실험실 환경 음성을 이용한 전화음성 인식에 관한 연구

윤 상호°, 지 상문 , 오 영환
한국과학기술원 전산학과

Telephone Speech Recognition Using Laboratory Environment Speech Data

Sahng-Ho Yoon°, Sang-Mun Chi , Yung-Hwan Oh
Department of Computer Science, KAIST

요 약

본 연구에서는 전화선을 통한 음성 인식을 위해 저잡음의 실험실 환경에서 수집된 음성 자료를 이용하는 접근을 하였다. 전화 음성과 실험실 음성 간의 특성 차이를 보정하기 위해 선형 회귀 분석법을 이용한 SDCN (SNR-dependent Cepstral Normalization)을 제안하였다. 두자료간의 보정은 동시 녹음된 실험실 환경의 음성과 전화음성의 SNR에 따른 두 자료간의 차이를 최소화하는 변환행렬을 구해, 이를 학습자료의 변환에 이용한다.

제안된 방법의 타당성을 확인하기 위해 두가지 인식 알고리즘인 DTW와 이산HMM에 대해 실험하였다. DTW를 통한 인식에서, 개선된 SDCN에 의한 특징벡터의 변환은 기존의 SDCN에 따른 특징변환보다 8 ~ 17 %의 인식률이 향상되었다. 이산HMM으로 인식할 때는 개선된 SDCN에 의한 전화음성과 실험실음성과의 유사도를 보다 잘 나타내기 위해 개선된 SDCN을 적용하고, VQ 코드열 상에서의 코드 사상법을 사용하여 인식률의 향상시켰다.

1 서론

인간과 기계간의 의사소통 수단으로 음성인식이 실용화되기 위해서는 환경변화나 배경잡음에 대한 적절한 대처가 필요하다. 특히 근래에 연구의 필요성이 증대되고 있는 전화 음성인식은 전송선의 특성에 따라 음성신호가 왜곡되므로 전송선의 특성을 고려하여야 한다. 전송선의 특성은 전송 대역폭의 제한 (300 ~ 3400 Hz)과 그에 따른 정보손실, 전송시 첨가되는 잡음, 시간에 따른 잡음의 특성변화 등이 있다.

또한, 전화 음성 인식의 어려움은 전송선에 따른 음성의 왜곡뿐만 아니라 대용량의 자료 수집이 용이하지 않다는 점이다. 따라서 본 논문에서는 전화 음성의 인식을 위해 학습 자료로 전화 음성을 이용하지 않고, 기존의 실험실 환경에서 수집된 자료를 이용하는 방법에 대해 연구하였다. 서로 다른 환경 하의 음성자료 간의 교차 인식이 가능할 경우, 기존의 수집된 대용량의 음성 자료를 활용하여 음성인식 시스템을 구현할 수 있으므로, 전송선 상의 자료수집과 시스템 구성에 따른 부가적인 시간과 비용을 절감할 수 있다. 이러한 교차인식은 전화를 통한 원격지 정보검색과 정보입력, 철도와 항공 티켓의 전화를 통한 무인예약 서비스와 국제전화 교환 등 많은 음성자료를 필요로 하는 응용분야에 효과적으로 적용할 수 있을 것이다.

본 논문에서는 실험실 환경에서 수집된 마이크 음성 자료를 사용하여 전화선 상의 음성을 인식하고자, 동시에 녹음된 마이크 음성과 전화음성 자료를 이용하여 특징벡터 영역에서 변환행렬을 구하고, 기존의 인식시스템의 학습자료를 변환하였다. 음향환경 적응(Environment Adaptation)의 방법으로 SNR값에 따라 동시 녹음된 자료로부터 캡스트럼 벡터의 평균차이를 구해 이를 차감하는 SDCN방법을 이용하였다 [4]. 본 논문에서는 마이크 학습자료를 전화음성의 특징을 갖게 변환하는 SDCN방법을 선형 회귀분석 방법을 사용하여 개선하고, 이를 적용하여 DTW와 이산HMM방법을 이용한 인식실험을 수행하였다. 이산HMM을 통한 인식은 양자화오류에 의한 인식률 하락을 줄이기 위해 코드 사상법을 적용하였다 [3].

2 전화음성의 전달 모델

전화선을 통해 전달되는 음성 $y(t)$는 발성음성 $x(t)$에 전화선의 전달함수 $h(t)$와의 convolution에 주변잡음 $n(t)$가 더해진다.

$$y(t) = x(t) * h(t) + n(t) \tag{1}$$

식 (1)는 파워 스펙트럼 상에서 본다면 식 (2)과 같이 표현된다. 식 (2)에서 $P_y(f)$는 주파수f에 따른 전송음성의 에너지이고 $H(f)$는 전화선 전송채널의 주파수 응답이다.

$$P_y(f) = P_x(f)|H(f)|^2 + P_n(f) \tag{2}$$

식 (2)을 log를 취한뒤 역FFT를 이용하여 캡스트럼 영역으로 사영하게 되면 식 (3)와 같다. 아래의 식 (3)에서 $y = IFFT\{\log P_y(f)\}$, $x = IFFT\{\log P_x(f)\}$, $q = IFFT\{\log|H(f)|^2\}$, $n = IFFT\{\log P_n(f)\}$ 일 때,

$$y = x + q + r(x, n, q) \quad (3)$$

$$r(x, n, q) = IFFT\{\log 1 + e^{DFT[n-x-q]}\} \quad (4)$$

전화선의 채널 특성은 한 번의 전화 통화 중에서는 크게 변하지 않으므로 q를 상수로 볼 수 있다. 식 (3)와 전화선의 채널 특성에 의해 전화선의 효과는 캡스트럼 영역에서 가산적으로 작용함을 알 수 있다. 다음에 제시되는 SDCN방법은 캡스트럼 영역에서 이루어지는 채널 특성에 의한 가산을 추정하고 이를 보상해준다.

3 개선된 SDCN의 개선과 코드열 변환

실험실의 마이크 음성자료는 SDCN등을 통하여 전화음성의 특징을 갖게 되고 이를 학습시켜 인식에 이용한다. 기존의 SDCN방법을 사용한 계수변환을 개선하여 교차 인식률의 향상을 얻고, 이산HMM을 인식에 효과적으로 적용하기 위해 코드 사상법을 이용한다.

3.1 개선된 SDCN을 이용한 자료 변환

SDCN (SNR-dependent coefficient normalization)은 SNR에 따른 인식계수 정규화이다. 이 방법은 동시 녹음된 두 음성자료를 가지고 있을 때 사용하는 방법으로, 한 종류의 자료가 다른 종류의 자료의 특성을 갖게 해주거나, 잡음이 섞인 음성을 깨끗한 음성으로 근사시키는데 사용될 수 있다.

SDCN은 보정계수 $q + r(x, n, q)$이 SNR에 따라 결정된다고 가정하고 이에 따른 보정을 한다. 보정계수를 얻기위해 평균 제곱 에러 최소화 (MMSE : minimize mean square error)를 사용한다. SNR에 따른 보정계수를 구하여 기준 음성특징을 목표 특징음성으로 다음과 같이 보정한다.

$$y' = x - w(SNR) \quad (5)$$

이 방법은 인식에서 사용되는 특징 영역인 캡스트럼 영역에서 직접 처리를 하게 된다. 근사된 벡터 y'는 기준음성 특징 벡터 x에서 보정벡터 w를 빼서 구해진다. 보정벡터는 동시 녹음된 두 음성자료에서, 각각의 캡스트럼 차수에 대해 SNR에 따른 평균차이로 구해진다. 식 (6)에서 N은 사용된 특징벡터의 갯수, $x_i[j]$와 $y_i[j]$는 i번째 특징벡터 x_i와 y_i의 j번째 원소이고, $w(j,k)$은 SNR값이 k일때, 특징벡터의 j번째 원소의 보정값이다. $\delta[i]$는 Kronecker의 델타이고 Δ_{SNR}은 SNR값을 20단계로 분할해준다 [2].

$$w(j,k) = \frac{\sum_{i=0}^{N-1}(x_i[j] - y_i[j])\delta[SNR_i - k\Delta_{SNR}]}{\sum_{i=0}^{N-1}\delta[SNR_i - k\Delta_{SNR}]} \quad (6)$$

SDCN방법은 선형 회귀분석 방법 (LMR: linear multiple regression)을 사용해 근사도를 개선할 수 있다. 개선된 방법에서는 각각의 SNR에 대응하는 변환행렬을 구해, 이를 적용하여 특징벡터를 변환시킨다. 식 (7)에서 X_i와 Y_i는 마이크 음성과 전화음성의 SNR값이 i인 캡스트럼 벡터들로 이루어진 $D \times N$ 행렬이다. D는 벡터의 차수이고 N는 해당되는 벡터의 갯수이다. 변환행렬 A_i와 B_i는 평균 제곱 에러인 $E(|Y' - Y|^2)$를 최소화하도록 LMR방법을 이용하여 구한 $D \times D$ 행렬과 $D \times 1$ 행렬이다.

$$Y_i \simeq A_i X_i + B_i \quad (7)$$

LMR방법은 목표 특징벡터를 추정함에 있어, 기존의 SDCN이 벡터의 원소가 같은 차수의 원소에만 의존하는 제약을 완화시켜 기준 특징벡터의 모든 차수의 원소를 고려한 것으로 목표 특징벡터를 더 잘 추정한다[표 1]. 보정벡터 $q + r(x, n, q)$는 식 (4)에서 캡스트럼 벡터 영역에서 변형된 $\log 1 + e^{DFT[n-s-q]}$에 IFFT 연산을 하므로, 캡스트럼의 모든 차수를 고려함이 보다 타당하다.

실험실에서 수집된 음성 특징계수는 [그림 1]과 같이 변환된다. 우선 동시 녹음된 자료를 이용하여 SDCN을 위한 변환행렬을 구한다. 그리고 구해진 변환행렬을 실험실 음성자료에 적용하여 변환 후 인식시스템을 학습시킨다.

전화음성의 주파수 대역은 300 ~ 3400Hz로 대역폭이 제한되어 있다. 이러한 주파수 대역의 차이는 상호 교차인식 때, 인식률을 낮추는 요인이다. 마이크 음성을 전화음성 주파수 대역으로 대역통과 필터 (BPF : band pass filter)를 적용하면 마이크 음성과 전화음성이 비슷한 대역특성을 갖게 되어, 교차인식시 인식률의 향상을 기대할 수 있다.

SDCN을 적용하여 마이크 음성에 전화음성의 특징을 갖게할 때, 먼저 마이크 음성을 전화음성 주파수 대역으로 대역통과 필터를 적용하고, SDCN을 이용해 특징계수 변환을 적용하였다.

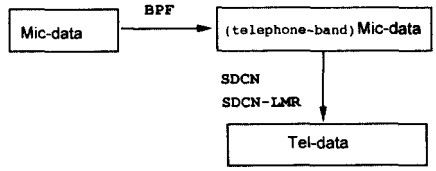

그림 1: 자료 변환 과정

3.2 코드열 변환

기존의 SDCN방법과 개선된 SDCN방법을 통한 마이크 음성의 변환된 특징계수는 전화음성의 특징계수에 보다 가까와진다[표 1]. 그러나 [표 2]에서 전화 음성과의 평균거리는 벡터양자화한 K-근접 코드에서 양자화 에러보다 크기 때문에 SDCN방법을 통해 변환된 자료를 코드화하게 되면, 전화음성을 코드화한 것과 다르게 나타난다. 이러한 코드의 불일치는 이산HMM과 같이 양자화가 필요한 방법으로 인식할 경우 인식률의 저하를 가져온다.

코드북 사상방법[3]은 두 코드열 간의 관련성 정보를 구해 코드를 변환한다. 주로 화자적응에 많이 사용되는 방법으로 두 화자간

표 1: 전화음성과의 평균거리

방법	MIC	SDCN	SDCN-LMR
평균거리	0.76	0.48	0.32

SDCN : (전화대역)마이크 음성 + SDCN
SDCN-LMR :(전화대역)마이크 음성 + 개선된SDCN

표 2: K-근접 코드의 각각의 양자화에러

	top1	top2	top3	top4
평균거리	0.22	0.28	0.32	0.35

서로 다른 특징공간 간의 연관관계를 구해 인식시 사용한다. 본 논문에서는 SDCN방법을 통한 자료와 K-근접 코드열과 전화음성자료와 K-근접 코드열 사이에 코드북 사상방법을 적용하여 두 코드열의 불일치를 완화시켰다.

코드북 사상방법은 두 코드열의 연관성을 구하는 단계와 이 정보를 이용해 코드를 변환하는 단계로 나누어진다. 두 코드열의 연관성은 두 코드열 간의 정합의 히스토그램으로 부터 구한다. 코드의 변환하는 절차는 다음과 같다. 벡터 X에 가장 가까운 코드 $a1$, $a2$, $a3$와 연관된 코드 $b1$, $b2$, $b3$를 구한다. 그리고 벡터 X와 $a1$, $a2$, $a3$코드와의 거리에 반비례하는 가중 $w1$, $w2$, $w3$에 의해 $b1$, $b2$, $b3$와 해당되는 벡터를 가중 평균하여 예측되는 벡터 Y를 구한다. 그리고 이 벡터 Y를 양자화한다[그림 2].

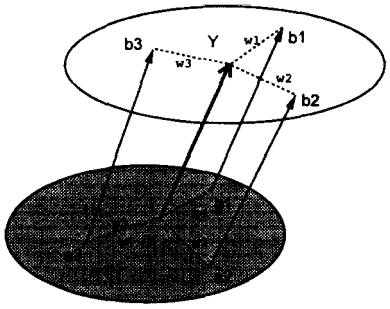

그림 2: 코드북 사상

4 실험 및 고찰

숫자음과 지역명에 대하여 개선된 SDCN방법의 유효성을 보이기 위해, 두가지 인식방법 이산HMM 방법과 DTW 방법을 이용하여 인식실험을 수행하였다. 마이크 음성은 개선된 SDCN을 적용하여 자료를 변환시킨 후 학습자료로 사용하였다. 이산HMM으로 인식할 때는 자료변환 후에 코드 사상방법을 더 적용하였다.

4.1 음성자료

본 연구에 필요한 전화음성 자료의 수집을 위해 마이크 음성과 전화음성을 동시에 수집할 수 있도록 그림과 같이 녹음환경을 구축하였다.

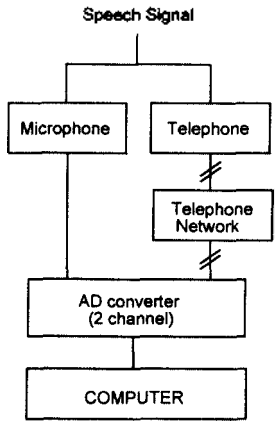

그림 3: 전화음성 데이타 수집 환경 개략도

실험에 사용한 음성자료로는 10개의 숫자음과 46개의 지역명이다. 10개의 숫자음은 35개의 4연속 숫자음을 녹음하여 각 숫자 별로 분할하였다. 학습자료로는 6명이 3회씩 발성한 것을 이용하였고, 인식자료로는 또다른 2명이 3회씩 발성한 음성자료를 이용하였다.

자료는 16kHz로 샘플링되고 16bit로 양자화되었다. 특징 파라미터 추출을 위해서, 분석구간은 256 sample, shift는 128 sample을 사용하였으며, 인식시 특징 파라미터로는 14차의 멜켑스트럼과 델타 멜켑스트럼, 정규화된 에너지, 그리고 델타 에너지를 이용하였다.

표 3: 숫자음 및 지역명 녹음자료

4연속 숫자음(35단어)	지역명 (46단어)
0287 5732 9601 4156	서울 대전 대구 부산 천안
1199 1398 6843 0712	여수 강릉 광주 목포 안동
5267 6633 2409 7954	동해 삼척 속초 태백 제천
1823 6378 8877 3510	구미 하양 진해 충무 나주
8065 2934 7489 2244	고양 예산 영덕 거제 이리
4621 9176 3045 8590	대천 인천 경주 원주 수원
5500 6972 5861 3649	포항 제주 창원 공주 김포
0316 7083 8194 9205	진주 남원 단양 해남 해남
1472 2538 4750	영암 울산 해남 완도 의정부 예 아니오

4.2 인식방법

인식방법으로는 DTW (dynamic time warping)방법와 이산HMM (hidden Markov Model)방법 두가지를 이용하였다. DTW는 시험패턴과 참조패턴을 시간축 상에서 비선형적으로 최적 정합하는 방법으로, 최적 정합에 의하여 거리가 가장 가까운 참조패턴으로 시험패턴을 인식한다. DTW에 의한 인식 실험에서는 6명

의 화자의 참조패턴을 단어별로 DTW에 의해 정합하여, DTW로 정합된 프레임별로 특징벡터를 평균한 참조패턴을 사용하였다.

이산HMM은 각 프레임을 K-근접 코드로 벡터양자화하는 방법을 사용하였다. 이 방법은 각각의 특징벡터마다 가장 가까운 K개의 코드를 고려하여 인식하는 방법으로 잡음이 존재할 때 코드의 민감성을 완화하여 보다 견고한 인식성능을 갖는다. 이때 K개의 코드의 가중은 해당되는 프레임과의 거리에 반비례한다. 실험에 사용한 k의 값은 4이다 [6].

4.3 실험결과

DTW방법을 이용한 인식에서는 개선된 SDCN방법을 사용하였을 때 지역명은 87.6%의 인식률을 보였다. 이는 기준패턴을 전화음성으로 하였을 경우의 95.3%보다는 낮은 인식률이지만 마이크 음성 자료를 사용하였을 때 보다는 57.5% 이상의 인식률 향상을 얻을 수 있다. 개선된 SDCN은 기존의 SDCN방법에 의한 기준패턴으로 인식한 것보다 17%의 향상을 보였다[표 4].

K-근접 이산HMM을 이용한 인식에서는 기존의 SDCN과 개선된 SDCN이 모두 51.9%와 57.1%의 낮은 인식률을 보였다. 이는 변환된 음성자료와 전화음성 자료의 평균거리가 약 0.32로 크기 때문에 두 자료의 코드가 많이 다르기 때문이다. 이를 보완하기 위해 코드 사상법을 적용하였을 경우, 인식률을 10%이상 올릴 수 있었다[표 5].

표 4: DTW방법을 이용한 인식

학습자료 시험자료	숫자음(10)	지역명(46)
MIC	26.4	30.1
TEL	70.7	95.3
SDCN	60.2	70.6
SDCN-LMR	68.1	87.6

MIC : 마이크 음성
TEL : 전화음성
SDCN : (전화대역)마이크 음성 + SDCN
SDCN-LMR : (전화대역)마이크 음성 + 개선된 SDCN

표 5: K-근접 이산HMM방법을 이용한 인식

학습자료 시험자료	숫자음	지역명
FMIC	31.0	13.6
TEL	80.2	85.9
SDCN	51.9	66.4
SDCN-LMR	57.1	62.9
SDCN-map	76.4	74.6
SDCN-LMR-map	72.6	79.7

FMIC : (전화대역)마이크 음성
SDCN-map : SDCN + 코드 사상법
SDCN-LMR-map: 개선된 SDCN + 코드 사상법

5 결론 및 검토

본 연구에서는 전화선을 통한 음성 인식을 위해 저잡음의 실험실 환경에서 수집된 음성 자료를 이용하는 접근을 하였다. 두 환경 하의 음성자료 간의 교차인식이 가능할 경우, 기존의 자료를 이용하게 되므로 새로운 자료수집과 시스템 구성에 따른 부가적인 시간과 비용을 절감할 수 있다.

본 논문에서는 전화음성의 교차인식을 위해 SDCN을 이용하여 음성인식 시스템의 학습자료를 변환하는 방법을 이용했다. 기존의 SDCN이 자료변환에 캡스트럼 벡타의 같은 차수의 원소에만 의존하는 제약을, 모든 차수의 원소를 고려하여 특징을 변환하는 SDCN 방법을 제안하고 이를 선형 회귀분석 방법을 이용하여 구현하였다.

DTW를 이용한 인식 실험을 통해 개선된 SDCN이 기존의 방법보다 8 ~ 17 % 좋은 성능을 나타내는 것을 알 수 있었다. 이는 제안한 방법의 유효함을 보여준다. 이산HMM을 이용한 인식에서는 양자화에러로 인해 교차 인식률이 낮았으나, 코드 사상법을 이용하여 성능 향상을 얻을 수 있었다.

차후의 연구로 선형 회귀분석을 비선형 회귀분석으로 전환하면 보다 높은 성능을 얻을 수 있을 것으로 판단된다.

참고 문헌

[1] 윤상호, 오영환, "전화음성 인식을 위한 특징추출 방법의 비교연구", 정보과학회 학술발표논문집 제21권 1호, pp. 279-282, 1994

[2] Alejandro Acero, "Acoustical and environmental robustness in automatic speech recognitionn", Kluwer Academic Publishers, 1993

[3] H. Hattori, S. Sagayama, "Speaker Adaptation Based on Vector Field Smoothing" IEICE trans. Inf. & Syst., vol E76, pp 227-234, Feb. 1993

[4] Alejandro Acero, Richard M. Stern, "Environmental robustness in automatic speech recognition", ICASSP-90, pp. 849-852, 1990

[6] H. P. Tseng, M. J. Sabin, and E. A. Lee, "Fuzzy vector quantization applied to hidden Markov modeling", *Proc. ICASSP '87*, pp. 641 - 643,1987

[7] C. Mokbel, G. Chollet, "Word Recognition in The Car", Proc. ICASSP-91, pp. 925-928. 1991