音聲適應 區間分割 멀티섹션 코드북을 利用한 孤立單語認識

°하경민, 조정호, 홍재근, 김수중
경북대학교 전자공학과

Isolated-Word Recognition Using Adaptively Partitioned Multisection Codebooks

Kyeong Min Ha, Jeong Ho Jo, Jae Kuen Hong, Soo Joong Kim
Dept. of Electronics Kyungpook National University

Abstract

An isolated-word recognition method using adaptively partitioned multisection codebooks is proposed. Each training utterance was divided into several sections according to its pattern extracted by labeling technique. For each pattern, reference codebooks were generated by clustering the training vectors of the same section. In recognition procedure, input speech was divided into the sections by the same method used in codebook generation procedure, and recognized to the reference word whose codebook represented the smallest average distortion.

The proposed method was tested for 100 Korean words and attained recognition rate about 96 percent.

Ⅰ. 序 論

지금까지의 音聲認識 시스템에서는 基準音聲과 入力音聲間의 時間整合을 위하여 DTW(dynamic time warping) 알고리즘[1]을 많이 사용하였다. 그러나 DTW는 認識率은 높으나 計算量이 많다는 단점이 있다. 1983년 Shore[2]등은 기준음성의 特徵벡터를 時間에 무관하게 코드북을 구성하는 벡터量子化 音聲認識器를 제안하였다. 그러나 이와 같은 인식방법은 어떤 기준음성이 다른 기준음성의 특징을 모두 포함하게 되면 誤認識할 가능성이 커진다. 1985년 Burton[3]등은 입력음성과 기준음성간의 대략적인 時間整合을 위하여 음성을 同一한 길이의 섹션으로 나누어 섹션별로 類似度를 비교하는 멀티섹션 코드북을 이용한 音聲認識器를 제안하였다. 그러나 同一한 길이로 섹션을 分割하는 방법은 기준음성 相互間의 길이 차이가 크면 認識率이 낮아진다.

本 硏究에서는 音聲의 音素構成과 類似하게 섹션을 分割하는 멀티섹션 코드북을 構成하여 音聲을 認識하는 方法을 提案하였다.

Ⅱ. 벡터量子化에 依한 音聲認識

1. 벡터量子化

음성인식에서의 벡터양자화는 입력된 음성의 특징벡터를 미리 저장해둔 특징벡터 중에서 가장 잘 整合되는 하나의 벡터로 寫像시켜 주는 것이다. 벡터양자화 코드북을 이용한 기본적인 음성인식 시스템의 구조도를 그림1에 나타내었다.

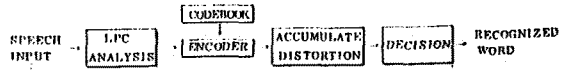

그림1. 벡터양자화를 이용한 음성인식 시스템
Fig. 1 The recognition system using vector quantization.

입력음성의 認識過程은 다음 3단계로 나눌 수 있다.

(1) 각 기준단어에 대하여 N개의 코드워드를 갖는 코드북을 구성한다.

(2) 기준단어 코드북과 입력음성간의 平均歪曲率

$$D_A(S,C)=\frac{1}{L}\sum_{j=1}^{L}\{\min_{1\le i\le N} d(Sj,Ci)\} \quad (1)$$

을 구한다. 여기서

L : 입력음성의 프래임 수
Ci : i번 째 코드워드, i=1,2, …… N
Sj : j번 째 입력벡터, j=1,2, …… L 이다.

(3) 기준단어 각각에 대하여 식(1)의 平均歪曲率을 구한 후 최소의 平均歪曲率을 나타내는 기준단어를 認識된 단어로 한다.

2. 歪曲率尺度

벡터양자화기 코드북 構成에 시험벡터 상호간의 類似度

測定을 위한 歪曲率尺度로는 다음과 같은 Itakura-Saito 歪曲率尺度[4]가 많이 쓰인다.

$$d_{IS}(f(\theta),\hat{f}(\theta))=\int_{-\pi}^{\pi}\frac{d\theta}{2\pi}\left\{\frac{f(\theta)}{\hat{f}(\theta)}-\ln\frac{f(\theta)}{\hat{f}(\theta)}-1\right\} \quad (2)$$

$f(\theta),\hat{f}(\theta)$를 선형예측모델로 나타내면 식(2)는 다음과 같다.

$$d_{IS}(f(\theta),\hat{f}(\theta))=\frac{\alpha}{\sigma^2}+\ln\frac{\hat{\sigma}^2}{\sigma^2}-1 \quad (3)$$

여기서 σ^2은 入力音聲을 예측모델 $f(\theta)$로 모델링 하였을 때의 자승 誤差이고, $\hat{\sigma}^2$은 基準音聲을 예측모델 $\hat{f}(\theta)$로 모델링 하였을 때의 자승 誤差이며, α는 入力音聲을 $\hat{f}(\theta)$로 모델링 하였을 때의 자승 誤差로서 다음과 같다.

$$\alpha=\gamma_a(o)\gamma_x(o)+2\sum_{n=1}^{M}\gamma_a(n)\gamma_x(o) \quad (4)$$

여기서 $\gamma_x(n)$은 입력음성의 자기상관함수이고 $\gamma_a(n)$은 기준모델 선형예측계수의 자기상관함수이다. 음성은 利得에 敏感한 特性을 가지므로, 이러한 特性을 考慮하여 修整된 I-S 歪曲率尺度[4]가 제안되었다.

$$d_{GO}=\ln(\alpha)-\ln(\sigma^2) \quad (5)$$

$$d_{GN}=\frac{\alpha}{\sigma^2}-1 \quad (6)$$

d_{GO}는 利得最適化(gain optimized) I-S 歪曲率尺度이고, d_{GN}는 利得定規化(gain normalized) I-S 歪曲率尺度이다.

本 硏究에서는 코드북을 構成할 때는 d_{GN}를 인식과정에서는 d_{GO}를 歪曲率尺度로 사용하였다.

3. 코드북 構成

1) 코드북 構成過程

음성인식에 응용되는 벡터양자화기 코드북 구성에는 주로 LBG 알고리즘[5,6]이 이용된다. 分離(split)方法을 이용한 LBG 알고리즘은 아래와 같이 6단계로 수행된다.

단계1 : L개의 試驗벡터 $T=\{x_i;\ i=0,1,\cdots L-1\}$에 대하여 전체 코드워드 수 $N=2^{Rm}$과 平均歪曲率 문턱값 δ를 정한다. 여기서 Rm은 코드북의 비트 수이다.

단계2 : 초기 비트 수를 R=0으로 두고 최초의 코드북 C_0을 試驗벡터 T의 중심 $\bar{y}$로 정한다. 즉, $C_0=\bar{y}$ 로 둔다.

단계3 : 현재의 코드북 $C_R=\{y_i;\ i=0,1,\cdots 2^R\}$의 각 코드워드 y_i를 $y_i+\varepsilon$과 $y_i-\varepsilon$으로 分離시켜 코드워드 수가 2배가 되는 $C_{R+1}(o)=\{y_i;\ i=0,1,\cdots 2^{R+1}\}$을 만든다. 여기서 ε은 分離常數이다. 비트수 R을 1증가 시키고, 코드북 개선 횟수 m을 0으로 둔다.

단계4 : m번 개선된 $C_R(m)=\{y_i;\ i=0,1,\cdots 2^R\}$의 각 코드워드와의 歪曲率이 최소인 試驗벡터의 集合 $S_i=\{x:d(x,y_i)\le d(x,y_j)\ all\ j\}$과 이때의 歪曲率 $D_m=\sum_{i=0}^{L-1}\min_{y\in C_R(m)}d(x_i,y)$ 을 구한다.

단계5 : $(D_{m-1}-D_m)/D_m>\delta$이면 새로운 코드북 $C_R(m+1)=\{cent(S_i);\ i=0,1,\cdots 2^R\}$을 만들고 개선 횟수 m을 1증가 시키고 단계4로 되돌아간다.
$(D_{m-1}-D_m)/D_m<\delta$이면 단계6을 수행한다.

단계6 : $C_R=C_{R(m)}$, $D_{(R)}=D_m$로 둔다. 만약 R=Rm이면 완성된 코드북을 C=C_R로 하고 과정을 끝내고, 그렇지 않으면 단계3으로 되돌아간다.

本 硏究에서는 선형예측계수의 자기상관함수를 코드워드로 사용하였다.

2) 單一섹션 코드북

단일섹션 코드북을 이용한 音聲認識器는 벡터量子化를 應用한 音聲認識器의 기본적인 모델이라 할 수 있다. 기준단어 수를 V라 할 때, 각 기준단어별로 얻은 시험발음에 대하여 특징벡터를 추출하여 N개의 코드워드를 갖는 코드북을 만든다. 이와 같이 각 기준단어별로 만들어진 V개의 코드북을 기준단어 코드북으로 한다.

認識過程에서는 입력음성과 각 기준단어 코드북과의 平均歪曲率을 구하여 최소의 平均歪曲率을 나타내는 기준단어를 인식된 단어로 한다. 입력음성과 기준단어 코드북과의 平均歪曲率은 입력음성의 각 프래임과 기준단어 코드북간의 최소 歪曲率의 합을 입력음성의 프래임 수로 나누어 구한다.

3) 멀티섹션 코드북

단일섹션 코드북을 이용한 음성인식 방법은 어떤 기준음성이 다른 기준음성의 특징을 모두 포함하면 두 음성을 區分하기 어렵게 된다. 이와 같은 문제를 해결하기 위하여 각 기준단어 코드북에 時間情報를 포함시킨 멀티섹션 코드북[3]이 제안되었다.

하나의 기준단어를 발음한 여러 개의 시험발음을 J개의 같은 길이의 섹션으로 분할한 후 각 섹션별로 K개의 코드워드를 갖는 섹션 코드북을 만들어 시간적인 순서대로 나열한 코드북 列을 기준단어 코드북으로 한다.

認識過程에서는 각 섹션별로 구한 입력음성과 기준단어 코드북과의 歪曲率을 누적하여 平均歪曲率을 구한다. 입력음성과 V개의 기준단어와의 平均歪曲率을 구하여 최소의 平均歪曲率을 나타내는 기준단어를 인식된 단어로 한다.

4. 音聲適應 區間分割 코드북

기존의 멀티섹션 코드북을 이용한 음성인식에서는 기준단어를 音素構成에 무관하게 동일한 길이의 섹션으로 분할하여 인식하는 것이므로 음성의 時間情報가 정확히 반영되지 않는다. 이와 같은 문제를 해결하기 위해서는 기준단어의 섹션을 音聲의 音素構成과 類似하게 分割하는 것이 바람직

하다.

本 硏究에서는 統計的 方法에 의한 래이블링[7]을 이용하여 音聲의 音素構成과 類似하게 섹션을 分割하여 음성을 인식하는 방법을 제안하였다.

기준단어 코드북을 構成하기 위하여 시험음성의 각 프래임을 廣帶域 에너지(wide band energy), 零通過率(ZCR) 두 帶域通過필터의 출력 에너지 등을 이용하여 無聲子音領域, 遷移領域, 母音領域으로 분류한 후 같은 래이블이 연속된 부분을 區劃化하여 음성의 패턴을 형성시킨다. 그 후 區劃의 境界를 섹션分割의 境界로 하여 섹션을 分割한다. 이와 같이 분할된 각 섹션별로 코드북을 만들어 순서대로 나열한 섹션 코드북 열을 기준단어 코드북으로 한다. 그림 2는 기준단어 코드북의 구성을 나타낸 것이다. 여기서 L1은 無聲子音領域, L2는 遷移領域, L3은 母音領域을 나타낸다.

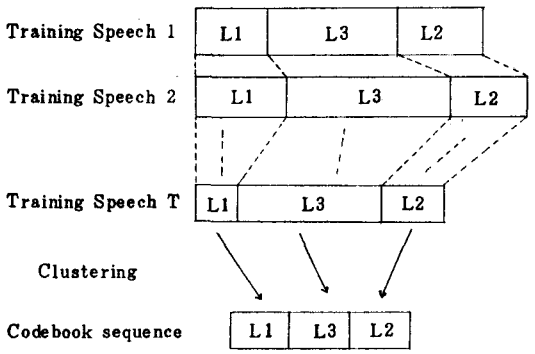

그림 2. 패턴 '132'의 기준 코드북 구성
Fig.2 The reference codebook of pattern '132'

인식과정에서는 입력음성에 대하여 패턴을 생성시키고 패턴에 일치하도록 섹션을 분할한 후 입력과 패턴이 같은 기준단어 코드북의 대응되는 섹션과의 왜곡률을 구하여 누적한다. 이와 같이 모든 섹션에 대하여 왜곡률을 구한 후 누적하여 平均歪曲率을 구한다. 입력음성과 각 기준단어 코드북과의 平均歪曲率을 구하여 최소의 平均歪曲率을 나타내는 기준단어를 인식된 단어로 한다. 그림 3은 인식과정의 구성도이다.

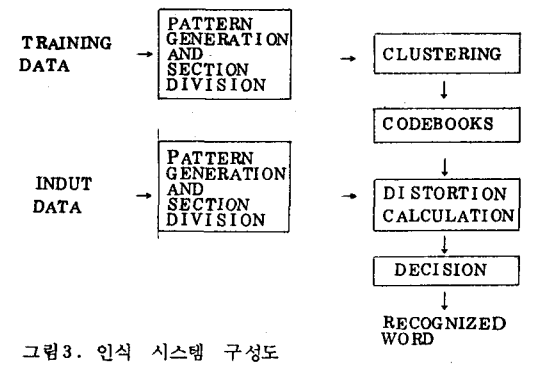

그림 3. 인식 시스템 구성도
Fig.3 The block diagram of recognition system.

Ⅲ. 實驗 및 結果

마이크를 통하여 입력된 음성을 180 H_3 — 3400 H_3로 帶域通過濾波시킨 후 10KHz로 표본화한 신호를 12비트로 量子化시켰다. 프래임 길이는 20 ms로 하고 Hamming 창함수를 사용하여 10차의 선형예측계수를 구하였다. 각 섹션당 코드워드 수는 無聲子音領域은 4개로 하고 遷移領域과 母音領域은 각각 8개로 하였다.

기준단어로는 長距離 自動電話 地域名 100개를 사용하였다. 기준단어 코드북 구성에는 하나의 기준단어당 8명의 성인 남성의 발음 15개와 10명의 성인 여성의 발음 15개를 사용하였다. 인식실험에서는 시험음성 발음에 참가한 남성 3명과 여성 3명의 200개의 발음과 시험음성 발음에 참가하지 않은 남성 3명과 여성 3명의 200개의 발음을 사용하였다. 표 1은 인식실험 결과이다.

Table 1. Recognition rate

		TEST 1	TEST 2
RECOGNITION RATE	GROUP 1	96 %	97.5 %
	GROUP 2	94 %	95 %

TEST 1; Multisection codebook of 4 sections.
TEST 2; Proposed method.
GROUP 1; Utterance in the training sequence.
GROUP 2; Utterance out of the training sequence.

Ⅳ. 結 論

本 硏究에서는 音聲의 섹션分割을 音聲의 音素構成과 類似하게 하여 각 섹션별로 類似度를 比較하는 音聲適應 區間分割 멀티섹션 코드북을 이용한 音聲認識方法을 제안하였다. 이 방법은 음성의 섹션분할을 음성의 음소구성과 유사하게 하므로 동일한 길이로 섹션을 분할하는 방법보다 認識率을 높일 수 있었다. 또한 입력음성과 동일한 패턴을 갖는 기준단어들만 비교하므로 비교되는 후보 기준단어수를 감소시켜 多語彙 認識시스템에도 적용할 수 있다.

제안된 방법을 長距離 自動電話 地域名 인식에 적용한 결과 약 96 %의 認識率을 얻었다.

참 고 문 헌

1. Itakura, F. "Minimum prediction residual principle applied to speech recognition", IEEE Trans. Acoustics, Speech, Signal Processing, Vol. pp. 67-72, 1975.
2. Shore, J.E. and Burton, D.K. "Discrete utterance Speech recognition without time alignment", IEEE

Trans. Inform. Theory, Vol. IT-29, pp. 473-491, 1983.

3. Burton, D.K. Shore, J.E. and Buck, J.T. "Isolated-word recognition using multisection vector quantization codebooks" IEEE, Trans. Acoustics, Speech, Signal Processing, Vol. ASSP-33, pp. 837-849, 1985.
4. Gray, R.M. Buzo, A. Gray, A.H, and Matsuyama, Y. "Distortion measures for speech processing" IEEE, Trans. Acoustics, Speech, Signal Processing, Vol. ASSP-28, pp. 367-376, 1980.
5. Linde, Y. Buzo, A. and Gray, R.M. "An algorithm for vector quantization" IEEE, Trans. Com., Vol. Corn-28, pp. 84-95, 1980.
6. Juang, B.H. Wong, D.y. and Gray, A.H. Jr. "Distortion performance of vector quantization for LPC voice coding" Trans. Acoustics, Speech, Signal Processing, Vol. ASSP-30, pp. 294-303. 1982.
7. Vysotsky, G.J. "A Speaker-independent discrete utterance recognition system, combining deterministic and probabilistic strategies" IEEE, Trans. Acoustics, Speech, Signal Processing, Vol. ASSP-32, pp. 489-499, 1984.