論 文

論文 78-15-6-3

숫자音聲 自動 認識에 關한 一實驗

(An Experiment of a Spoken Digits-Recognition System)

吳 永 煥*, 安 居 院 猛**

(Oh, Yung Hwan and Agui, Ta Keshi)

要 約

本 論文은 複數話者를 對象으로 한 숫자音聲自動 시스템의 開發을 위한 基礎 實驗 結果의 報告다. ZCR, 對數 에너지등의 파라메터에 의한 無聲子音의 分類, 線形豫測에 의한 formant 周波數의 推定 및 그를 利用한 母音 및 有聲子音의 認識을 行했다. 成人 男性 한 사람의 숫자音에 對한 認識實驗의 結果, 音素(phoneme) 結合時의 過渡 部分이나, 音素 認識 段階에서의 局所的 誤認識을 吸收 할 수 있는 algorithm을 採用함으로써 良好한 認識 結果를 얻을 수 있었다. 앞으로, 複數話者를 對象으로 한 認識實驗, 認識시스템의 改善과 함께 國語의 音聲學的 諸性質의 硏究를 해 나갈 豫定이다.

Abstract

This paper describes a speech recognition system for ten isolated spoken digits.

In this system, acoustic parameters such as zero crossing rate, log energy and three formant frequencies estimated by linear prediction method were extracted for classification and/or recognition purpose(s). The former two parameters were used for the classification of unvoiced consonants and the latter one for the recognition of vowels and voiced consonants.

Promising recognition results were obtained in this experiment for ten digit utterances spoken by a male speaker.

1. 序 論

電子計算機의 發達에 따라, 從來의 Vocoder, sonagraph등으로 代表되는 analog 的 音聲分析으로부터 digital 技術에 依存하는 音聲分析·合成등이 높은 比重을 차지하게 되었다. 人間 相互間의 情報傳達手段中에서 가장 自然스러운 形態의 音聲을 人間과 機械間의 情報傳達에 利用하려는 man-machine communication 시스템[1] (computer에의 入力手段, 無人豫約시스템, 機械 操作등)이나 話者確認(speaker verification)[1]등의 硏究는 世界 各國에서 活潑히 進行되고 있으며, 制限 된 發聲者에 의한 限定語의 自動認識은 相當한 成果를 올리고 있다.

*正會員, 大田工業專門學校
(Dept. of Electronics Engineering, Member, Daejon Technical Junior College)
**非會員, 東京工業大學
(Tokyo Institute of Technology, Japan)
接受日字 : 1978年 5月 29日

한편, 現在까지 報告 되어 온 國語의 音聲 내지는 音聲 自動 認識에 關한 硏究의 槪略은 다음과 같다.

(가) 子音 ; 주로 言語學的 觀心으로 부터 發聲時의 發聲器官의 觀測에 의한 調音(articulation)構造의 比較에 關한 硏究가 大部分으로, 國語 特有의 된 소리, 거센소리등(즉, /ㄱ, ㄲ, ㅋ/, /ㄷ, ㄸ, /ㅌ, /ㅂ, ㅃ, ㅍ/)이 硏究 對象이다.

(나) 母音 ; 주로 金博, 藤崎등에 의해 아, 어, 오, 우, 으, 이, 애, 에의 8母音을 中心으로 이루어졌다. 즉, 發聲器官의 調音 現象에 關한 硏究[5,6], 發聲者 13名分의 母音으로부터, analysis-by-synthesis(A-b-S)法에 의한 formant 周波數 F_1, F_2, F_3의 抽出 및 그에 의한 認識 實驗[7]등이다.

限定語 音聲認識의 一般的 對象으로는, 숫자, alphabet, Fortran 音聲등을 들 수 있다. 特히, 숫자 音聲認識은 各國에서 오래 前 부터 硏究되어 왔으나,[8,9] 國語에 對한 硏究結果는 아직 報告되어 있지 않다. 本 論文은 콤퓨터에 의한 多數話者音聲自動 認識의 可能性을 찾기 위한 基礎的 硏究로, 線形豫測(linear

predictive coding ; LPC)에 의한 formant 周波數의 推定 및 그에 의한 母音 및 有聲子音의 認識, zero crossing rate(ZCR)와 log energy등의 파라메터에 의한 子音의 分類에 관한 認識實驗 結果의 報告다.

2. 認識시스템

音聲認識 시스템의 block diagram을 그림 1에 보인다. 音聲信號를 前處理(pre-processing)한 다음, 特徵 파라메터(feature parameter)를 抽出했다. 이들 파라메터를 使用해 以後 記述하는 方法으로 認識을 행했다.

그림 1. 음성생성 모델
Fig. 1. Speech recognition system.

2.1. 파라메터의 抽出

2.1.1. Formant 周波數의 推定[10]

本 實驗에서 formant 周波數 推定에 使用한 線形豫測에 關해 簡積히 記述한다. 音聲 波形의 sample data $s(n)$을, 入力信號인 Gaussian impulse train $e(n)$이 all-pole(autoregressive) model인 시스템 $1/A(n)$을 通過한 出力으로 보면,

$$s(n)=e(n)/A(n) \tag{1}$$

$$A(n)=1+a_1B+a_2B^2+\cdots\cdots+a_MB^M \tag{2}$$

(但, B는 backward shift operator)

이라 쓸 수 있다. (그림 2)

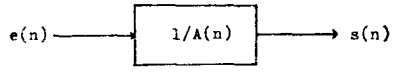


그림 2. 음성생성 모델
Fig. 2. Linear speech production model.

여기서 시스템을 all-pole model로 假定한 理由는, 有聲音, 周期가 p인 impulse train이 無聲音은 flat spectrum인 randum noise가 各各 autoregressive (AR)特性을 지닌 聲道(vocal tract)를 通過한 出力으로 보는 線形音聲生成 model[10]에 基礎를 두고 있다.

한편, (1)式에서 시스템의 入力인 音源 impulse train $e(n)$을 出力인 音聲信號 $S(n)$으로부터 推定하기 위해 inverse filter $A(n)$을 考慮하면,

$$e(n)=s(n)A(n) \tag{3}$$

즉,

$$e(n)=s(n)+\sum_{i=1}^{M}a_is(n-i) \tag{4}$$

이다. M개의 前 data로부터 推定한 sample data $\hat{s}(n)$을

$$\hat{s}(n)=-\sum_{i=1}^{M}a_is(n-i) \tag{5}$$

로 表示하면, (4)式으로 부터

$$e(n)=s(n)-\hat{s}(n) \tag{6}$$

이라 쓸 수 있다. 여기서, $e(n)$은 實際 音聲信號 $s(n)$과 推定信號 $\hat{s}(n)$과의 差로 볼 수 있으므로 豫測誤差(prediction error)라 부른다. 誤差$e(n)$의 제곱합 을

$$\alpha=\sum_{n=1}^{N}e(n)^2 \tag{7}$$

最小로 하는 豫測係數(predictor coefficient) $-a_i$를 決定하기 위해서는 (8)의 正規 方程式(normal equation)을 풀면 된다.

$$\sum_{i=1}^{M}a_ir(|i-j|)=-r(j)\quad(j=1,\cdots M) \tag{8}$$

(8)式은 Yule-walker 方程式이라 부르며,

$$r(l)=\frac{1}{N}\sum_{n=1}^{N-l}s(n)s(n+l)\quad(l\geqq 0) \tag{9}$$

으로, 이는 自己相關係數(autocorrelation coefficient)다. 한편, 音聲信號의 推定 power spectrum $\hat{P}$는 model에 의한 推定 power $1/A[\exp(j\theta)]$와 豫測誤差 $e(n)$의 power σ와의 곱으로 表示 할 수 있다. 즉,

$$|\hat{P}[\exp(j\theta)]|^2=\frac{\sigma^2}{|A(e^{j\theta})|^2}=\left|\frac{\sigma}{A(z)}\right|^2_{z=e^{j\theta}} \tag{10}$$

(10)式의 power spectrum $\hat{P}$상에서 energy가 集中해 있는 formant 周波數(pole의 位置)를 推定 할 수 있다. (이를 peak picking 法[10]이라 부른다)

한편, (10)式의 分母

$$a_Mz^M+a_{M-1}z^{M-1}+\cdots\cdots+a_1z+1=0 \tag{11}$$

의 根

$$z_i=C_ie^{j\lambda i} \tag{12}$$

(但, $C_i>0$, $-\pi\leqq\lambda_i\leqq\pi$, $i=1,\cdots\cdots,M$)

로 부터 formant 周波數 F_i와 bandwidth 를 求할 수 있다[12].

$$F_i=\frac{\lambda_i}{2\pi T}\quad B_i=\frac{\log C_i}{\pi T} \tag{13}$$

(但, T는 sampling 周期)

이를 root solving 法이라 부른다.

本 實驗에서는 mini-computer를 使用하는 關係로 記憶容量 및 計算精度등을 考慮해, (11)式의 高次 方

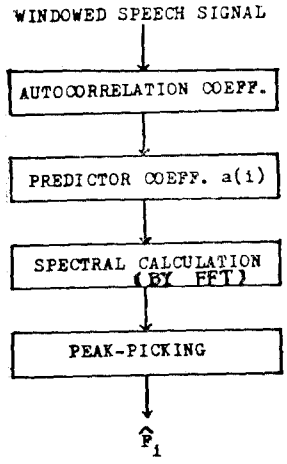
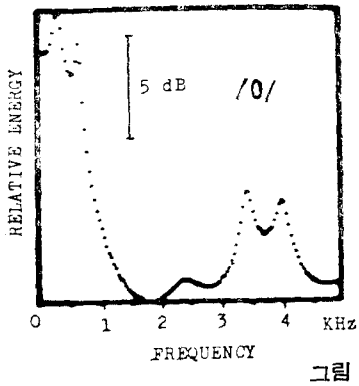


그림 3. Formant주 파수 추정의 유통도
Fig. 3. Flow chart for formant frequencies estimation.



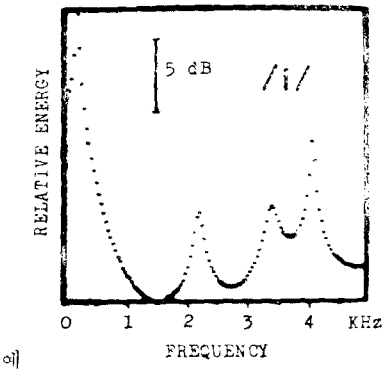


그림 4. 추정 스펙트럼의 display 예
Fig. 4. Example of CRT displayed spectra.

程式을 푸는 代身, 高速 Fourier 變換(FFT)에 의해 計算한 (10)式의 power spectrum 上의 peak의 位置로부터 formant를 推定하는 peak-picking 法을 利用했다. 以上의 LPC에 의한 formant 周波數 抽出의 flow chart를 그림 3에, 計算機에 의해 CRT에 display한 power spectrum의 一例를 그림 4에 보인다. 여기서, peak중 周波數가 낮은 쪽으로 부터 順序대로 제 1, 제2, 제3, formant 周波數로 抽出한다.

2.2.2. 다른 파라메터의 抽出

母音 및 有聲子音은 無聲子音에 比해, 振幅 및 에너지가 크며, 周波數 spectrum上에서의 formant 特性도 比較的 安定하다. 한편, 無聲子音은 周波數 特性이 不規則하므로 formant 周波數 以外의 파라메터에 의한 認識方法이 바람직하다. 本 實驗에서는 分析 frame內의 ZCR 및 에너지를 利用해, 分類 내지는 認識을 행했다. 여기서 ZCR은 分析 區間內에서 波形이 zero軸과 交叉하는 回數를 말하며, log energy LE로는 分析 frame內의 sample data x_i에 對해,

$$LE = 10\log \sum_{i=1}^{N} x_i^2 \qquad (14)$$

을 計算해서 使用했다.

3. 認識實驗

3.1. 데이터의 前 處理

本 實驗에서 行한, 데이터 前 處理의 flow chart를 그림 5에 提示한다. 成人 男性 한 사람이, 조용한 방에서 따로 떼어 發聲한 10개의 숫자음(영, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구)을 tape recorder에 錄音해, cut-off frequency 5kHz인 low-pass filter (LPF)를 通過시킨 다음, sampling 周波數 10kHz, 12bit로 A/D 變換해 minicomputer의 磁氣 테이프(MT)에 記錄했다. Formant 周波數의 推定에는, 切斷에 의한 spectral distortion을 줄이기 위해, 데이터 x_i에 (15)式의 Hamming window를 걸은 後 分析했다.

$$W = 0.54 - 0.46\cos\,[2\pi(n-1)/N] \qquad (15)$$
$$(n = 1, 2, \cdots\cdots, N)$$

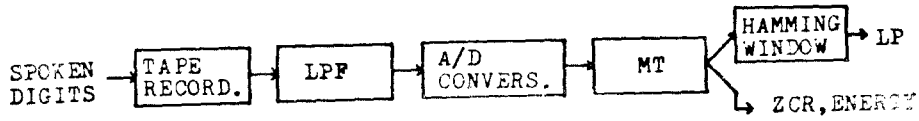


그림 5. 전처리의 유통도
Fig. 5. Flow chart of pre-processing.

本 實驗에서의 frame length $N=256$(25.6ms), frame interval(分析 frame의 移動間隙)은 128 sample (12.8ms), 豫測次數 $M=15$이며, 分析 숫자 音聲의 平均持續 時間은 約 0.5sec 程度였다.

3.2 分析結果

2.1의 algorithm에 의해 推出한 formant 周波數의 推定値 F_1, F_2, F_3의 軌跡(trajectory)을 그림 6에, 2.2에 의해 求한 ZCR 및 에너지를 그림 7에 보인다. 그림 6에 보인 바와 같이, "영"과 "육"에서의 重母音 '여'와 '유'의 경우, '이'와 '어' 또는 '우'사이, "삼"에서의 '아'와 'ㅁ'사이등 formant 周波數의 差가 큰 두 音素間의 過渡部分에 있어서는 單位 frame 當 formant 周波數의 變動率은 他에 비해 크다. 그러므로, 本 實驗에서와 같은 音素認識을 主體로 한 認識시스템에서는 이와 같은 過渡部分이 誤認識의 原因이 되기 쉬우므로, 이를 處理 할 algorithm의 補完이 必要하게 된다. 本 시스템에서는3.3에서 記述한 바와 같이, 一定 frame 以上 同一 音素가 繼續 되었을 때에 限해 그 音素가 存在하는 것으로 보았다. 이에 의해 同一 音素의 持續時間의 比較的 짧은 過渡部分의 認識結果에의 影響을 吸收 할 수 있다.

또한, 그림 7의 結果를 보면, 有聲音(母音 및 有聲子音)과 無聲子音(unvoiced consonant)의 區別은, ZCR과 에너지의 두 파라메터에 適當한 threshold를 設定함으로써 可能하리라 생각된다. 3.3에 實驗結果를 보인다.

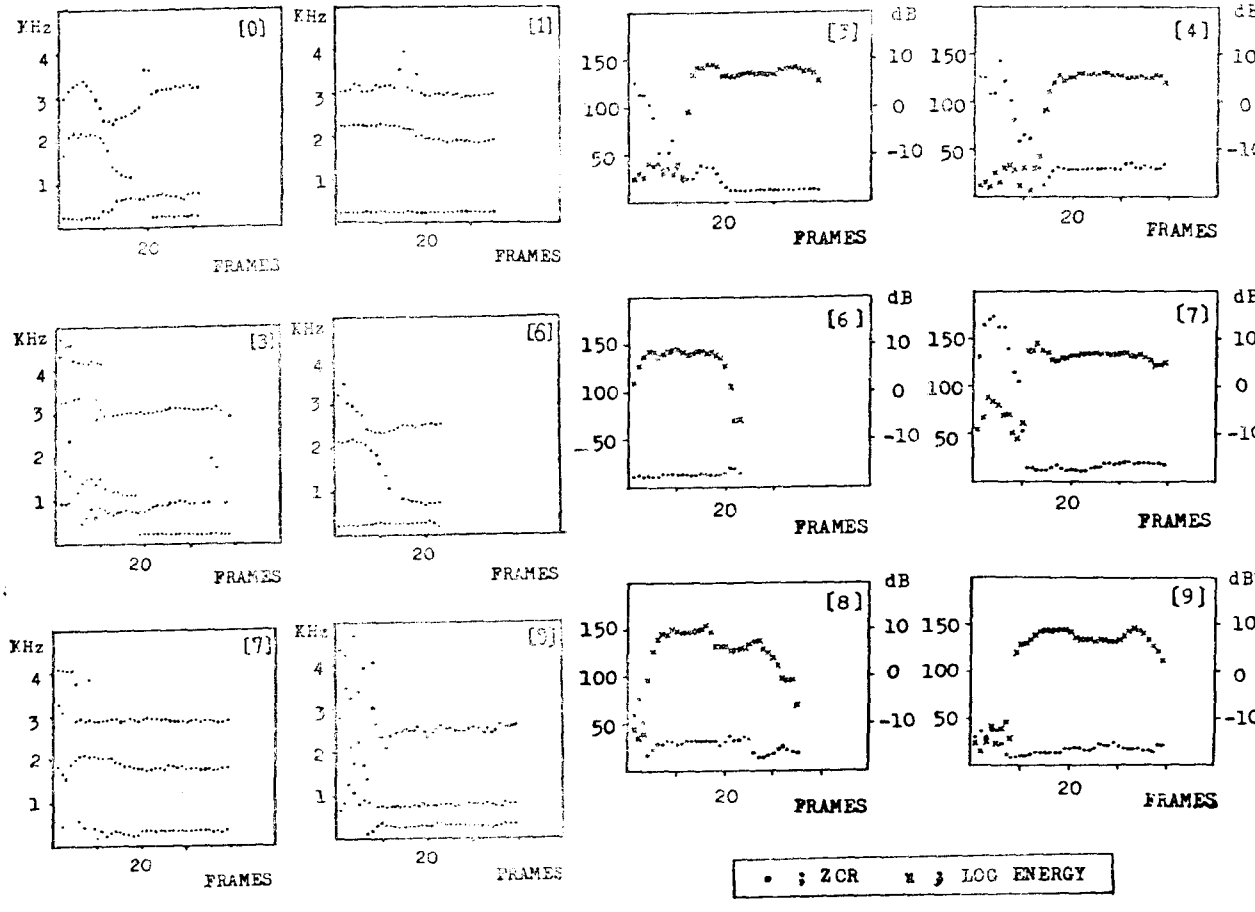


그림 6. 추정 formant 주파수의 예
Fig. 6. Example of estimated formant trajectories.

그림 7. 숫자음의 ZCR과 에너지
Fig. 7. Example of ZCR and energy of Korean digits.

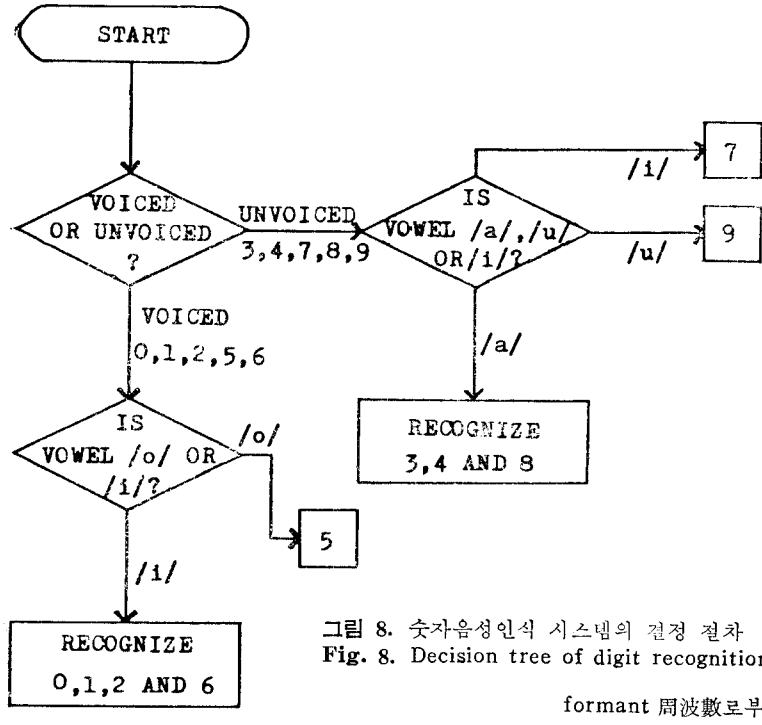


그림 8. 숫자음성인식 시스템의 결정 절차
Fig. 8. Decision tree of digit recognition system.

3.3 分類 및 認識

本 實驗의 認識 algorithm을 그림 8에 보인다.

(가) 有聲·無聲音의 分類

先頭部分의 無聲子音과 母音을 區別함에 있어서는 各 숫자 音聲의 1 frame 當 平均 ZCR ZCR을 計算해, threshold level $TH=1.3\times ZCR$ 보다 크고, log energy LE가 1보다 작은 frame을 無聲子音 區間으로 判定했다. 숫자音聲은 摩擦音(fricatives) ㅊ(7), ㅅ(3,4)과 破裂音(explosives) ㅣ(8), (9) 이 先頭部分에만 나타나고 있다. "육"의 'ㄱ'은 無音(voiceless)部로서 나타나므로 有聲·無聲音의 分類는 語頭만으로 足했다. 本 實驗에서는 위의 條件을 滿足하는 區間이 3~9 frame이었으며, 2 frame(38.4 ms) 以上 持續 되었을 때, 無聲子音이 認識 된 것으로 했다.

(나) 母音·有聲子音의 認識

2.1.1에 의해 推定한 第 1, 第 2, 第 3 formant 周波數 $\hat{F}_1$, $\hat{F}_2$, $\hat{F}_3$와 標準패턴 F_1, F_2, F_3와의 city-block distance D_j를 最小로 하는 音素를 認識音素로 했다.

$$D_j=\sum_{i=1}^{3}|F_{ij}-\hat{F}_{ij}| \quad (j=1,\cdots\cdots,8) \qquad (16)$$

本 實驗에서 準備한 母音 및 有聲子音의 標準 패턴은 아, 어, 오, 우, 이, ㄹ, ㅁ, ㅇ의 8種類다. 標準 formant 패턴은 本 實驗에 使用한 data로부터, 音素別로 가장 定常(stationary)이라 생각되는 部分의 formant 周波數로부터 作成했다. 誤認識을 줄이고 또한 安定된 音素만을 抽出하기 위해, 같은 有聲音素가 3 frame(51.2ms) 以上 나타날 때에 限해 最終的으로 그 音素가 存在하는 것으로 判定했다.

0 ; (iiiiiiiiiiiəoəəəiiiliŋŋŋŋŋŋŋŋŋŋŋŋ)
1 ; (iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii)
2 ; (iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii)
3 ; (ccccccccccciaaaaaammmmmmmmmmmm mimmiimmi)
4 ; (ccccccccccccimiaaaaaaaaaaaaaaaaaaaaaaa)
5 ; (unoooooooooooooooooooooooooooooooo)
6 ; (iiiiiiiiilluuiiuuuuuuuuu)
7 ; (cccccccccciiiiiiiiiiiiiiiiiiiiiiiiiiii)
8 ; (ccciaaaaaaaaaaaaaaaaIIIIIIIIIIIIIIIIII)
9 ; (ccccccciiouuuuuuuuuuuuuuuuuuuuuuuuuuu)

(1)

0 ; iəŋ	5 ; 0
1 ; il	6 ; iu
2 ; i	7 ; cil
3 ; cam	8 ; cal
4 ; ca	9 ; cu

(2)

그림 9. 인식된 음소열(1) 및 최종 인식결과(2)
Fig. 9. Recognized phoneme strings(1) and final results of recognition (2).

以上의 認識手法에 의한 숫자음의 音素認識(또는

分類)結果 및 最終的으로 認識된 音素列(phoneme string)을 그림 9에 보인다. 여기서는 便宜上 다음과 같은 音素記號로 表示했다.

/ㄱ, ㅅ, ㅊ, ㅖ/→c, /아/→a,
/어/→ə,/오/→o, /우/→u
/ㄹ/→l, /ㅁ/→m, /ㅇ/→ŋ

4. 結論 및 檢討

本 論文에서는, 國語 限定語 音聲自動認識 시스템의 一例로서, 숫자音聲認識에 關해 檢討했다. 成人 男性 한 사람의 音聲에 對한 認識實驗의 結果, 線形豫測에 의한 formant 周波數 推定時의 一部 peak의 脫落이나 附加에 의한 音素認識 段階에서의 局所的인 認識을 除外하고는 語頭의 子音 分類를 包含해 全體的인 숫자音聲認識이 이루어져, 複數話者등을 對象으로 한 認識시스템의 擴張등에 밝은 資料를 주고 있다. 局所的 音素誤認識은 豫測次數 M의 增加, 音聲學的 知識을 利用한 formant trajectory의 正確한 推定, 새로운 類似度(similarity)에 의한 音素認識등에 의해서도 解決할 수 있다고 보나, 國語 音聲學 全般에 걸친 充分한 基礎 確究가 必要하다. 그러나 숫자音聲은 比較的 적은 音素에 의한 單純한 構造이므로, 認識시스템의 改善등에 의해, 보다 效果的인 認識手法을 期待 할 수 있을 것이다.

앞으로, 多數話者 숫자音聲自動認識 시스템을 目標로 한 硏究와 아울러, 國語의 音聲學的 諸性質에 關해서도 檢討해 나갈 豫定이다.

References

1. Special Issue on MAN-MACHINE COMMUNICATION BY VOICE, Proc. of IEEE, Apr. 1976.
2. Kim, C-W; Cineradiographic study of Korean stops and a note on "Aspiration", Quaterly progress report, research Lab. of Electronics, MIT, pp. 259~272, 1967.
3. ——; On the autonomy of tensity in stop classification, Word, 21, No.3, Dec.1965.
4. Umeda, H. and Umeda, N.; A coustical features of Korean "forced" consonants, language research, 48, pp.23~33, 1965. (in Japanese)
5. Kim, B., Fujisaki, H. and Sawashima, M.; observation of jaw, tongue and lip control in articulation of Korean vowels, Tech. Report on Speech of Acous. Soc. of Japan, S74~56, 1975. (in Japanese)
6. Fujisaki, H. and Kim, B.; Articulatory description of the Korean vowel system, Ann. Report of the Eng. Res. Inst., Univ. of Tokyo, Vol.32, pp. 219~226, 1974.
7. ——; Analysis and recognition of Korean vowels, Univ. of Tokyo, Vol.32, pp.227~232, 1974.
8. Proc. IEEE pp.493~495, Apr. 1976.
9. Sambur, M.R. and Rabiner, L.R.; A speaker-independent digit-recognition system, the bell system technical journal, Vol. 54, No.1, pp.81~102, 1975.
10. Markel, M.D. and Gray, Jr., A.H.; Linear preiction of speech, Springer-Verlag, N.Y., 1976.
11. Agui, T. and Oh, Y.; Analysis of acoustical parameters in vowels, Ann. convention of Inst. of electrical eng. of Japan, Vol.10, p. 1086, 1978. (in Japanese)
12. 藤村; 音聲科學, 東京大學出版會, p.212, 1972.