한국어 연속 숫자음 전화 음성 인식에서의 오인식 유형 분석*

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

<차 례>

<Abstract>

Analysis of Error Patterns in Korean Connected Digit Telephone Speech Recognition

Min Sung Kim, Sung Yun Jung, Jong Mok Son, Keun Sung Bae, Sang Hun Kim

Channel distortion and coarticulation effect in the Korean connected digit telephone speech make it difficult to achieve high performance of connected digit recognition in the telephone environment. In this paper, as a basic research to improve the recognition performance of Korean connected digit telephone speech, recognition error patterns are investigated and analyzed. Korean connected digit telephone speech database released by SiTEC and HTK system are used for recognition experiments. Both DWFBA and MRTCN methods are used for feature extraction and channel compensation, respectively. Experimental results are discussed with our findings.

* Keywords: Connected digit telephone speech recognition, DWFBA, MRTCN

* 본 논문은 한국전자통신연구원 네트워크기술연구소 음성정보연구센트의 연구비 지원으로 수행되었습니다.

1. 서 론

오늘날 음성인식 기술의 발달로 이를 이용한 다양한 서비스가 제공되고 있다. 전화망 환경에서는 음성 다이얼링이나 증권 안내, 자동 응답 시스템 등의 분야에서 음성인식이 적용되어 부분적으로 실용화되기도 했다. 최근에는 이동전화 사용의 급격한 증가와 단말기를 이용한 다양한 서비스를 제공받고자 함에 따라 전화망 환경에서 음성인식 기술의 적용이 더욱 중요시되고 있다. 하지만 호형성시마다 달라지는 전화 채널 특성과 부가 잡음의 영향으로 전화망 환경에서의 음성인식의 경우 PC 환경에서 만큼의 성능을 얻지 못하고 있다[1, 2]. 특히, 한국어 연속 숫자음 인식의 경우 조음 효과와 인식이 어려운 숫자 쌍들의 영향으로 만족할 만한 성능을 얻지 못하고 있는데, 주파수 대역의 에너지 비를 이용한 특징 파라미터를 사용함으로써 이러한 오인식 쌍들의 인식률을 다소 개선할 수 있다고 하나 전화 음성에 대해서는 채널왜곡으로 인해 그러한 개선 효과를 얻지 못하고 있다[3].

전화망에서의 인식 성능 향상을 위해서는 채널 보상 기법이나 모델 적응 방법이 많이 연구되어 왔다. 채널보상 기법은 켑스트럼 영역에서 전화 음성의 켑스트럼 평균이 채널의 영향을 나타낸다고 보고 이를 특징 파라미터에서 빼 줌으로써 채널에 의한 왜곡을 보상하는 기법이다[4, 5]. 이외에도 특징 파라미터 추출단계에서 채널과 잡음에 강하도록 DWFBA (Direct Weighted Filter Bank Analysis)와 같은 특징 파라미터를 추출하는 방법이 연구되기도 했다[6]. 본 연구에서는, 전화 음성 연속 숫자음의 인식 성능을 향상시키는 방법을 찾기 위한 기초연구로, 기존 인식 시스템에서 얻어지는 인식 결과를 바탕으로 1-best의 인식 결과와 2-best, 3-best의 인식 결과를 가지고 오인식의 유형과 오인식이 발생했을 경우의 앞뒤 숫자음 분포 및 N-best 결과를 가지고 오인식이 발생했을 경우의 로그 유사 확률값의 분포를 분석하였다. 이는 전화 음성 연속 숫자음 인식에서 주로 발생하는 오인식의 문제를 보다 구체적으로 검토하여 인식 성능을 향상시키는데 기초 자료로 활용하기 위함이다.

특징 파라미터로는 잡음 및 채널 특성에 좋은 성능을 보인 DWFBA 방식의 38차 MFCC (Mel Frequency Cepstrum Coefficient)를, 채널 보상 기법으로는 켑스트럼 영역에서 전체 음성의 평균과 분산으로 정규화하는 MRTCN (Modified Real Time Cepstrum Normalization) 방식을 적용하여 SiTEC (Speech Information Technology and Industry Promotion Center)에서 배포한 전화 음성 4연숫자음 2000명 화자의 DB (DataBase)에 대해 실험한 결과를 이용하였다. 인식 실험 결과 1-best에서는 91.52%의 문장인식률, 즉 4연숫자음 인식률을 얻었으며 4자리 중 1자리를 틀린 4연숫자음이 가장 많았으며, 오인식이 발생했을 경우 앞뒤의 묵음이나 특정 숫자음의 조음현상에 기인한 경우가 많았다. 또한, N-best결과에서는 "일"과 "이", "일"과 "칠", "이"와 "일", "오"와 "구"가 가장 많이 발생하는 오인식 쌍임이 확인되었다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 전화 음성의 인식 실험에 많이 사용되는 기법들에 대해 설명하고, 3장에서는 본 연구에서 사용된 인식 실험 환경에 대해 기술한다. 4장에서는 실험 결과 및 그 분석 결과를 제시하고, 5장에서 결론을 맺는다.

2. DWFBA와 MRTCN

본 논문의 전화 음성 연속 숫자음 인식에서 사용된 특징추출 방법인 DWFBA 방식과 채널 보상 기법인 MRTCN 방법에 대해 설명한다. 그리고 실험에 사용된 전화 음성 DB에 대해 기술한다.

2.1. DWFBA

DWFBA는 켑스트럼이 채널 및 주변 잡음의 간섭에 강인하도록 하기 위해 log 필터 뱅크 에너지의 높은 에너지 부분을 강조해 주는 것인데, MFCC 추출과정 중 DCT (Discrete Cosine Transform) 전에 임계대역의 로그 에너지에 비례하도록 하는 가중함수를 곱하여 특징 파라미터를 추출하는 방법이다[6]. 관련 수식은 식 (1), (2)와 같다. 여기서 L과 Q는 특징 파라미터의 차수, 임계대역의 수를 나타내며, $e(i)$는 i번째 임계대역의 에너지이다.

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

$$(2) \quad w(i) = \frac{\log[e(i)+1.0]}{\sum_{i=1}^{Q}\log[e(j)+1.0]}$$

2.2. MRTCN

채널 보상 기법은 전화 채널 특성이 켑스트럼 영역에서 합의 형태로 나타나고 그 특성이 단시간에 큰 변화가 생기지 않고 거의 일정하기 때문에 전화 채널의 변화 특성을 켑스트럼의 평균으로 보고 전체 켑스트럼의 평균값을 빼 줌으로써 채널에 의한 영향을 줄여 주는 기법이다. 본 논문에서 적용한 MRTCN은 4연숫자음 마다 켑스트럼을 구해 전체 켑스트럼의 평균을 추정하고 전체 켑스트럼의 분산도 같은 방식으로 추정하여 켑스트럼 영역에서 정규화해 줌으로써 채널에 의한 영향을 줄여 주는 보상기법으로 식 (3)~(5)로 표시된다[5]. 여기서, $c_p(t)$는 왜곡

된 p번째 음성신호에서 t번째 프레임의 켑스트럼 벡터, $\hat{C}_p(t)$는 보상기법이 적용된 p번째 음성신호의 t번째 프레임에 대한 켑스트럼 벡터를 나타낸다. 따라서 p는 음성신호의 인덱스이고, t는 프레임 인덱스가 된다. $\hat{m}_p$와 $\hat{V}_p$는 p번째 음성구간에서 추정된 켑스트럼의 평균 벡터와 분산 벡터를 나타내며, V_p 는 p번째 음성구간의 분산 벡터이다. MRTCN의 적용은 식 (5)와 같이 이루어진다. 여기서 α는 전체 추정 계수이다.

$$\hat{m}_p = (1-\alpha)\hat{m}_{p-1} + \alpha\left[\frac{1}{T}\sum_{t=1}^{T} c_p(t)\right] \quad (3)$$

$$\hat{V}_p = (1-\alpha)\hat{V}_{p-1} + \alpha V_p \quad (4)$$

$$\hat{C}_p(t) = [c_p(t) - \hat{m}_p]/\sqrt{\hat{V}_p} \quad (5)$$

3. 인식 실험

음성인식 실험에서 음향모델의 생성 및 훈련과 인식은 HTK (Hidden Markov ToolKit)를, 특징 파라미터 추출은 직접 구현한 프로그램을 사용하였다[9]. 연속 숫자음 인식 시스템을 구현하기 위해 <표 1>과 같이 개별 숫자음 11개와 묵음 모델을 포함하는 총 17개의 유사 음소를 정의하였다. 음향모델로는 CHMM (Continuous Hidden Markov Model)을 사용했으며 트라이폰(triphone) 단위로 인식 실험을 하였다. 트라이폰 모델을 사용함에 따라 훈련시켜야할 모델의 수가 증가하므로 모델별 훈련 데이터의 수를 확보하기 위해 TBC (Tree Based Clustering) 기법을 이용하였다[9,10]. HMM의 상태수는 5개, 상태당 mixture의 수는 9개로 정하여 실험하였으며, 언어모델로는 연속 숫자음은 개별 숫자음들의 나열이므로 비교적 간단한 언어모델인 FSN (Finite State Network)을 적용하였다.

실험에서 사용한 음성 DB는 SiTEC에서 제작한 전화 음성 4연숫자음 DB이다. 전화 음성은 총 2000명 화자의 음성으로 이루어졌으며 유선/무선 전화 및 cellular, PCS 전화망 환경에서 녹음되어 linear PCM (Pulse Code Modulation)으로 저장되어 있다. 녹음된 4연숫자음의 종류는 1620개이며, 화자당 32개 정도의 4연숫자음으로 구성되어 있다. 특히, 숫자 6의 경우 "륙"과 "육"으로 구분하여 레이블링 되어있다. 음성 DB는 훈련용으로는 1800명의 58388개, 테스트용으로는 200명의 6468개의 4연숫자음으로 구성되어 있다[8]. DB에서 테스트용 4연숫자음은 각 전화망 환경에서 남녀 화자수가 각각 25명으로 구성되어 있다.

전화 음성 분석프레임의 길이는 20 ms, 프레임의 이동은 10 ms로 하였다. 전처

리 계수를 0.97로 하였으며 프레임별로 해밍윈도우 처리하였다. 프레임별로 1차의 에너지와 12차의 DWFBA를 구했으며, 총 13차의 특징 파라미터에 MRTCN 보상기법을 적용한 후 delta 및 delta-delta 특징 파라미터를 구해 총 38차의 특징 파라미터로 인식 실험을 하였다. MRTCN 적용시 전체 켑스트럼 평균과 표준편차 추정계수로는 0.125를 사용하였다. SiTEC 전화 음성 DB의 구성에 따라 훈련 데이터로는 58388개, 테스트 데이터로는 6468개의 4연숫자음을 사용하였다.

<표 1> 기본 유사 음소

기 호	음 소	기 호	음 소
g	ㄱ(초성)	l	ㄹ(초성)
s	ㅅ(초성)	nge	ㅇ(종성)
ch	ㅊ(초성)	me	ㅁ(종성)
p	ㅍ(초성)	ge	ㄱ(종성)
o	ㅗ	le	ㄹ(종성)
i	ㅣ	yu	ㅠ
a	ㅏ	u	ㅜ
sil	묵음	yeo	ㅕ
		sp	짧은묵음

4. 인식 결과

인식 실험 결과 1-best 결과에서는 6468개의 4연숫자음 중에서 5919개를 바르게 인식하여 91.52%의 문장인식률을 보였다. 2-best 결과와 3-best 결과에서는 각각 253개, 102개의 바르게 인식된 4연숫자음이 포함되어 95.43%, 97.01%의 문장인식률을 보였다.

4.1. 1-best 결과 분석

1-best 결과 91.52%의 문장인식률을 얻었는데, 오인식된 연속 숫자음에서 개별 숫자의 오인식 분포를 보면 <표 2>와 같다. 4개의 숫자중에 1개만 오인식된 경우가 437개로 오인식 전체의 약 80%를 차지함을 볼 수 있다. 또한, <표 3>에서 SiTEC DB에서 각 숫자음의 빈도는 비슷하나 오인식된 수를 보면 숫자음에서 오

인식이 잘되는 숫자음 "일", "이", "오"의 오인식률이 특히 높음을 알 수 있다. 이러한 숫자음들은 특정 숫자음으로 오인식되는 경우가 많은데 "일"은 "이"이나 "칠"로, "이"는 "일"로, "오"는 "구"로 오인식되는 경우가 많다.

다음으로 오인식이 발생되었을 경우 앞뒤 숫자음에 의한 영향에 대해 알아보기 위해 발생한 오인식 숫자음 중 발생 횟수가 많은 오인식에 대해 앞뒤 숫자음을 <표 4>에 나타내었다. 숫자음의 앞뒤에 묵음이 오는 경우 그 숫자음은 오인식될 확률이 높음을 볼 수 있다. 이는 묵음 부분에서 음성 부분으로 전환할 때 화자의 발성 습관에 따라 신호의 특성이 급격하게 바뀌는 경우가 많을 수 있고, 마지막 숫자음을 발성할 때에는 끝을 흐려 발음하는 경우가 많아서 오인식이 많이 발생하는 것으로 생각된다. 앞뒤 묵음의 영향을 줄이기 위해서 끝점 검출 기법을 이용한다면 4연숫자음의 인식률은 향상될 것으로 사료된다. 연속 숫자음 발음시에는 조음효과에 의해 숫자음 "오"의 경우, 종성 자음 'ㄱ'을 가지는 단어 뒤에서는 오인식률이 높음을 볼 수 있다.

<표 2> 오인식된 개별 숫자 수에 따른 에러 분포

4연숫자음에서 오인식된 개별 숫자 수	1개	2개	3개	4개
오인식된 4연숫자음 수	437 개	69 개	24 개	18 개

<표 3> SiTEC DB에서 각 숫자음별 빈도와 오인식된 수
(숫자음 빈도에서 차지하는 비율)

개별 숫자음	숫자음의 빈도수	오인식한 수(개)
영	2175	40(1.8%)
일	2186	131(6%)
이	2198	100(4.5%)
삼	2149	27(1.2%)
오	2216	147(6.6%)
others	.....	.....

전화 채널별로 오인식된 4연숫자음의 수를 <표 5>에 나타내었다. 표에서 유선전화 음성에 비해 다른 전화 채널에서의 음성에 오인식된 수가 많음을 알 수 있다. 이는 유선전화망에 비해 다른 전화망의 채널특성이 더 불완전하여 채널 보상 기법을 적용해도 채널의 영향이 남아있기 때문이다. 또한 성별에 따른 오인식된 4연숫자음 수를 보면 남성에 비해 여성의 오인식률이 더 높음을 알 수 있다.

<표 4> 오인식 유형 발생시 앞뒤 숫자음 분포

오인식 유형	앞 숫자음 (갯수)	뒤 숫자음 (갯수)	오인식 유형	앞 숫자음 (갯수)	뒤 숫자음 (갯수)
이→일	일(9) (18.7%)	SIL(12) (25%)	일→칠	SIL(11) (47.8%)	삼(4) (17.3%)
	오(8) (16.7%)	사(8) (16.7%)		삼(4) (17.3%)	영(3) (13%)
	칠(6) (12.5%)	칠(5) (10.4%)		육(2) (8.7%)	칠(2) (8.7%)
	SIL(5) (10.4%)	영(5) (10.4%)		공(2) (8.7%)	
일→이	SIL(9) (29%)	SIL(10) (32.2%)	오→구	SIL(13) (27%)	구(7) (14.6%)
	오(5) (15.1%)	오(4) (12.9%)		육(8) (16.7%)	륙(7) (14.6%)
	공(3) (9.7%)	육(3) (9.7%)		륙(7) (14.6%)	SIL(5) (10.4%)
	이(3) (9.7%)	륙(3) (9.7%)		팔(4) (8.3%)	영(4) (8.3%)

<표 5> 전화 채널에 따른 오인식된 4연숫자음의 수(개)

전화망 환경	남성 화자	여성 화자	합 계(%)
유선전화	48	57	105(19.19)
무선전화	47	76	123(22.49)
PCS	64	92	156(28.52)
Cellular	92	71	163(29.80)
전체	251	296	547

4.2. 2-best와 3-best의 결과 분석

1-best와 2-best결과를 이용해 오인식 유형이 발생하였을 경우 두 결과에서의 로그 유사 확률값의 차이를 알아보았다. 인식 결과 중에서 1-best에서 틀리고 2-best에서 맞은 4연숫자음에 대해 로그 유사 확률값의 차이를 조사해 보면 <표 6>과 같다. 표에서 보면 "이→일", "오→구", "일→이", "일→칠"의 합이 전체 오인식의

50% 이상을 차지함을 알 수 있다. 이는 이들의 오인식만 특징 파라미터 추출 단계에서 줄일 수 있다면 큰 인식 성능을 가져옴을 나타낸다. 마찬가지로 3-best 결과를 분석해 보면 오인식 유형의 발생 횟수는 2-best의 결과와 비슷하지만 "일→칠"의 유형 대신 "육→영" 유형이 많이 발생함을 알 수 있었다. "육"이 "영"으로 오인식되는 경우는 숫자음들 사이에서 "융"으로 발음되는 경우가 많기 때문으로 생각된다.

<표 6> 2-best와 1-best의 오인식 유형에 따른 분석

오인식 유형	2-best에서 맞은 갯수 (%)	오인식된 수	평균 확률값	표준 편차
이→일	48 (18.97%)	100	132	144
오→구	48 (18.97%)	147	231	181
일→이	31 (12.25%)	131	225	239
일→칠	23 (9.09%)	131	247	203

5. 결 론

전화 음성 연속 숫자음의 인식은 전화 채널에 의한 영향과 연속 숫자음의 조음현상으로 인식률을 높이는데 어려움이 있다. 특히 "일"과 "이", "일"과 "칠", "이"와 "일", "오"와 "구" 같은 오인식 쌍들에 의한 인식률 저하가 크게 나타난다. 본 연구에서는 이 문제를 극복하기 위한 기초연구로 SiTEC 전화 음성 DB를 이용한 4연숫자음 인식 실험 결과에서 1-best와 N-best의 결과를 분석하고 오인식 유형에 따른 로그유사도의 차이를 알아보았다. 1-best 결과에서는 4연숫자음 중에서 1개 숫자가 오인식된 경우가 가장 많이 발생하였으며, 앞뒤의 묵음에 의한 영향과 앞뒤 숫자와의 조음현상이 오인식에 가장 큰 영향을 미치고 있음을 알 수 있었다. 오인식이 발생한 경우, 2-best와 3-best의 로그유사 확률값 차이의 결과를 1-best의 결과와 비교해 보았다. 로그유사 확률값 차이의 평균값과 표준편차가 작다는 의미는 특징 파라미터 추출부분에서 어느 정도 보상을 해주거나 오인식 쌍들의 구별도를 높이는 새로운 특징 파라미터를 적용하면 제대로 인식할 수 있다는 것을 의미한다. 본 연구결과를 참고로 하여 향후 특징 파라미터 추출 단계에서 오인식되는 숫자들의 변별력을 높여줄 수 있는 방법에 대한 연구를 수행할 계획이다.

참 고 문 헌

[1] P. J. Moreno, "Speech Recognition in Telephone Environment", Master Thesis, Carnegie Mellon University, 1992.

[2] 김성탁, 김상진 et al., "전화망 환경에서의 연속 숫자음 인식 성능평가", *한국음향학회 논문집*, 21권, 1호, pp.253-256, 2002.

[3] 김승희, 김형순, "음향학적 파라메터를 이용한 한국어 연결숫자인식의 성능개선", *한국음향학회 논문집*, 18권, 1호, pp.44-47, 1999.

[4] J. D. Veth, L. Boves, "Comparision of Channel Normalization Tech-nique for Automatic Speech Recognition Over the Phone", *Proc. ICSLP*, pp.2332-2335, 1996.

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

[6] 정성윤, 김민성 et al., "한국어 연속 숫자음 전화 음성의 인식 성능개선", *대한전자공학회 추계학술발표대회 논문집*, 25권, 2호, pp.582-585, 2002.

[7] 최종연구보고서, "전화망 환경에서의 연속 숫자음 신호왜곡 연구", 한국전자통신연구원, 2002.

[8] http://www.SiTEC.or.kr/index.asp.

[9] S. Young, G. Evermann, D. Kershaw, *The HTK Book (for HTK Version 3.1)*, Cambridge University Engineering Department.

[10] http://inc2.ucsd.edu/~owkwon/srhome/lecture/robust_speech_recog.html

접수일자: 2003년 05월 24일
게재결정: 2003년 06월 12일

▶ 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

▶ 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

▶ 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

▶ 김상훈(Sang-Hun Kim)
주소: 702-701 대전광역시 유성구 가정동 161번지 한국전자통신연구원 컴퓨터 소프트웨어 연구소 음성언어정보연구센터 음성DB연구팀
소속: 음성DB연구팀
전화: 042) 860-5141
E-mail: ksh@etri.re.kr