채널보상기법 및 특징파라미터에 따른 한국어 연속숫자음 전화음성의 인식성능 비교

정 성 윤, 김 민 성, 손 종 목, 배 건 성, 김 상 훈 *
경북대학교 전자공학과, 한국전자통신연구원 *

Comparison of the recognition performance of Korean connected digit telephone speech depending on channel compensation methods and feature parameters

Sung Yun Jung, Min Sung Kim, Jong Mok Son, Keun Sung Bae, Sang Hun Kim *
School of Electronic and Electrical Engineering, Kyungpook National University
Electronics Telecommunications Research Institute *
E-mail : yunij@mir.knu.ac.kr

Abstract

As a preliminary study for improving recognition performance of the connected digit telephone speech, we investigate feature parameters as well as channel compensation methods of telephone speech. The CMN and RTCN are examined for telephone channel compensation, and the MFCC, DWFBA, SSC and their delta-features are examined as feature parameters. Recognition experiments with database we collected show that in feature level DWFBA is better than MFCC and for channel compensation RTCN is better than CMN. The DWFBA+Delta_Mel-SSC feature shows the highest recognition rate.

I. 서론

전화음성의 인식률은 전화망 환경에서 수반되는 신호의 왜곡 및 잡음으로 인해 일반 마이크 음성의 인식률에 비해 아직 만족스럽지 못한 수준이며, 특히, 한국어 연속숫자음의 경우 다양한 조음효과로 인해 인식에 어려움이 많다. 전화망을 통한 은행계좌번호, 주민등록번호 등의 조회에 음성인식 기술을 활용하기 위해서는 전화음성의 채널왜곡 및 잡음의 영향을 최소화하여 한국어 연속숫자음의 인식률을 향상시킬 수 있는 기법에 대한 연구가 선행되어야 한다.

본 연구에서는 유/무선 전화망 환경에서 한국어 연속숫자음의 인식성능을 개선하기 위한 선행연구로 특징파라미터 및 보상방법에 따른 연속숫자음 전화음성의 인식실험을 수행하고 인식률을 비교하였다. 이를 위해 소량의 4연속숫자음의 전화음성을 자체 수집/분석하였으며, 기존의 채널보상기법 중에 비교적 적은 변이를 보인 CMN(Cepstral Mean Normalization) 및 RTCN(Real-Time Cepstral Normalization)을 보상기법으로 선택하였다[1]. 인식률을 향상시키기 위해서 특징파라미터로 기존의 MFCC(Mel Frequency Cepstrum Coefficient)에 포만트 성분과 비슷한 특성을 나타내는 SSC(Subband Spectral Centroid)[2]라는 보조 특징파라미터를 추가하였으며, 필터뱅크 출력에 가중치 함수를 줌으로써 상대적으로 높은 에너지를 갖는 스펙트럼을 강조시키는 DWFBA(Direct Weighted Filter Bank Analysis)기반의 특징파라미터를 사용하였다[3].

전술한 특징파라미터 추출방법과 채널보상방법에 따른 인식실험을 위해 HTK(Hidden markov model Tool Kit) V3.1 기반의 연속숫자음 인식시스템을 구현하였다.

자체 수집한 전화음성 DB를 이용한 인식실험 결과, 기존의 MFCC에 비해 새로 도입한 특징파라미터인 DWFBA가 0.73%의 인식률 증가를 나타내었고, 여기에 보조 특징파라미터로 Delta_Mel-SSC를 추가로 사용한 경우에는 2.92%의 인식률 증가를 더 얻을 수 있었다. 또한 적용한 채널보상기법 중에는 RTCN이 CMN에 비해 대체로 나은 인식률 증가를 보였었다. 최종적으로, DWFBA에 보조특징파라미터로 Delta_Mel-SSC를 추가하고 RTCN을 채널보상기법으로 적용한 인식실험의 결과는 기존의 MFCC만 적용한 경우에 비해서는 8.85%, CMN과 함께 적용한 경우에 비해 1.77%의 인식률 증가를 나타내었다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 소용량의 유/무선 전화음성 DB를 자체 수집한 과정을 기술하고, 3장에서는 본 논문에서 적용한 채널보상기법 및 특징파라미터 추출 방법들을 기술한다. 그리고, 4장에서는 Baseline 인식기를 기반으로 특징파라미터 및 보상기법에 따른 인식실험 결과를 검토한 후, 5장에서 결론을 맺는다.

II. 전화음성 DB 수집

분석 및 인식 실험에 사용하기 위한 연속숫자음의 전화음성 수집을 위해 한국전자통신연구원(Electronics Telecommunications Research Institute:ETRI)에서 제공한 1000개의 4연속숫자음 목록 중에서 160개를 표 1에 주어진 것과 같이 영과 공을 포함하여 다양한 숫자음이 고르게 분포되도록 선정하였다. 전화음성은 매 통화시 변경되는 전화망의 경로에 따라 채널 특성이 변화하면서 음성신호를 왜곡시키는데, 이러한 특성을 반영하기 위해 한 통화당 8개의 4연속숫자음을 설정하여 모두 20 통화분을 준비하였다.

표 1 . 선정된 4연속숫자음의 예

칠일공육	오공이육	이공이공
이영오이	이사육공	육삼구일
칠팔이이	일오팔육	.....

전화음성의 녹음은 Dialogic 사의 전화 인터페이스 카드(모델명:D/41EPCI)를 사용하여 PC에서 자동으로 녹음할 수 있도록 시스템을 구현하였다. 전화음성은 8kHz 샘플링에 μ-law 방식으로 녹음되는데, 녹음이 이루어진 시간을 기준으로 자동으로 파일이름이 결정되어 저장된다. 8kHz, μ-law 형식으로 저장된 음성파일은 나중에 μ-law expanding을 통해 8kHz, 16-bit Linear PCM(Pulse Code Modulation)으로 변환되어 분석 및 인식용 파일로 저장된다.

160 종류의 4연속숫자음에 대해, 연구실에서 9명(남자 5명, 여자4명)의 화자가 유/무선전화를 통해 1회 발성한 총 1140개의 연속숫자음을 녹음하였다. 그리고, 음소별 변이 분석을 위해, 수작업으로 음소 레이블링을 수행하였다.

III. 채널보상 및 특징파라미터 추출

3.1 채널보상기법

(1) CMN

CMN의 기본 개념은 시간영역에서 컨벌루션 형태로 나타나는 채널특성이 켑스트럼 영역에서 합의 형태로 표현되므로 채널의 변이에 해당되는 켑스트럼의 평균값을 제거하는 것이다. 채널특성은 짧은 시간에 큰 변화가 생기지 않고 거의 일정하게 나타나기 때문에, 켑스트럼 영역에서는 전체 켑스트럼의 바이어스 성분으로 볼 수 있다. 즉, 음성신호가 임의의 채널을 통해 녹음 되었을 때 켑스트럼 도메인에서는 채널특성이 음성신호의 켑스트럼에 합해진 형태로 나타나기 때문에, 켑스트럼의 바이어스(평균값)를 제거해주는 것만으로도 채널왜곡으로 인한 인식성능 감소를 상당히 줄일 수 있게 된다.

$C=[c_1, c_2, c_3, \ldots.]$ 가 MFCC 벡터이고, $C_{t,i}$ 를 t번째 4연속숫자음의 음성신호에서 i 번째 프레임의 MFCC 벡터, X_t 를 t번째 음성신호의 모든 프레임에 대한 MFCC 평균벡터라고 하면, CMN은 식 (1)과 같이 구해진다.

$$C_{t,i}' = C_{t,i} - X_t \qquad (1)$$

(2) RTCN

CMN은 계산량이 매우 작으면서 그 왜곡보상 능력은 큰 방법이다. 하지만, 음성 데이터의 구간이 짧을 경우 그 평균값을 구하는 것이 어려운 문제로 남게 된다. 만약, 너무 짧은 구간의 음성데이터를 사용하여 평균값을 계산할 경우 음성신호 자체의 특성이 채널특성으로 나타날 수도 있어 오히려 인식성능을 감소할 수도 있게 된다. RTCN은 짧은 구간 음성 데이터의 켑스트럼 평균값과 그 이전 데이터에서 얻어진 켑스트럼 평균값을 사용하여 전체 켑스트럼의 평균값을 추정해

사용하는 방법으로, 식 (2)와 같이 구할 수 있다.

$$C_{t,i}' = C_{t,i} - X_t' \quad (2)$$

여기서, X_t' 는 t 번째 음성 데이터와 그 이전의 음성 신호를 모두 고려하여 추정된 켑스트럼 평균 벡터로서, 식 (3)과 같은 방법으로 구해진다.

$$X_t' = \alpha X_t + (1-\alpha) X_{t-1}' \quad (3)$$

여기서 $\alpha = 0.125$ 로 설정하였다.

3.2 특징파라미터 추출

(1) DWFBA

DWFBA는 켑스트럼이 채널, 주변 잡음의 간섭에 덜 민감하도록 하기위해 log 필터 뱅크 에너지의 높은 에너지 부분을 강조해 주는 것인데, 그림 1에서와 같이 DCT(Discrete Cosine Transform) 전에 각 critical band 의 log 에너지에 비례하도록 하는 가중함수를 곱하여 특징파라미터를 추출하는 방법이다.

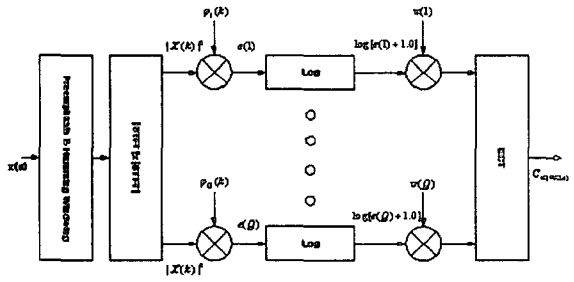

그림 1. DWFBA 기반의 특징파라미터 추출 블럭도

먼저, 입력신호 $x(n)$에 대해 프레임 단위로 프리엠퍼시스와 해밍윈도우를 거쳐 일반적인 MFCC 추출과정과 마찬가지로 필터뱅크 출력을 얻는다. 그런 다음 식 (4)와 (5)에 따라 DWFBA기반의 특징파라미터를 구하게 된다. 식 (5)에서 가중인자는 각각의 critical band 의 log 에너지에 비례함을 알 수 있으며, L은 켑스트럼의 차수이다.

$$c_m = \sum_{i=1}^{Q} w(i) \log[e(i)+1.0] \cos[m(\frac{2i-1}{2})\frac{\pi}{Q}], \quad 1 \le m \le L \quad (4)$$

$$w(i) = \frac{\log[e(i)+1.0]}{\sum_{j=1}^{Q} \log[e(i)+1.0]} \quad (5)$$

(2) SSC

SSC는 음성신호의 전체 주파수 대역을 몇 개의 서브밴드 영역으로 나누고, 각 서브밴드 영역 내에서 주파수의 무게중심(centroid)을 구하는 것인데, 정확한 포만트의 위치는 아니지만 포만트와 유사한 성분을 나타내게 된다. 이는 FFT(Fast Fourier transform)기반의 파워스펙트럼만의 정보를 이용하는 기존의 MFCC 추출 방법에, 우세한 파워스펙트럼 영역에 해당하는 주파수의 위치 정보를 추가로 이용하여 인식성능 향상에 도움을 주려는 것이다.

SSC를 구하기 위해 주파수 밴드를 나눌 때, [Hz] 단위로 균일하게 나눈 경우와 식 (6)을 이용하여 Mel 단위로 균일하게 나눌 수가 있는데, 본 논문에서는 Mel 단위에서 균일하게 나눈 Mel-SSC를 적용하였다.

$$mel = 2595 \log_{10}(1 + \frac{f}{700}) \quad (6)$$

여기서, mel 은 멜주파수이고, f 는 실제주파수를 나타낸다.

SSC를 구하는 식은 식 (7)과 같다. 본 연구에서 서브밴드의 수는 3으로 하였고, 파워스펙트럼의 dynamic range를 조절하는 상수는 1로 설정하였다.

$$C_m = \frac{\int_{l_m}^{h_m} f w_m(f) P^r(f) df}{\int_{l_m}^{h_m} w_m(f) P^r(f) df} \quad (7)$$

$P(f)$: Power spectrum

r : Power spectrum의 dynamic range 조절 상수

h_m, l_m: m번째 서브밴드의 lower, higher edge

SSC의 경우에도 동적인 특성을 잘 나타내는 Delta 특징파라미터를 정의하여 사용하였다. SSC의 Delta 성분은 MFCC의 경우와 동일하게 설정하였는데, 식 (8)과 같이 현재 구간을 포함하여 전후 2구간의 정적인 SSC의 차분을 사용하였다.

$$\Delta c_j(n) = c_j(n+2) - c_j(n-2) \quad (8)$$

IV. 인식실험 및 결과

Baseline 4연속숫자음 인식기는 공개 소프트웨어인 HTK(Hidden markov Tool Kit)를 사용하여 구현하였다. 음성신호는 20ms 의 윈도우 구간에 10ms 씩 중첩 이동하면서 특징을 추출하였다. 특징 파라미터로는 1차의 에너지, 12차의 멜켑스트럼 및 이들의 차분, 차차분을 포함한 총 39차의 파라미터를 사용하였으며, 음향모델은 트라이폰(Triphone) HMM 모델을 적용하였

다. 또한, 4연속숫자음 인식의 특성을 고려하여, 언어 모델은 FSN(Finate State Network)을 사용하였다. 그리고, 4연속숫자음에서 모두 15개의 음소를 정의하였고, 연속 HMM 모델의 state 수는 3, mixture 수는 8로하였다.

인식실험에 사용된 음성데이터는 160 종류의 4연속숫자음에 대해 9명의 화자(남자5명, 여자4명)가 발성한 1440개이다. 이 중 7명이 발성한 1120개의 숫자음성을 훈련에 사용하였고, 2명이 발성한 320개의 숫자음성을 테스트에 사용하였다. 수집된 전화음성 DB가 그리 크지 않으므로 Leave-one-out 방식을 적용하여 훈련과 인식실험을 수행하였다. 즉, 2명씩의 테스트 화자를 3개의 그룹(A,B,C)으로 나누어 각각의 경우에 대해 나머지 7명을 훈련용 음성으로 설정하여 음향모델을 만들고 인식실험을 행한 후, 3개 그룹의 평균인식률을 취하도록 하였다.

특징파라미터 추출방법 및 보상기법에 따른 인식실험의 결과는 표 2, 표 3과 같다. 39차의 MFCC를 사용한 Baseline 인식기를 기준으로 비교하면, 새로 도입한 특징파라미터인 DWFBA가 약 0.7%의 인식률 증가를 나타내었고, 여기에 보조 특징파라미터로 Delta_Mel-SSC 성분을 추가한 경우에 추가적으로 약 3%의 인식률 증가를 나타내었다. 또한 적용한 채널보상기법 중에는 RTCN이 가장 큰 인식률 증가를 나타내었다. 최종적으로, DWFBA에 보조특징파라미터로 Delta_Mel-SSC를 추가하고 RTCN을 채널보상기법으로 적용한 인식실험의 결과는 기존의 MFCC만 적용한 경우에 비해서는 8.85%, CMN과 함께 적용한 경우에 비해 1.77%의 인식률 증가를 나타내었다.

V. 결론

본 논문에서는 유/무선 전화망 환경에서 한국어 연속숫자음의 인식성능을 향상 시킬 수 있는 특징파라미터와 채널보상기법에 대해 실험하였다. 자체 수집한 전화음성 DB를 사용한 인식실험에서, 특징파라미터로는 기존의 MFCC에 비해 DWFBA 기반의 특징파라미터가, 채널보상방법으로는 RTCN이 CMN에 비해 더 좋은 인식률을 보였었다. 또한, 보조 특징파라미터로 Mel-SSC는 예상과는 달리 인식률 향상에 도움을 주지 못했으며, 이에 비해 Delta_Mel-SCC가 비교적 좋은 성능을 나타냄을 알 수 있었다. 그러나 이러한 특징파라미터 및 보상기법의 조합에 따른 인식률의 증가가 다른 DB에 대한 실험에서도 동일한 경향을 나타낼지는 알 수 없다. 따라서, 앞으로 대용량의 연속숫자음 전화음성 DB에 대해서도 같은 방법으로 인식실험을 수행하여 본 연구에서의 실험결과를 검증할 계획이다.

표 2. 특징파라미터 추출방법에 따른 인식률(%)
4연속숫자열 인식률(개별숫자 인식률)

특징파라미터	A그룹	B그룹	C그룹	평균
MFCC	85.00 (95.31)	71.25 (91.95)	82.19 (94.77)	79.48 (94.01)
DWFBA	85.94 (96.17)	73.13 (92.18)	81.56 (94.69)	80.21 (94.34)
DWFBA +Mel-SSC	84.69 (95.63)	67.50 (90.39)	80.63 (94.30)	77.60 (93.44)
DWFBA +△Mel-SSC	90.31 (97.27)	75.63 (93.05)	83.44 (95.16)	83.13 (95.16)

표 3. 특징파라미터에 보상기법을 적용한 인식률(%)
4연속숫자열 인식률(개별숫자 인식률)

특징파라미터	A그룹	B그룹	C그룹	평균
MFCC+CMN	90.00 (97.27)	82.50 (95.16)	87.19 (96.33)	86.56 (96.25)
DWFBA+ △Mel-SSC+CMN	91.25 (97.58)	82.50 (94.84)	87.50 (96.48)	87.08 (96.30)
MFCC+RTCN	89.69 (97.27)	81.88 (94.84)	87.81 (96.48)	86.48 (96.20)
DWFBA+ △Mel-SSC+RTCN	91.25 (97.66)	84.06 (95.63)	89.69 (96.88)	88.33 (96.72)

본 연구는 한국전자통신연구원 네트워크기술연구소 음성정보연구센터의 연구비 지원으로 수행 되었습니다.

참고문헌

[1] 정성윤, 김민성, 손종목, 배건성, "한국어 숫자음 전화음성의 채널왜곡에 따른 특징파라미터의 변이 분석" 대한전자공학회 하계종합학술대회, 제25권 제1호, 2002
[2] Wei-Wen Hung, Hsiao-Chuan Wang, "On the use of weighted filter bank analysis for the derivation of robust MFCCs", IEEE Signal Processing Letters, Vol. 8, No. 3, March 2000
[3] K.K. Paliwal, "Spectral subband centroid features for speech recognition" ,ICASSP98,Vol.2, pp. 617-620