한국어 숫자음 전화음성의 채널왜곡에 따른 특징파라미터의 변이 분석

정 성 윤, 손 종 목, 김 민 성, 배 건 성
경북대학교 전자공학과

Variation Analysis of Feature Parameters According to the Channel Distortion of Korean Telephone Digit Speech

Sung-Yun Jung, Jong Mok Son, Min Sing Kim, Keun Sung Bae
School of the Electronic & Electrical Engineering, Kyungpook National University
E-mail : yunij@mir.knu.ac.kr

Abstract

The final purpose of this paper is the enhancement of speech recognition rate under the matched telephone environment between training data and test data. To analyze the effect by the distortion of the changing telephone channel on every call, MFCC is used as the feature parameter and CMN, RTCN, and RASTA are used as channel compensation techniques. For each case, the variation of feature parameters of all phones is analyzed. And, we find recognition rates according to each compensation method using the continuous HMM recognizer, and examine the relationship between variation and recognition rate

I. 서론

음성인식 기술의 발달과 함께 이를 다양한 서비스에 응용하려는 요구가 증가하면서, 전화망 환경에서는 음성다이얼링이나 증권안내, 자동응답시스템 등의 분야에 음성인식 기술이 적용되어 부분적으로 실용화의 성과를 얻고있다. 최근에는 이동전화 사용의 급격한 증가와 단말기의 소형화에 따라 무선전화망 환경에서 음성인식 기술의 적용이 더욱 중요시 되고있다. 그러나, 전화음성의 인식률은 전화망 환경에서 수반되는 신호의 왜곡 및 잡음으로 인해 일반 마이크 음성의 인식률에 비해 아직 만족스럽지 못한 수준이며, 특히, 한국어 연속 숫자음의 경우 다양한 조음효과로 인해 인식에 어려움이 많다. 앞으로, 유선전화 또는 이동전화를 이용하여 컴퓨터-전화 통합시스템을 이용한 주식거래, 신원조회, 정보검색 등과 같은 다양한 서비스를 제공하기 위해서는 유/무선 전화망 환경에서 음성인식 성능을 향상시키기 위한 연구가 절대적으로 필요하다. 특히, 전화망을 통한 연속 숫자음의 인식은 신원조회 등과 같은 보안이 요구되는 서비스에 필수적이므로, 전화음성의 채널왜곡 및 잡음의 영향을 보상하여 인식률을 향상시킬 수 있는 기법에 대한 연구가 선행되어야 한다.

전통적으로 환경이나 채널의 보상기법에 관한 연구는 크게 2가지 영역, Feature-domain 과 Model-domain에서 접근되어 왔다. Model-domain 접근방법의 목적은 잡음환경의 테스트 음성에서의 통계치와 일치하도록 미리 훈련된 reference HMM의 파라미터들을 변경하는 것이다. 따라서, 인식 성능은 잡음환경에 일치된 조건하에서 얻을 수 있는 성능에 의해 결정되어진다. PMC(Parallel Model Combination), CDCN(Codeword-Dependent Cepstral Normalization), SM(Stochastic Matching)등이 있다. Feature-domain 접근방법은 인식과정 전에 전 처리단에서 잡음환경에 강인한 특징추출 파라미터나 채널잡음에 의한 영향을 보상해 주는 것으로써, CMN(Cepstral Mean

Normalization), RTCN(Real Time Cepstral Normalization), RASTA(RelAtive SpecTrAl)등의 Cepstral smoothing 기법들이 대표적이다. [1],[5]

본 논문에서는 MFCC를 특징파라미터로 사용하여 유/무선전화음성에 대해 CMN, RTCN, RASTA를 보상기법으로 적용하여 각각의 경우에 대해, 매 통화시 변화하는 특징파라미터의 변이를 분석한다. 그리고, Continuous HMM방식의 HTK 인식기를 사용하여 각 경우의 인식률을 구하여 보상기법에 따른 변이와 인식률과의 관계를 확인한다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 유/무선 전화음성의 신호왜곡 특성을 분석하기 위해 분석용 소용량의 유무선 전화음성 DB를 수집하는 내용을 기술하고, 3장에서는 보상기법에 따른 채널왜곡 특성을 분석하고, 4장에서는 Baseline 인식기를 기반으로 각 보상기법에 따른 인식실험 및 결과를 검토한 후, 5장에서 결론을 맺는다.

II. 전화음성 DB 수집

분석에 사용될 연속 숫자음은 ETRI 측에서 제공한 1000개의 4연속 숫자음 목록 중에서 160개를 임의 선정하여 표 1과 같이 영과 공을 포함하여 사용하였다. 전화음성은 매 통화시 변경되는 전화망의 경로에 따라 채널 특성이 변화하면서 음성신호를 왜곡시키는데, 이러한 특성을 분석하기 위해 한 통화당 8개의 4연속 숫자음을 정하여 모두 20 통화분을 준비하였다.

표 5 . 4연속 숫자음성의 선정 예

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

전화음성의 녹음은 Dialogic 사의 전화 인터페이스 카드를 사용하여 PC에서 자동으로 전화음성을 녹음할 수 있도록 시스템을 구현하였다. 전화음성은 전화 인터페이스 카드에서 샘플링 8kHz, μ-law 포맷으로 녹음되고, 녹음이 이루어진 시간을 기준으로 자동으로 파일이름이 결정되어 저장된다. 8kHz, μ-law 포맷으로 저장된 음성파일은 나중에 μ-law expanding을 통해 8kHz, 16-bit Linear PCM으로 변환되어 분석용 파일로 저장된다.

160개의 4연속 숫자음에 대해, 연구실에서 10명(남자 5명, 여자5명)의 화자가 유선전화를 통해 2회 발성하여 녹음하였고, 5명의 화자가 무선전화를 사용하여 2회 녹음하였다. 이 중 음소별 분석을 위해, 유선전화를 통해 녹음한 5명의 전화음성과 무선전화를 통해 녹음한 5명의 전화음성에 대해 음소 레이블링을 수행하였다.

III. 전화음성의 신호왜곡 분석

전화음성의 신호왜곡 분석은 MFCC를 특징파라미터로 사용하여 CMN, RTCN, RASTA를 보상기법으로 적용하여 각각 경우에 대해, 매 통화시의 특징 파라미터 차수에 대한 음소들의 변이를 기법에 따라 분석하였다.

3.1 기존의 채널 보상 기법

(1) CMN (Cepstral Mean Normalization) [4],[7]

CMN의 기본 개념은 시간영역에서 컨벌루션의 형태로 나타나는 채널특성이 켑스트럼 영역에서 합의 형태로 나타난다는 것에 있다. 채널특성은 단시간에 큰 변화가 생기지 않고 거의 일정하게 나타나기 때문에, 켑스트럼 영역에서는 전체 켑스트럼의 바이어스 성분으로 볼 수 있다. 즉, 음성신호가 임의의 채널을 통해 녹음되었을 때 켑스트럼 도메인에서는 채널특성이 음성신호의 켑스트럼에 합해진 형태로 나타나기 때문에, 켑스트럼의 바이어스(평균값)를 제거해주는 것만으로도 채널왜곡으로 인한 인식성능 감소를 상당히 줄일 수 있다. CMN의 일반적인 적용과정은 다음과 같다. 신호가 주어졌을 때 단구간 신호분석을 통하여 T개의 켑스트럼을 계산하며, 켑스트럼의 평균은 식(1)과 같이 구할 수 있다.

$$x_{cmn} = \frac{1}{T}\sum_{t=1}^{T} x_t \qquad (1)$$

여기서 x_t 는 시간 t 에서의 켑스트럼 벡터이다. CMN이 각 켑스트럼 벡터의 바이어스를 제거해 주는 과정이므로, 식(2)와 같이 CMN을 적용하여 정규화된 켑스트럼을 구할 수 있다.

$$x_t' = x_t - x_{cmn} \qquad (2)$$

(2) RTCN(Real-Time Cepstral Normalization)

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 나타나게 되어 오히려 인식성능을 감소할 수도

있게 된다. RTCN은 짧은 순간 켑스트럼의 평균값을 사용하여 전체 켑스트럼 바이어스를 추정해 사용하는 방법으로, 본 연구에서는 식(3)과 같은 방법으로 적용하였다.

$$x_{rtcnm} = \alpha x_{mt} + (1-\alpha) x_{rtcn(t-1)} \qquad (3)$$

여기서, x_{rtcnt} 는 t번째 추정과정에서의 추정 켑스트럼 평균이고, x_{mt} 는 t번째 음성신호의 켑스트럼 평균이다.

(3) RASTA(RelAtive SpecTrAl)

RASTA 필터는 음성에 비해 완만히 변하는 채널 왜곡 및 음성에 비해 빠르게 변하는 부분을 억제하여 음성부분을 강조시키는 방법으로 필터뱅크 출력에너지의 시간궤적에 대해 필터링을 수행할 수 있다. RASTA 필터의 전달함수는 아래의 식 (4)와 같다.[3]

$$H(z) = \frac{0.2 + 0.1z^{-1} - 0.1z^{-3} - 0.2z^{-4}}{1 - 0.94z^{-1}} \qquad (4)$$

는 RASTA 필터의 주파수응답에서 저역차단 주파수를 결정하는 필터계수로, 음성에 비해 느리게 변하는 채널왜곡을 제거한다.

3.2 분석 결과

전화음성의 신호왜곡 특성분석은 특징 파라미터를 기준으로 기존의 보상기법인 CMN, RTCN, RASTA를 적용하였다. 매 통화시 변하는 전화채널의 왜곡에 대한 변이를 각각의 기법에 대해, 모든 음소들의 특징파라미터 차수별 변이로 분석하였다. 화자간의 변이는 고려하지 않고, 채널변이만을 분석하기 위해 한 명의 화자가 발성한 20통화분의 전화음성에 대해 분석하였다.

분석과정은, 먼저 20통화분의 전화음성에 대해 레이블링 정보를 바탕으로 모든 음소에 대해 각 차수별 분산값을 구한다. 그리고 전체 음성에 대해 각 차수별 분산값을 구하여 Global 분산값을 정한다. Global 분산값이 정해지면, 각 차수별 분산값에 역의 가중치를 주어 정규화 과정을 행한다. 이렇게 Global 분산값으로 정규화된 각 차수별 분산값을 모든 음소에 대해 구하여 보상기법에 따른 변이를 분석하게 된다.

그림 1과 그림 2는 각각 음소 /i/와 음소 /o/에 대한 보상기법별 분산값을 나타낸 것이다. 그림에서, 대부분의 차수에 대해 MFCC가 가장 변이가 큰 것을 알 수 있고, RTCN이 변이가 작은 것을 확인할 수 있다. 그러나, 차수에 따라 각 기법의 변이 정도는 다르다. 9차 이상의 높은 MFCC 차수에서는 RASTA가 가장 변이가 크고, 12차에서는 MFCC가 가장 변이가 작음을 알 수 있다.

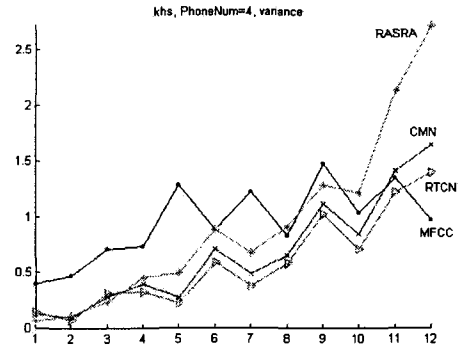

그림 1. 음소 /i/ 에 대한 보상기법별 분산값

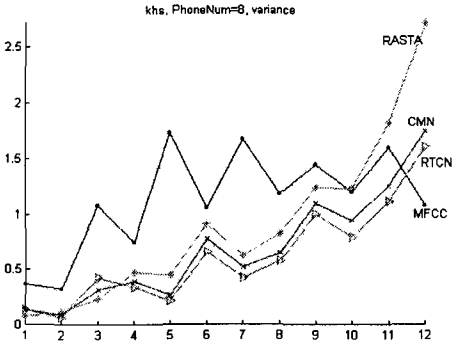

그림 2. 음소 /o/ 에 대한 보상기법별 분산값

Ⅳ. 인식실험 및 결과

Baseline 4연속 숫자음 인식기 구현은 공개 소프트웨어인 HTK(Hidden markov Tool Kit)를 사용하였다. 음성신호는 20ms 의 윈도우 구간에 10ms 씩 중첩이동하면서 특징을 추출하였다. 특징 파라미터로는 38차의 멜켑스트럼을 사용하였으며, 음향모델은 트라이폰(Triphone) HMM모델을 적용하였다. 또한, 4연속 숫자음 인식의 특성을 고려하여, 언어모델은 FSN(Finate State Network)을 사용하였다. 그리고, 4연속 숫자음에서 표 2와 같이 모두 15개의 음소를 정의하여, 3 states, 8 mixture 의 연속 HMM 모델을 적용하였다.

표 2. 15개의 음소 집합

번호	기호	음소	번호	기호	음소
1	a	ㅏ	9	p	ㅍ
2	ch	ㅊ	10	s	ㅅ
3	g	ㄱ	11	sil	묵음
4	i	ㅣ	12	sp	짧은포즈
5	le	ㄹ	13	u	ㅜ
6	me	ㅁ	14	yeo	ㅕ
7	nge	ㅇ	15	yug	육
8	o	ㅗ			

인식실험에 사용된 음성데이터는 160개의 4연속 숫자음에 대해 10명의 화자(남자5명, 여자5명)가 2번 발성한 3200개이다. 이 중 남, 여 각 4명이 발성한 2560개의 숫자음성을 훈련에 사용하였고, 남, 여 각각 1명이 발성한 640개의 숫자음성을 테스트에 사용하였다.

각 보상기법에 따른 인식성능은 표 3과 같다. 38차의 MFCC를 사용한 Baseline 인식기를 기준으로 비교하면 RTCN을 적용한 경우가 가장 인식성능이 좋고, RASTA를 적용한 경우가 인식성능이 가장 낮은 것으로 나타났다. 3장의 변이분석의 결과는 RTCN, CMN, RASTA, MFCC 순으로 변이가 많은 것으로 나타났는데, 이는 인식률의 결과와 정확히 일치하지는 않는다.

RTCN의 경우에는 변이도 가장 적었고, 인식성능도 가장 높은 것으로 나타난 반면, CMN 이나 RASTA는 변이가 MFCC 보다 작은데도 불구하고 인식률면에서 MFCC 보다 떨어지는 것으로 나타났다. 이것은 화자의 변이가 포함된 인식실험에 따른 결과 때문이다. 3장에서 분석한 분산값은 한명의 화자에 대한 결과이기 때문에 인식률과의 결과를 정확히 비교하려면, 향후, 화자에 대한 변이도 함께 포함시켜 분석해야 한다.

표 3. 보상기법에 따른 인식결과

인식률 / 보상기법	인식률 (%)	
	4연 숫자열	개별 숫자
Baseline (MFCC38차)	90.78	97.54
MFCC+CMN	87.97	96.84
MFCC+RASTA	85.63	95.98
MFCC+RTCN	91.25	97.77

V. 결론

본 논문에서는 매 통화마다 변화하는 채널의 변이를 4연속 숫자 전화음성에 대해 특징파라미터를 기반으로 기존의 보상기법인 CMN, RTCN, RASTA에 따라 비교 분석하였고, 인식 실험을 통해 인식률도 확인하였다. 변이분석의 결과, RTCN의 경우가 가장 변이가 작았고 또한 인식성능도 가장 높았다. 그러나 CMN이나 RASTA는, 채널의 변이 분석의 결과와 인식성능과의 관계가 일치하지 않음을 확인할 수 있었는데, 이는 화자의 변이를 고려하지 않은 상태에서 변이분석이 이루어졌기 때문이라 사료된다. 따라서 향후, 화자의 변이를 포함한 채널왜곡 특성을 분석하여, 인식성능과의 관계를 확인해야 한다.

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

참고문헌

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

[2] C. Mokbel, J. Monne and D. Jouvet, "On -line adaptation of a speech recognizer to variations in telephone line condition," Proc. Eurospeech, pp.1247-1250, 1993

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

[4] A. Acero, "Environmental Robustness in Automatic Speech Recognition," Proc. ICASSP, pp.849-852

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

[6] J. G. Wilpon, C. H. Lee, and L. R. Rabiner, "Improvements in the Connected Digit Recognition Using Higher Order Spectral and Energy Feature," Int. Conf. on Acoustics, Seech, and Signal Processing, vol.1, pp.349-352, 1991.

[7] 김상진, 서영주, 한민수 "LCMS를 이용한 한국어 연속 숫자인식에 관한 연구," 한국음향학회 논문집, Vol.20, pp.43-46 , 2001