的인 情報處理가 可能하며

③ 몸을 움직이면서 情報發生이 可能하며(運動의 自由)

④ 他 入力手段에 比해 高速이며(區分發聲의 境遇,타이핑의 2배, 連續音聲의 境遇는 약4배 程度임)

⑤ 信賴性이 높으며

⑥ 遠隔入力이 可能하고(電話回線등)

⑦ 突發事態時에도 臨機入力이 可能하며

⑧ 長時間에 걸친 連續入力이 可能하며

⑨ 말의 內容에 더해 話者確認도 同時에 可能하며

⑩ 入力費用이 節減되어 經濟的인 点을 들 수 있다.

音聲認識의 應用分野를 크게 나누면 두 가지로 볼 수 있다. 하나는 機械의 音聲에 依한 制御이고, 다른 하나는 音聲에 依한 데이타의 入力이다. 前者에 屬하는 應用으로 物品 分類시스템, robot制御, 生産工程의 制御, 身体障碍者의 補助用品, 장난감, CAI등의 教育시스템, 家電製品等의 音聲制御가 있으며, 後者에 屬하는 應用으로는 航空管制, CAD시스템, CNC의 프로그램制御, 檢査·接受·發送, 컴퓨터의 터미날등에서의 音聲入力을 들 수 있다.

한편, 話者認識의 應用分野로는, ① 情報檢索시스템에서의 個人file의 保護 ② 信用카드나 身分證등의 安全裝置 ③ 統制區域에의 出入管理 ④ 電話를 使用한 犯罪의 防止 및 搜査 等으로 社會에서의 要求가 크다.

Ⅵ. 맺 음 말

將來의 音聲處理 機器의 需要豫測[15]을 보면 1982년부터 1987年까지의 5년간에 2600萬弗에서 7억 8천만弗로 30倍의 伸張이 있는데, 그 중에서 音聲認識 機器는 1982年의 500萬弗에서 1987年의 2億7千萬弗로 約 54倍의 伸張率이 있을 것으로 보여 他 電子機器 中에서도 아주 높은 增加率을 보이는 有望한 分野다.

또 다른 豫測[16]은 1萬弗 以下의 連續音聲認識 裝置가 市販되면 그 需要는 가히 爆發的일 것으로 보고 있다. 現在, 연속 음성인식기의 早期 実用可能性은 稀薄하지만, 低價格, 高速의 單語認識機를 中心으로 하여 需要는 꾸준히 增大될 것이며, 그와 並行하여 연속음성기의 研究도 많이 進前될 것이다.

우리들이 指向하는 "사람과 機械間의 말에 의한 自由로운 情報交換(man-machine communication by voice)에 必須 不可缺한 音聲處理 技術에 대한 보다 많은 関心과 支援이 이루어져야, 眞情한 意味에서 컴퓨터가 大衆化될 날이 앞당겨 질 것이다.

參 考 文 献

[1] Dudley, H.; "The vocoder", Bell Lab. Record, vol. 17, pp. 122-126, 1939.

[2] Potter, R.K., Kopp, G.A., Green, H.C.; *Visible Speech,* D. van Nostrand Co., N.Y., 1947.

[3] Dudley, H. and Tarnoczy, T.H.; "The speaking machine of Wolfgang von Kempelen", *J. Acoustical Society of America,* vol. 22, pp. 151-166, 1950.

[4] Rabiner, L.R. and Schafer, R.W.; *Digital processing of speech signals,* Prentice-Hall, New Jersey, 1978.

[5] Flanagan, J.L.; *Speech analysis, synthesis and perception,* Springer-Verlag, N.Y., 1972.

[6] Markel, J.D. and Gray, Jr. A.H.; *Linear prediction of speech,* Springer-Verlag, N. Y., 1976.

[7] 中田和男; 音声, Corona社, 1977.

[8] 斉藤, 中田; 音声情報處理의 基礎, Ohm社, 1981.

[9] Doddington, G.R.; "A method of speaker verification", *J. Acount. Soc. America,* vol. 49-1, 1970.

[10] 新美康永; 音声認識, 共立出版, 東京, 1979.

[11] Teja, E.R. and Gonnella, G.; *Voice technology,* Prentice-Hall Co., Reston, 1983.

[12] Kim, B. and Fujisaki, H.; "Analysis and recognition of Korean vowels', Annual Report of Eng., Univ. of Tokyo, vol. 32, pp. 227-232, 1973.

[13] 安居院, 呉永煥; "Formant tracking의 一手法과 그의 韓國語 數字音認識에의 應用", 日本電子通信學會誌 vol. J63-A no. 4, pp. 322-323, 1980.

[14] 呉永煥; "音素間 類似度의 整數空間에의 投影에 의한 單語音聲의 認識", 韓國情報科學會誌 10卷 4號, pp. 254-259, 1983.

[15] Nance, J.; "Implementation strategies for voice-processing terminals", Mini-Micro Systems, pp. 183-194, Nov.,1983.

[16] 中田; 패턴認識과 그의 應用, Corona社, pp. 155-156, 東京, 1978. *