연속 숫자음 전화음성의 인식 성능 향상에 관한 연구

김민성, 정성윤, 손종목, 배건성
경북대학교 전자·전기공학부

A Study on the Performance Improvement of Connected Digit Telephone Speech Recognition

Min Sung Kim, Sung Yun Jung, Jong Mok Son, Keun Sung Bae
School of Electronics and Electrical Engineering, Kyungpook National University
kmslove@mir.knu.ac.kr, ksbae@ee.knu.ac.kr

요 약

전화음성의 경우 전화 회선의 채널 대역폭 제한과 통화로 형성시 달라지는 채널의 특성으로 인하여 마이크 음성에 비하여 인식 성능이 많이 저하된다. 본 연구에서는 연속 숫자음 전화음성의 인식율 향상을 위해 채널 왜곡 보상 기법들을 적용하고, HTK 기반의 인식 실험을 통해 보상 기법에 따른 인식 성능을 비교하였다. 채널 왜곡 보상 기법으로 CMN, RASTA, RTCN 등을 적용하고, 각 보상 기법에 따라 HMM의 state 수, mixture 수를 바꾸어 가며 인식 실험한 결과를 제시한다.

I. 서 론

전화음성의 연속 숫자음 인식은 증권 거래를 비롯한 금융 거래와 음성을 이용한 사용자 비밀번호 인증에 이용될 수 있으므로 이에 대한 연구가 꾸준히 이루어지고 있다. 전화음성의 경우 대역폭 제한과 채널의 영향으로 인한 왜곡으로 연속 숫자음 인식을 할 때, 마이크 음성에 비해 낮은 인식율을 나타내며, 이에 따라 인식율 향상을 위한 다양한 보상 기법이 연구 되어 왔다[1, 2].

본 연구에서는 전화음성 연속 숫자음의 인식율을 향상시키기 위한 기초연구로 MFCC (Mel Frequency Cepstral Coefficient)를 특징벡터로 하고 기존의 채널 보상 기법들을 적용하여 HMM 모델의 state 수와 mixture 수를 변화시키면서 인식율의 변화를 조사하였다. 이때, 인식 실험에 사용한 음성 데이터는 동일한 환경의 전화음성 데이터를 훈련과 인식 과정에 적용하였다. 또한 연속 숫자음 인식에서 주로 오인식 되는 숫자 "일"과 "이", "일"과 "칠", "오"와 "구"의 오인식율에 대해서 왜곡 보상 기법에 따른 변화를 살펴보았다.

본 논문의 구성은 다음과 같다. 우선, 2 장에서는 채널 왜곡 보상을 위해 적용한 기법에 대해 설명하고, 3 장에서는 인식 실험에 사용된 HMM 모델에 대해 언급한다. 4장에서는 실험내용과 인식결과에 대해 검토한 후, 5장에서 결론을 맺는다.

II. 전화 음성 왜곡 보상 기법

1. CMN (Cepstrum Mean Normalization)

전화음성은 전화 채널의 대역폭 제한과 통화로 형성시의 채널의 변이로 인해 왜곡이 생기며, 이는 음성신호에 convolution noise 형태로 나타나게 된다. 채널은 시간에 따라 그 특성이 급격히 변화하지 않으므로 음성 구간내에 각 음소가 고르게 분포한다고 가정하면 전체 음성에 대한 켑스트럼의 평균은 채널의 영향인 convolution noise를 나타내게 되고, 이때 평균을 제거함으로써 채널의 영향을 줄일 수 있다[2]. CMN 과정은 식 (1)과 같이 표현된다.

$$C(t)_{cmn} = C(t)_y - \frac{1}{T}\sum_{t=1}^{T} C(t)_y \qquad (1)$$

여기서, $C(t)_y$는 왜곡된 전화음성의 켑스트럼, T는 전체 전화음성의 길이, $C(t)_{cmn}$는 CMN 처리된 켑스트럼,

t는 현재 분석 프레임의 인덱스를 나타낸다. 본 연구에서는 4연 숫자음 단위로 켑스트럼의 평균을 구하여 CMN 기법을 적용해서 실험 하였다.

2. RASTA (RelAtive SpecTrAl)

RASTA 기법은 음성 신호에 비해 완만히 변하는 채널 왜곡 영향을 줄이기 위해 특징벡터 각 계수의 시간 궤적에 대해 필터링을 수행하는데, 이때 RASTA 필터의 전달함수는 식 (2)와 같다.

$$H(Z) = \frac{0.2 + 1.0z^{-1} - 0.1z^{-3} - 0.2z^{-4}}{1 - \alpha z^{-1}} \quad (2)$$

여기서 α 는 RASTA 필터의 주파수응답에서 저역차단 주파수를 결정하는 필터 계수로, 음성에 비해 느리게 변하는 채널 왜곡을 제거한다. 본 연구에서는 α 값으로 0.98을 사용해서 실험하였다[3].

3. RTCN (Real Time Cepstrum Normalization)

CMN의 경우 음성의 길이가 충분히 길어야만 바이어스 없는 켑스트럼을 얻을 수 있고, CMN의 여러가지 가정도 성립될 수 있다. 그러나 실시간 처리를 요하는 전화음성 인식에서는 이러한 조건을 만족시키기가 어렵고, 또한 음성 자체의 특성에 해당되는 짧은 음성 구간에서의 켑스트럼 평균을 채널 왜곡으로 오인할 수도 있다. 따라서, 짧은 음성 구간에서 구한 켑스트럼의 평균을 사용하여 충분한 길이를 갖는 음성 신호 전체의 켑스트럼 평균을 추정하게 되는데, 이는 식 (3)과 같다.

$$\bar{x}_t = \alpha x_t - (1-\alpha)\bar{x}_{t-1} \quad (3)$$

여기서, $\bar{x}_t$ 는 t번째 추정과정에서의 전체 켑스트럼 평균의 추정치이고, x_t 는 t번째 음성신호의 켑스트럼 평균이다. 왜곡된 음성의 켑스트럼에서 추정된 $\bar{x}_t$ 를 빼줌으로써 왜곡을 보상하는 방법이 RTCN이다. 본 연구에서는 α 로 1/8을 사용했다.

III. HMM (Hidden Markov Model)

각 음소의 음향학적 특성을 나타내기 위한 모델로 HTK 3.1 기반의 HMM 인식기를 사용하였다. HMM의 형태로는 CHMM(Continuous HMM)을 이용했고 각 state 마다 mixture 수를 1개에서 9개까지 늘려가면서 실험을 하였으며, state 수도 3개에서 5개까지 늘려가면서 실험을 하였다. 그리고 상태천이 모델로는 Left to Right를 이용했다[4]. 연속 숫자음 인식을 위하여 FSN (Finite State Network)을 구성하였으며, 4연 숫자를 발성할 때 나타나는 음운현상을 고려하기 위해서 cross word를 적용하고여 인식 및 훈련 과정에서 cross word에 따른 영향을 고려하였다[5].

한국어 숫자음의 기본 유사음소로 13개를 정하고, 4연 숫자음 구성시 발생하는 묵음 모델과 짧은 묵음 모델을 포함하여 총 15개의 기본 유사음소를 사용하였다. 4연 숫자음에서 나타나는 음운현상을 고려하기 위해 triphone을 사용하였으며, 이로 인한 훈련 파라미터의 증가를 막기 위해 tree-based state clustering을 이용해서 모델간 파라미터를 공유시켰다. 특히 4연 숫자음 발성시에 숫자들 사이에 나타나는 짧은 묵음 모델은 묵음 모델의 3번째 state와 파라미터를 공유시키고 1 state 모델로 구현하였다[4, 5].

IV. 실험 환경 및 결과

본 논문에서는 Dialogic사의 전화인터페이스 보드를 사용하여 시내 및 시외 전화음성을 8kHz, 8bits, μ-law로 녹음한 후 8kHz, 16bits PCM(Pulse Code Modulation)으로 바꾸어서 사용하였다. 실험에 사용된 음성은 남녀 각각 5명씩, 화자당 160개씩 두 번 발성한 총 3,200개의 4연 숫자음 음성을 사용했다. 훈련 데이터로는 남녀 각각 4명을, 테스트 데이터로는 남녀 각각 1명의 음성을 사용하였다. 이때 발성한 160개의 음성은 11개의 숫자음(영, 일, ... 구, 공)이 고르게 분포하도록 선택하였다.

특징 파라미터 추출과정은 다음과 같다. PCM으로 변환된 음성을 전처리과정(계수=0.97)을 거친 후에 분석구간별로 hamming window를 이용해서 windowing을 하였다. 이때, 분석구간의 길이는 20ms로 했으며, 분석구간의 이동은 10ms로 정했다. 각 분석구간별로 19개의 필터뱅크의 출력을 이용해서 12차의 MFCC를 구하고 log energy를 포함시켜 13차 특징벡터를 구했다. 이 특징벡터에 왜곡 보상 기법을 적용한 후 특징벡터의 Delta, Delta-Delta 특징벡터를 구한 후, log energy는 제외시키고

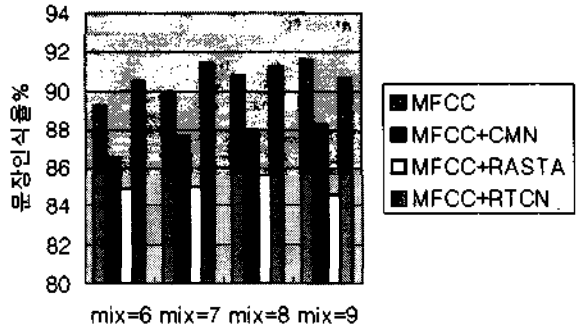
총 38차의 특징 벡터를 사용했다.

실험은 HMM 모델의 state 수와 state당 mixture수를 변화시켜 가면서 인식율을 관찰하는 방법으로 수행하였다. 또한 전화음성의 왜곡을 보상하는 기존의 방법으로 처리한 음성에 대해서도 같은 방법으로 실험하였다.

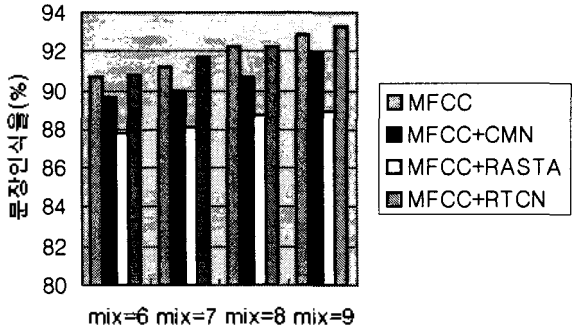
예비 실험에서, 음향모델로 triphone을 사용한 경우 문장 인식율, 즉, 4연 연속 숫자음의 인식율이 90%를 넘게 나왔는데 비해 같은 조건으로 monophone으로 인식 실험한 경우는 문장 인식율이 77% 정도에 그쳐 triphone 모델이 연속 숫자음에서 발생하는 음운현상을 잘 모델링하는 것으로 나타났다.

그림 1은 state 수와 mixture 수에 따른 인식결과를 보인 것이다. state 수는 3개에서 5개까지, mixture 수는 6개에서 9까지 정하여 실험을 하였는데, 인식율을 비교해 보면 MFCC 특징벡터에 RASTA 기법을 적용한 경우가 다른 방법에 비해 인식율이 다소 낮게 나타났음을 볼 수 있다. CMN의 경우도 2장 1절에서 설명한 바와 같이 짧은 4연 숫자음의 평균 켑스트럼을 구해서 cepstral normalization을 수행했기 때문에 인식율이 다소 낮아진 것이라 생각된다.

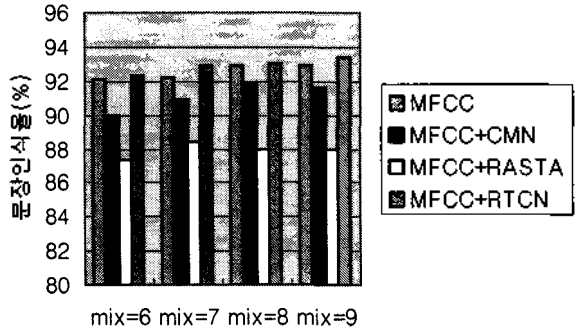
CMN의 단점을 보완하는 보상 기법인 RTCN을 적용한 경우는 다른 기법보다 높은 인식율을 보였다. state 수를 증가시킬 경우 인식율은 보상 기법에 관계없이 다소 증가하지만 인식율의 증가분은 감소하므로 연산량과 메모리량을 고려하여 적당한 state 수를 선택해 줄 필요가 있다. Mixture 수를 증가시키는 경우 보상 기법에 상관없이 인식율은 증가했지만 mxiture당 훈련데이터가 적게 되어 훈련이 잘못된 경우 인식율이 떨어지는 경우도 있었다.



(a) state 3개일 경우



(b) state 4개일 경우.



(c) state 5개일 경우

그림 1. state 수, mixture 수에 따른 문장인식율 비교

연속 숫자음 인식에서 주로 에러를 발생시키는 숫자음은 종성의 'ㄹ'의 유무를 판단하지 못해서 생기는 "일"과 "이"의 오인식과 초성의 'ㅊ'의 유무를 판단하지 못해 생기는 "일"과 "칠"의 오인식 그리고 연속 숫자 사이의 "오"를 "구"로 오인식하는 것이므로 confusion matrix에서 해당되는 경우만을 표 1 및 표 2에 나타내었다. 표 1과 표 2는 각각 state 3개, state 5개인 경우이며, mixture 8개일 때의 confusion matrix의 일부를 나타낸 것이다. 표를 보면 "일"과 "이", "일"과 "칠", "오"와 "구"의 오인식율이 낮아지면 전체 인식율이 상당히 증가할 수 있음을 알 수 있다. 예를 들어, 표 1의 RTCN 보상 기법을 적용한 경우의 오인식율과 그림 1(a)의 인식율을 비교해 보면 인식율이 높은 RTCN 보상 기법을 적용한 경우가 다른 보상 기법보다 오인식율이 낮게 나타났음을 확인할 수 있다.

또한 표 1와 표 2를 보면 숫자 "이"를 "일"로 인식하는 경우가 state 5개일 때 state 3개일 때 보다 줄어 들었음을 알 수 있고, 그림 1의 (a), (c)를 비교하면 인식율이 state 5개일 때 더 높게 나왔다. 같은 state 수를 갖는

모델에서 보면 전체 인식율인 높은 보상기법은 혼돈되는 숫자의 오인식율이 낮음을 보여 주고 있다.

표 1. state 3개 일 때 confusion matrix의 일부.

	일->이	이->일	오->구	일->칠
MFCC	6/252	7/228	3/208	8/212
MFCC+CMN	1/252	22/228	6/208	5/212
MFCC+RASTA	4/252	19/228	7/208	4/212
MFCC+RTCN	0/252	14/228	6/208	6/212

표 2. state 5개일 때 confusion matrix의 일부.

	일->이	이->일	오->구	일->칠
MFCC	10/252	6/228	4/208	7/212
MFCC+CMN	3/252	16/228	6/208	7/212
MFCC+RASTA	17/252	10/228	3/208	3/212
MFCC+RTCN	2/252	8/228	4/208	9/212

V. 결론

본 논문에서는 연속 숫자음 전화음성 인식에서 기존의 왜곡 보상 기법을 적용해 인식 실험을 해 보았다. 유사 음소모델로는 triphone을 사용하고 tree-based state clustering을 적용하여 모델간 파라미터를 공유시켰다. 또한 state와 mixture 수를 다르게 하여 실험을 하였으며 채널의 왜곡을 보상하기 위해 보상 기법들을 적용해 실험을 하였다. 실험 결과 보상 기법 중에서 RTCN 보상 기법이 가장 좋은 성능을 보였으며, state수를 증가시키고 mixture수를 증가시키면 인식율은 증가하지만 증가분은 감소하여 state 6개이고 mixture 9개인 경우, RTCN 보상 기법을 적용한 경우 93.44%의 문장 인식율을 나타냈다.

채널 보상 기법중 인식율과 오인식율을 비교해 보면 인식율이 높은 기법은 오인식율이 낮으므로 인식율을 향상시키기 위해서는 오인식 되는 "일"과 "이", "일"과 "칠", "오"와 "구"를 구분할 수 있는 모델이나 보상 기법을 찾는 것이 해결책임을 알 수 있다.

추후 특징추출 부분에서의 새로운 보상기법과 LDA와 같은 기법을 적용하고 HMM 모델상에서의 보상 기법을 적용하여 인식 성능을 향상시키고자 한다.

본 연구는 한국전자통신연구원 네트워크기술연구소 음성정보연구센터의 연구비 지원으로 수행되었으며, 지원에 감사드립니다.

참고 문헌

[1] P.J. Moreno, "Speech Recognition in Telephone Environment," MS. Thesis, CMU

[2] J. D. Veth and L. Boves, " Comparision of channel normalization technique for automatic speech recognition over the phone," Proc. ICSLP, pp.2332-2335, 1996

[3] H. Hermansky and N. Morgan, " RASTA Processing of speech," IEEE Trans. Speech Audio Processing, Vol.2, No.4, pp.578-589, 1994

[4] L. Rabiner and B. H. Juang, "Fundamentals of speech recognition," Prentice Hall International, Inc. 1993

[5] Steve Young and Gunnar Evernann, " The HTK book (For Version 3.1)," Cambridge University Engineering Department , 2001

[6] 김상진, 서영주, 한민수 "LCMS를 이용한 한국어 연속 숫자인식에 관한 연구," 한국음향학회 학술발표대회 논문집 제20권, pp. 43, 2001

[7] 박성준, 김재인, 전주식, "한국어 연속 숫자음 인식 연구," 음성통신 및 신호처리 학술대회, pp. 179, 2000

[8] J.Zhao amd A. Ganapathiraju, "Decision Tree-Based State Tying for Acoustic Modeling," Department of Electrical and Computer Engineering Mississippi State University.