한국어 숫자음 전화음성의 채널왜곡에 따른 특징파라미터의 변이 분석 및 인식실험*

정성윤(경북대), 손종목(경북대), 김민성(경북대), 배건성(경북대)

<차 례>

<Abstract>

Analysis of Feature Parameter Variation for Korean Digit Telephone Speech according to Channel Distortion and Recognition Experiment

Sung-Yun Jung, Jong-Mok Son, Min-Sung Kim, Keun-Sung Bae

Improving the recognition performance of connected digit telephone speech still remains a problem to be solved. As a basic study for it, this paper analyzes the variation of feature parameters of Korean digit telephone speech according to channel distortion. As a feature parameter for analysis and recognition, MFCC is used. To analyze the effect of telephone channel distortion depending on each call, MFCCs are first obtained from the connected digit telephone speech for each phoneme included in the Korean digit. Then CMN, RTCN, and RASTA are applied to the MFCC as channel compensation techniques. Using the feature parameters of MFCC, MFCC+CMN, MFCC+RTCN, and MFCC+RASTA, variances of phonemes are analyzed and recognition experiments are done for each case. Experimental results are discussed with our findings and discussions

* 주제어: 전화음성 연속숫자음 인식, 변이분석, 채널보상, CMN, RASTA, RTCN

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

1. 서 론

음성인식 기술의 발달과 함께 이를 다양한 서비스에 응용하려는 요구가 증가하면서, 전화망 환경에서는 음성다이얼링이나 증권안내, 자동응답시스템 등의 분야에 음성인식 기술이 적용되어 부분적으로 실용화의 성과를 얻고 있다. 최근에는 이동전화 사용의 급격한 증가와 단말기의 소형화에 따라 무선전화망 환경에서 음성인식 기술의 적용이 더욱 중요시되고 있다. 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 주민등록번호를 이용한 신원조회 또는 신용카드의 번호인식 등과 같은 보안이 요구되는 서비스에 필수적이므로, 한국어 연속 숫자음에 대한 전화음성의 채널왜곡 및 잡음의 영향을 줄여서 인식률을 향상시킬 수 있는 기법에 대한 연구가 중요해진다.

전통적으로 주변잡음이나 채널의 보상기법에 관한 연구는 크게 2가지 영역, Feature-domain과 Model-domain에서 접근되어 왔다. Model-domain 접근 방법은 주로 잡음환경인 테스트 음성에서의 통계적인 특성과 유사해지도록 미리 훈련된 reference HMM의 파라미터들을 변경하는 것이다. 따라서 인식 성능은 잡음환경에 일치된 조건하에서 얻을 수 있는 성능에 의해 결정되어지는데, 대표적인 방법들로써 PMC(Parallel Model Combination), CDCN(Codeword-Dependent Cepstral Normalization), SM(Stochastic Matching) 등이 있다. Feature-domain 접근방법은 인식과정 전에 전 처리 단에서 잡음환경에 강인한 특징파라미터를 추출하거나 채널잡음에 의한 영향을 보상해 주는데, CMN (Cepstral Mean Normalization), RTCN(Real Time Cepstral Normalization), RASTA(RelAtive SpecTrAl) 등의 Cepstral smoothing 기법들이 대표적이라고 할 수 있다[1,5].

본 연구는 한국어 숫자음의 전화음성 인식률을 향상시키기 위한 선행 연구로서, MFCC(Mel Frequency Cepstral Coefficient)를 기본 특징파라미터로 사용하여, 유/무선 환경에서 수집된 전화음성에 대해 CMN, RTCN, RASTA를 보상기법으로 적용하여 각각의 경우에 대해, 매 통화 시 변화하는 특징파라미터의 변이를 비교, 분석한다. 그리고 Continuous HMM 방식의 HTK 인식기를 이용한 4연 숫자음 인식실험을 통해 각 보상기법에 따른 변이와 인식률과의 관계를 조사하였다. 본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 유/무선 환경에서의 전화음성의 신호왜곡 특성을 분석하기 위해 전화음성 DB를 수집하는 내용을 기술하고, 3

장에서는 보상기법에 따른 전화음성의 신호왜곡을 분석한 결과를 제시하며, 4장에서는 Baseline 인식기를 기반으로 각 보상기법에 따른 인식실험 및 결과를 검토한 후, 5장에서 결론을 맺는다.

2. 전화음성 DB 수집

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

<표 1> DB 수집에 선정한 4연 숫자음의 리스트

7106(칠 일 공 육)	5026(오 공 이 육)	2020(이 공 이 공)	2552(이 오 오 이)
5225(오 이 이 오)	6391(육 삼 구 일)	7822(칠 팔 이 이)	1586(일 오 팔 육)
6426(육 사 이 육)	0707(영 칠 영 칠)	6646(육 육 사 육)	8466(팔 사 육 육)
8999(팔 구 구 구)	9600(구 육 공 공)	1808(일 팔 공 팔)	9000(구 영 영 영)
.....			

전화음성의 녹음은 Dialogic사의 전화 인터페이스 카드를 사용하여 PC에서 자동으로 전화음성을 녹음할 수 있도록 시스템을 구현하였다. 전화음성은 샘플링 주파수 8kHz, μ-law 포맷으로 녹음되고, 녹음이 이루어진 시간을 기준으로 자동으로 파일 이름이 결정되어 저장된다. μ-law 포맷으로 저장된 음성파일은 나중에 μ-law expanding을 통해 8kHz, 16-bit Linear PCM으로 변환되어 분석용 파일로 저장된다. 160개의 4연 숫자음에 대해, 연구실에서 10명(남자 5명, 여자5명)의 화자가 유선전화를 통해 2회 발성하여 녹음하였고, 5명의 화자가 이동통신 전화를 사용하여 2회 녹음하였다. 이 중 음소별 분석을 위해, 유선전화를 통해 녹음한 5명의 전화음성 1회 발성분과 이동전화를 통해 녹음한 5명의 전화음성 1회 발성분에 대해 음소 레이블링 작업을 수행하였다.

3. 전화음성의 신호왜곡 분석

전화음성의 신호왜곡 분석은 MFCC를 특징파라미터로 사용하여 CMN, RTCN,

RASTA를 보상기법으로 적용하여 각 경우에 대해, 매 통화시의 특징 파라미터 차수에 대한 음소들의 변이를 분석하였다.

3.1. 기존의 채널 보상 기법

3.1.1. CMN (Cepstral Mean Normalization) [4,7]

CMN의 기본 개념은 시간영역에서 컨벌루션의 형태로 나타나는 채널특성이 켑스트럼 영역에서 합의 형태로 나타나며, 채널특성은 단시간에 큰 변화가 생기지 않고 거의 일정하게 나타나기 때문에 켑스트럼 영역에서는 전체 켑스트럼의 바이어스 성분으로 볼 수 있다는 것이다. 즉, 음성신호가 임의의 채널을 통해 녹음되었을 때 켑스트럼 도메인에서는 채널특성이 음성신호의 켑스트럼에 합해진 형태로 나타나기 때문에, 켑스트럼의 바이어스(평균값)를 제거해 줌으로써 채널왜곡으로 인한 영향을 상당히 줄일 수 있게 된다. CMN의 일반적인 적용과정은 다음과 같다. 신호가 주어졌을 때 단구간 신호분석을 통하여 얻어지는 T개의 전체 켑스트럼에 대해 식(1)과 같이 평균을 구할 수 있다.

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

여기서 x_t는 시간 t에서의 켑스트럼 벡터이다. CMN 과정은 각 켑스트럼 벡터의 바이어스를 제거해 주는 과정이므로, 식(2)와 같이 CMN을 적용하여 정규화된 켑스트럼을 구할 수 있다.

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

3.1.2. RTCN (Real-Time Cepstral Normalization)

CMN은 간단하면서도 그 왜곡보상 능력은 큰 방법이다. 그러나, 평균값을 구하기 위한 음성 데이터의 구간이 충분히 길어야 하며, 너무 짧은 구간의 음성데이터를 사용하여 평균값을 계산할 경우 음성신호 자체의 특성이 채널특성으로 나타나게 되어 오히려 인식성능이 감소할 수도 있게 된다. RTCN은 짧은 구간의 음성 데이터로부터 얻어지는 켑스트럼의 평균값을 사용하면서 그 이전의 켑스트럼 특성을 포함하도록 전체 켑스트럼 바이어스를 추정해 사용하는 방법으로, 식(3)과 같이 표현된다.

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

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

3.1.3. RASTA (RelAtive SpecTrAl)

RASTA 필터는 음성에 비해 완만히 변하는 채널 왜곡 및 음성에 비해 빠르게 변하는 부분을 억제하여 음성부분을 강조시키는 방법으로 특징벡터 각 계수의 시간궤적에 대해 필터링을 수행한다. RASTA 필터의 전달함수는 아래의 식 (4)와 같다[3].

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

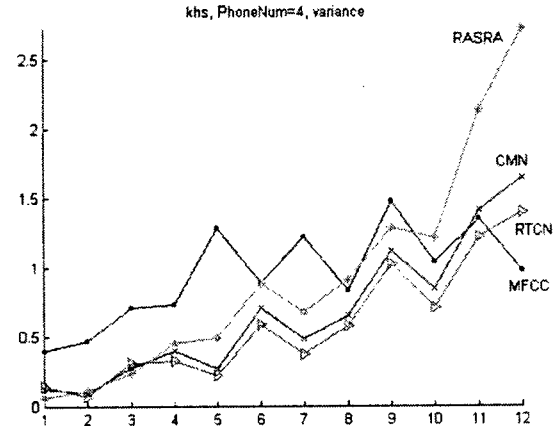
α는 RASTA 필터의 주파수응답에서 저역차단 주파수를 결정하는 필터계수로, 음성에 비해 느리게 변하는 채널왜곡을 제거한다. 본 논문에서는 α값을 0.94로 설정하였다.

3.2. 분석 결과

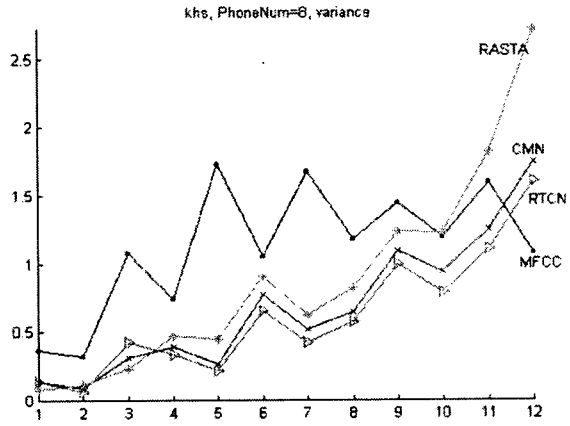
전화음성의 신호왜곡 특성 분석은 특징파라미터인 MFCC를 기준으로 기존의 보상기법인 CMN, RTCN, RASTA를 적용하여 한국어 숫자음에 포함된 모든 음소들의 차수별 변이를 분석하였다. 화자간의 변이는 고려하지 않고, 채널변이만을 분석하기 위해 한 명의 화자가 발성한 20 통화분의 전화음성에 대해 분석하였다. 먼저, 20통화분의 전화음성에 대해 레이블링 정보를 바탕으로 모든 음소에 대해 각 차수별 분산값을 구한다. 그리고 전체 음성에 대해 각 차수별 분산값을 구하여 Global 분산값을 정한다. Global 분산값이 정해지면, 각 차수별 분산값에 역의 가중치를 주어 정규화 과정을 행한다. 이렇게 Global 분산값으로 정규화 된 각 차수별 분산값을 모든 음소에 대해 구하여 보상기법에 따른 변이를 분석하였다.

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 다르게 나타났었다. 9차 이상의 높은 MFCC 차수에서는 RASTA가 가장 변이가 크고, 12차에서는 MFCC가 가장 변이가 작음을 알 수 있다. 그리고 RTCN

과 CMN은 분산값의 차이가 크지 않음을 볼 수 있다. 다른 음소에 대한 분석 결과도 그림 1, 2와 유사한 패턴을 나타내었다.



<그림 1> 음소 /i/ 에 대한 보상기법별 분산값



<그림 2> 음소 /o/ 에 대한 보상기법별 분산값

4. 인식실험 및 결과

Baseline 4연 숫자음 인식기 구현은 공개 소프트웨어인 HTK(Hidden markov Tool Kit)를 사용하였다. 음성신호는 20ms의 윈도우 구간에 10ms 씩 중첩이동하면서 특징을 추출하였다. 특징 파라미터로는 38차의 MFCC를 사용하였으며, 음향모델은

트라이폰(Triphone) HMM모델을 적용하였다. 또한, 4연 숫자음의 특성을 고려하여, 언어모델은 FSN(Finate State Network)을 사용하였다. 그리고 한국어 연속 숫자음에 포함된 음소를 표 2에서와 같이 모두 15개의 유사음소 단위로 정의하여, 3 states, 8 mixture의 연속 HMM 모델을 적용하였다.

<표 2> 숫자음 인식에 사용된 15개의 유사음소

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	short pause
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 4연 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 인식실험은 Leave-one-out 방식을 적용하여, 훈련음성 및 테스트 음성 DB가 작은점에 대한 약점을 보완하도록 하였다. 즉, 인식실험에 사용될 9명의 음성데이터를 테스트음성으로 적용될 화자의 조합별로 4개의 그룹(A, B, C, D)으로 나누어 훈련용 음성데이터는 유선전화 4명분과 무선전화 3명분의 총 7명분의 음성 데이터로, 인식테스트용 음성 데이터는 유, 무선 각 1명분으로 총 2명의 음성 데이터가 Leave-one-out 방식을 적용하여 사용되었다. 4개 그룹의 인식결과는 표 3과 같다.

4개 그룹의 평균인식률을 비교하면 RTCN을 적용한 경우가 가장 인식성능이 좋고, MFCC만을 사용한 경우가 상대적으로 가장 낮은 인식성능을 나타내었다. 3장의 변이분석의 결과에서 RTCN, CMN, RASTA, MFCC 순으로 변이가 많은 것으로 나타났는데, 이는 인식률의 결과와 일치하는 경향을 보이고 있다. 즉, RTCN을 적용한 경우에는 변이도 가장 작고 인식성능도 가장 높은 것으로 나타났고, MFCC는 가장 큰 변이에 가장 낮은 인식률을 나타내었다. 그리고 변이분석에서 큰 차이가 없었던 CMN과 RTCN의 경우에도 서로간의 인식률 차이가 크지 않음을

알 수 있다. 그러나 각 그룹별 인식결과는 변이분석의 결과와 다소 다른 경향을 보이는 부분이 있는데, 이는 훈련 데이터와 테스트 데이터의 수가 작은 탓에 인식 성능이 나쁜 특정화자의 영향이 크게 나타났기 때문이라 판단된다.

<표 3> Leave-one-out 방식을 적용한 보상기법별 인식율

인식률 / 보상 기법	인식률 (%)				
	A그룹	B그룹	C그룹	D그룹	**평균**
Baseline (MFCC38차)	83.13	70.31	82.81	95.31	**82.89**
MFCC+CMN	88.75	79.06	85.94	97.30	**87.76**
MFCC+RASTA	85.00	71.25	86.25	96.25	**84.95**
MFCC+RTCN	87.81	83.44	87.81	97.31	**88.83**

인식결과에서 38차 기반의 MFCC에 RTCN을 적용한 경우에 가장 좋은 인식성능을 나타냄을 확인하였다. 실제 11개의 숫자음중에서 가장 많이 혼동되는 단어쌍들은 /일/→/이/, /이/→/일/, /일/→/칠/, /오/→/구/, /육/→/영/ 등이다. 이 숫자쌍들은 인식성능에 가장 많은 영향을 끼치고 인식율 저하의 주요 요인이 된다. 인식결과의 Confusion Matrix로부터 이들의 오인식된 단어쌍들에 대한 총 오인식 개수를 표 4에 나타내었다. 표 4로부터 오인식 단어쌍들의 오인식 단어수가 전반적으로 작은 보상기법은 RASTA를 적용한 방법이다. 보상기법에 따라 오인식되는 숫자음의 분포가 많이 다름을 볼 수 있는데, RASTA의 경우 다른 숫자음들에 비해 /이/를 /일/로 오인식하는 경우가 훨씬 큼을 볼 수 있다.

<표 4> 보상기법에 따른 오인식 단어쌍 들의 단어 갯수

오인식 단어 / 보상기법	일→이	이→일	일→칠	오→구	육→영
MFCC	16	13	17	16	23
MFCC+CMN	11	26	12	10	13
MFCC+RASTA	9	34	9	13	14
MFCC+RTCN	14	23	13	12	15

5. 결 론

본 논문에서는 매 통화마다 변화하는 채널의 변이를 4연 숫자음 전화음성에 대해 특징파라미터를 기반으로 기존의 보상기법인 CMN, RTCN, RASTA 적용에 따라 비교 분석하였고, Leave-one-out 방식의 인식 실험을 통해 분석된 변이와 인식률과의 관계를 검토하였다. 변이분석과 인식실험의 결과, 보상기법으로 RTCN을 적용한 경우가 가장 작은 변이와 가장 높은 인식률을 나타내었다. 인식결과에 대한 Confusion Matrix로 부터 인식성능 저하에 절대적인 영향을 끼치는 오인식 단어쌍들이 /일/→/이/, /이/→/일/, /일/→/칠/, /오/→/구/, /육/→/영/ 이고, 이들의 성능 향상만이 궁극적으로 전화음성 연속숫자음의 인식성능을 향상시킬 수 있음을 알 수 있었다. 실험결과를 바탕으로 하여 향후, /이/와 /일/, /오/와 /구/ 등의 오인식 단어쌍들의 혼동을 줄일 수 있는 특징파라미터 연구를 통해 전화음성 연속숫자음의 인식성능을 향상시키고자 한다.

참 고 문 헌

[1] Moreno, P.J. (1993), Speech Recognition in Telephone Environment, MS. Thesis, CMU.

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

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

[4] Acero, A. (1990), Environmental Robustness in Automatic Speech Recognition, *Proc. ICASSP*, pp.849～852.

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

[6] Wilpon, J. G., C. H. Lee, and L. R. Rabiner (1991), 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.

[7] 김상진, 서영주, 한민수(2001), LCMS를 이용한 한국어 연속 숫자인식에 관한 연구, 「한국음향학회 논문집」 Vol.20, pp.43～46.

접수일자: 2002년 5월 7일
게재결정: 2002년 5월 24일

▶ 정성윤(Sung-Yun Jung)
주소: 702-701 대구광역시 북구 산격동 1370번지 경북대학교 공과대학 전자공학과 10호관 724호
소속: 경북대학교 전자공학과 신호처리연구실
전화: 053) 940-8627
Fax: 053) 950-5505
E-mail: yunij@mir.knu.ac.kr

▶ 손종목(Jong-Mok Son)
주소: 702-701 대구광역시 북구 산격동 1370번지 경북대학교 공과대학 전자공학과 10호관 724호
소속: 경북대학교 전자공학과 신호처리연구실
전화: 053) 940-8627
Fax: 053) 950-5505
E-mail: sjm@palgong.knu.ac.kr

▶ 김민성(Min-Sung Kim)
주소: 702-701 대구광역시 북구 산격동 1370번지 경북대학교 공과대학 전자공학과 10호관 724호
소속: 경북대학교 전자공학과 신호처리연구실
전화: 053) 940-8627
Fax: 053) 950-5505
E-mail: kmslove@mir.knu.ac.kr

▶ 배건성(Keun-Sung Bae)
주소: 702-701 대구광역시 북구 산격동 1370번지 경북대학교 공과대학 전자공학과 10호관 719호
소속: 경북대학교 전자공학과 신호처리연구실
전화: 053) 950-5527
Fax: 053) 950-5505
E-mail: ksbae@ee.knu.ac.kr