원거리 음성인식을 위한 MLLR적응기법 적용

권석봉(ICU), 지미경(ICU), 김회린(ICU), 이용주(원광대)

<차 례>

1. 서론
2. 환경보상 기법
 2.1. 신호 보상
 2.2. 특징 보상
 2.3. MLLR 기반의 모델보상
3. 실험 및 결과
 3.1. Database
 3.2. 실험결과
4. 결론

<Abstract>

MLLR-Based Environment Adaptation for Distant-Talking Speech Recognition

Suk-bong Kwon, Mikyong Ji, Hoi-Rin Kim, Yong-Ju Lee

Speech recognition is one of the user interface technologies in commanding and controlling any terminal such as a TV, PC, cellular phone etc. in a ubiquitous environment. In controlling a terminal, the mismatch between training and testing causes rapid performance degradation. That is, the mismatch decreases not only the performance of the recognition system but also the reliability of that. Therefore, the performance degradation due to the mismatch caused by the change of the environment should be necessarily compensated. Whenever the environment changes, environment adaptation is performed using the user's speech and the background noise of the changed environment and the performance is increased by employing the models appropriately transformed to the changed environment. So far, the research on the environment compensation has been done actively. However, the compensation method for the effect of distant-talking speech has not been developed yet. Thus, in this paper we apply MLLR-based environment adaptation to compensate for the effect of distant-talking speech and the performance is improved.

* Keywords: Distant-talking speech recognition, Environment compensation, Model compensation

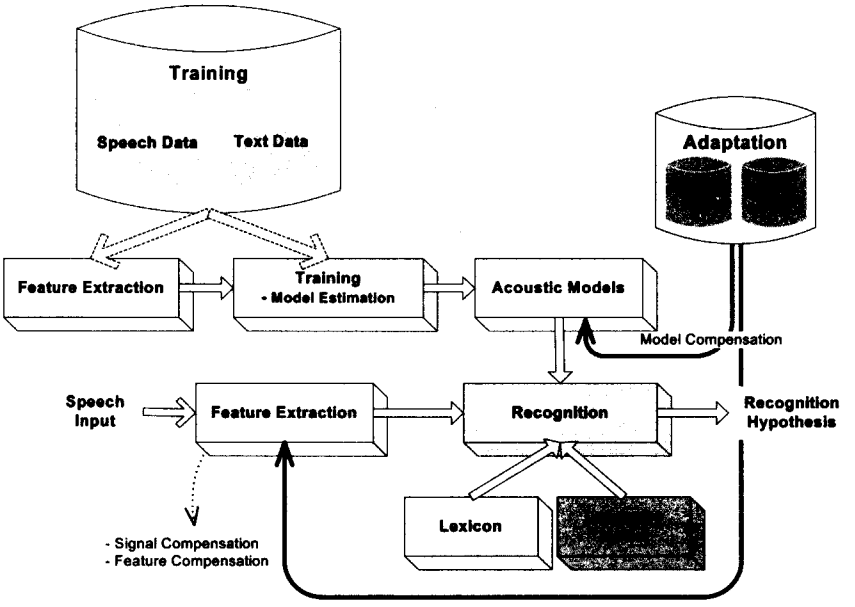
1. 서 론

임의의 단말기를 음성으로 제어하는 원거리 음성인식 시스템은 사용에 용이하고 정확해야 할 뿐 아니라 실제상황에서 잘 동작해야 한다. 이 중 신뢰도 높은 원거리 음성인식 시스템을 위해 가장 중요한 요소는 robustness이다. 대부분의 경우, 실제 음성인식 시스템이 사용될 환경이 학습 환경과 다르기 때문에 원거리 음성인식의 성능저하가 현저하다[1]. 특히 마이크와 화자 사이의 거리가 먼 원거리 음성입력의 경우, 입력신호 외에 주변잡음, 마이크, 채널왜곡 등에 의해 실제 주변환경의 급격한 변화가 발생하여 시스템의 급격한 성능저하를 초래하게 된다. 즉 원거리로부터 음성을 입력하는 과정에서 입력신호 외의 다른 소리가 존재할 수 있어 인식률을 저하시키기 때문에 잡음을 제거하여야 하며 시스템의 마이크가 먼 거리에 있고, 사방에 소음이 많은 경우 사람은 평상시 보다 힘을 주어 발화하게 되는데 이러한 상황은 훈련 상황과 인식 상황을 다르게 하기 때문에 인식률을 저하시킬 수 있다(Lombard effect). 따라서 실제 환경에서 잘 동작하는 원거리 음성인식을 위해 변화된 환경의 영향을 배제해 주는 환경보상이 요구된다. 즉, 신뢰도 높은 원거리 음성인식 시스템을 위해 변화된 환경 또는 화자의 영향을 배제해 주는 환경보상 기법이 필수적이다. 본 논문에서는 신뢰도 높은 원거리 음성인식 시스템을 위해 MLLR(Maximum Likelihood Linear Regression)[2] 기반의 모델보상 기법을 적용하여 원거리 효과를 제거함으로써 원거리 음성인식을 성능을 향상시켰다.

2. 환경보상 기법

실제 음성인식 시스템이 사용될 환경은 인식 시스템을 학습시킬 때의 상황과 다를 수 있다. 음성입력 시 입력신호 외에 주변잡음, 마이크, 채널 등의 영향 등이 전체 인식 시스템의 성능을 크게 좌우한다. 이러한 학습상황과 인식상황의 불일치로 인해 발생하는 문제를 해결하기 위해 실제 시스템이 사용될 모든 환경에 대해 음성인식 시스템을 학습시키는 것은 현실적으로 불가능하다. 따라서 인식 시에 이런 환경의 영향을 배제하는 과정을 환경보상이라고 한다. 환경보상 방법은 분류하는 방법에는 화자적응과 같이 인식모델을 새로운 환경에 적응시키는 방법과 수학적 모델을 이용하여 환경의 영향이 배제된 음성신호를 추정하거나 또는 음성특징을 추정하는 방법으로 나눌 수 있고, 환경보상 방법이 적용되는 단계에 따라 신호단계 보상, 특징단계 보상과 모델단계 보상으로 나눌 수 있다. 본 논문에서는 모델 단계에서 적용되는 MLLR 기반의 환경보상 기법을 적용하여 원거리 효과를 제거함으로써 원거리 음성인식에 대한 성능을 향상시키려 하였다. <그림 1>은 음성

인식 과정 및 환경보상에 관한 블록다이어그램이다.



<그림 1> 인식과정 및 환경보상에 관한 블록도

2.1 신호 보상

원거리에서 음성을 입력하는 과정에서 입력신호 외에 다른 소리(잡음)가 존재할 수 있고 이로 인한 학습상황과 실험상황의 차이가 발생하여 음성인식의 성능을 저하시킬 수 있다. 입력신호 외의 신호 즉 잡음을 신호 단계에서 제거함으로써 인식성능을 향상시킨다. Aurora-front end의 Wiener 필터[3] 등이 대표적이다.

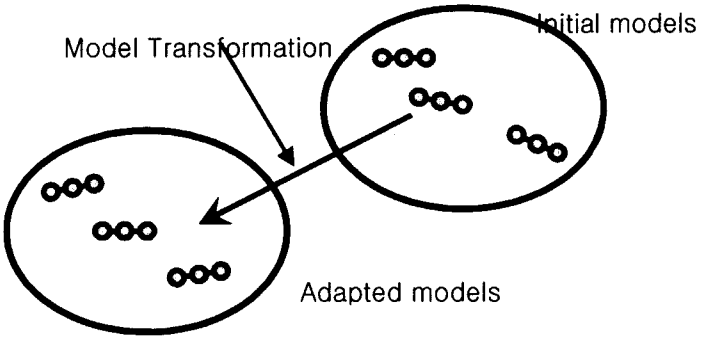
2.2 특징 보상

원거리 음성입력 과정에서 발생하는 학습상황과 인식상황 사이의 불일치를 인식과정 전에 전처리 단계에서 잡음환경에 강인한 특징벡터를 추출하거나 채널왜곡에 의한 영향을 보상하여 주는데 효과적이다. CMN(Cepstral Mean Normalization), SS(Spectral Subtraction), RASTA(RelAtive SpecTrAL) 등의 기법들이 대표적이다[4].

2.3 MLLR 기반의 모델보상

모델보상이란 인식환경이 학습 환경과 다른 환경으로 변화하여 발생하는 성능저하를 모델단계에서 보상하여 주는 방법이다. 모델보상 방법으로는 MLLR[2],

PMC(Parallel Model Compensation)[5], SM(Stochastic Matching)[6] 기법 등이 있다. 본 논문에서는 MLLR 기반의 모델보상 기법을 이용하여 상대적으로 적은 데이터 양을 사용하여 새로운 환경(원거리 음성)에 적응하여 새로운 환경에 적합한 음성 모델을 만들어 빠르고 효과적으로 환경을 보상해 줌으로 성능을 향상시켰다. 본 논문에서는 PBW 및 POW DB를 사용하여 초기모델을 생성하고 거리(30, 100, 200, 300 cm) 별로 수집된 데이터의 일부 적은 양을 사용하여 거리보상이 된 새로운 음성모델을 만들어 나머지 데이터에 대해 그 성능을 평가하였다. 아래의 <그림 2>는 MLLR 기반의 모델보상에 관한 블록다이어그램이다. 초기 모델을 환경적응 용 DB(원거리 DB)를 사용하여 환경에 적합하도록 모델을 변환시켜 환경 적응된 모델을 생성하였다.



<그림 2> MLLR 기반의 모델보상

3. 실험 및 결과

3.1. Database

원거리 음성인식을 위한 훈련 DB는 PBW(16 kHz) 및 POW(16 kHz) DB를 8 kHz로 다운 샘플링 하여 사용하였다. 원거리 DB(16 kHz)는 8 kHz로 다운 샘플링 하여 그 일부를 MLLR 기반의 거리보상 기법의 적응용 DB로 사용하였고 나머지를 성능평가를 위해 사용하였다. 원거리 DB는 음악이나 전화벨, 선풍기 등의 비음성 잡음과 음성 잡음이 어느 정도 존재하는 환경에서 수집되었다. 전체 19명의 화자가 46개의 단어를 한번씩 발성하였고 거리 별로 설치된 마이크에 의해 동시에 녹음되었다. 총 3,496 개(4(거리: 30, 100, 200, 300 cm) x 19(화자 수) x 46(단어 수))의 발화를 환경적응 및 평가용 DB로 사용하였다. 성능 평가를 위해 사용된 원거리 DB의 전체 어휘목록은 <표 1>과 같다.

<표 1> 원거리 음성인식 성능평가를 위해 사용된 전체 어휘 목록

번호	단어목록	번호	단어목록
1	예	24	저장
2	아니오	25	크게
3	취소	26	작게
4	확인	27	빨리
5	종료	28	느리게
6	처음으로	29	이리와
7	이전	30	저리가
8	다음	31	앞으로
9	반복	32	뒤로
10	음성시작	33	돌아
11	음성끝	34	서
12	대기	35	이메일
13	열기	36	전자우편
14	찾기	37	인터넷
15	아래로	38	날씨
16	위로	39	뉴스
17	왼쪽으로	40	교통정보
18	오른쪽으로	41	주식시세
19	재생	42	일정
20	중지	43	스케줄
21	추가	44	전화
22	삭제	45	영화
23	변경	46	스포츠결과

3.2. 실험결과

원거리 음성인식 시스템의 거리보상을 위해 MLLR 기반의 환경보상 기법을 적용하고 그 인식성능을 HTK를 사용하여 평가하였다. 거리보상을 위한 적응용 DB의 양을 달리하였을 때와 인식대상의 어휘를 좁혀 그 성능을 평가하고 비교하였다. <표 2>의 실험 1은 MLLR 기반의 거리보상 기법 적용하여 46개 단어의 거리별 인식성능을 보여주고 있다. MLLR 기반의 거리보상 기법을 적용하기 위해 전체 19명의 화자 DB 중 1명의 화자 분인 46 단어 발화(한 번씩 발성) 분을 적응용 DB로 사용하였고, 나머지 18명의 DB를 인식성능 평가용으로 사용하였다. MLLR 기반의 거리보상을 했을 경우 baseline보다 높은 성능을 보여주고 있으며, 거리가 30cm일 때 가장 높은 성능향상을 보이고 있고, 100cm일 때 가장 낮은 성능향상을 보이고 있다.

<표 2> MLLR보상기법을 적용한 성능평가 (실험1, 실험2)

종류 / 거리(cm)	Baseline	MLLR 거리보상 - 실험1*	MLLR 거리보상 - 실험2*
30	89.45	93.12 (34.79)	89.49 (0.38)
100	78.89	83.31 (20.94)	78.23 (-3.13)
200	69.47	79.08 (31.48)	73.64 (13.66)
300	67.40	77.54 (31.10)	72.19 (14.69)

실험1* : MLLR 기반의 거리보상 방법 적용 (19명의 화자 중 1명의 화자 분, 46개의 단어를 적응용 DB로 나머지 화자의 DB를 성능평가용으로 사용하였음.

실험2* : MLLR 기반의 거리보상 방법 적용 (19명의 화자 중 1명의 화자를 선택하고 인식대상의 모든 음소를 포함하는 14개의 단어를 선택하여 적응용 DB로 나머지 화자의 DB를 성능평가용으로 사용하였음

( ) : ERR(Error Reduction Rate)

<표 2>의 실험 2는 MLLR 기반의 거리보상 기법에서 보상을 위한 적응용 DB 어휘 목록과 인식 DB의 어휘 목록을 달리하였을 때의 성능이다. 평가대상의 어휘에 포함된 모든 음소를 포함하도록 14개의 어휘(<표 3>)를 선택하여 1명의 화자 분에서 뽑아낸 DB로 환경적응을 수행하고 나머지 18명 화자 분으로 성능을 평가하였다. 평가대상의 어휘 중 14개만을 적응용 DB로 사용하였을 때 큰 폭의 성능 하락을 보였다. 그러나 거리가 멀어질수록 거리 보상 효과가 크다는 것을 알 수 있다. <표 2>의 실험 1과 실험 2를 통해 적응용 DB가 적을 때는 원거리의 인식 성능향상이 좋고, 적응용 DB가 증가함에 따라 근거리 및 원거리에 대한 인식 성능향상이 고루 증가함을 알 수 있다.

<표 3> 거리보상 성능평가에 사용된 14개 어휘 목록

번호	단어목록	번호	단어목록
1	예	32	뒤로
4	확인	39	뉴스
6	처음으로	40	교통정보
11	음성끝	41	주식시세
17	왼쪽으로	43	스케줄
19	재생	45	영화
27	빨리	46	스포츠결과

<표 4>의 실험 3은 실험 2에서 선택된 14개의 단어(<표 3>)를 적응 및 인식 어휘 대상으로 하여 인식성능을 평가하였다. 인식대상의 어휘 14개만을 대상으로 baseline 시스템의 인식성능을 평가하였고, 한 사람의 발성 분에서 <표 3>의 어휘

14 개를 사용하여 MLLR 기반의 거리보상을 수행하였다. 표에서 보는 바와 같이, 앞선 실험들과 같은 결과를 볼 수 있다. 실험을 통해 MLLR 기반의 환경보상 기법으로 원거리 효과를 효과적으로 보상할 수 있었다.

<표 4> 실험 2에서 선택된 어휘 14개만을 대상으로 한 MLLR기반의 거리보상 기법의 성능비교 (실험 3)

종류 / 거리(cm)	Baseline	MLLR 거리보상 - 실험3*
30	96.62	96.83 (6.21)
100	89.31	92.06 (25.72)
200	82.10	87.30 (29.05)
300	84.03	86.51 (15.53)

실험3* : MLLR 기반의 거리보상 방법 적용 (19명의 화자 중 1 명의 화자 분의 DB에서 실험 2에서 선택된 14개의 단어를 이용하여 거리적응으로 사용하였고 나머지 화자의 해당 어휘를 성능평가용으로 사용하였음

( ) : ERR(Error Reduction Rate)

4. 결 론

본 논문에서는 원거리 음성인식 시스템의 문제점인 실제 인식환경의 변화에 따른 학습 환경과 인식환경의 불일치에 의한 성능저하 현상을 MLLR 기반의 모델 보상기법에 의해 환경변화를 보상해줌으로써 전체적인 원거리 인식 시스템의 성능을 평가하였고, 적응 데이터를 달리하여 그 성능을 평가하였다. 실험 결과, MLLR 기반의 거리보상에 대해 성능향상이 컸다. 즉, Baseline 시스템에 MLLR 기반의 거리보상 기법만 적용하여도 실제상황에서의 환경변화를 효과적으로 제거할 수 있다는 것을 보여주고 있다.

향후에는 다양한 환경에서 어휘를 수집하여 인식 전체 과정에 대해 다각도의 시뮬레이션을 통해 원거리 음성인식의 성능을 평가하고자 한다. 또한 MLLR 기반의 환경보상 기법을 보완하여 효과적으로 환경변화를 보상할 수 있도록 할 것이며 원거리 음성의 특징에 관한 조사를 지속할 것이다. 또한 원거리 음성인식의 성능을 높이기 위한 다른 방법으로 여러 개의 마이크로부터 들어오는 입력 또는 그것들의 인식결과를 효율적으로 통합할 수 있는 방법을 모색할 것이다.

감사의 글

본 연구는 한국과학기술부 기초과학연구사업 중 지방연구중심대학육성사업인 헬스케어기술개발사업단의 지원에 의해 수행되었으며 이에 감사드립니다.

참고문헌

[1] S. Molau, D. Keysers, H. Ney, "Matching training and test data distributions for robust speech recognition", *Speech Communication*, vol. 41, pp. 579-601, Nov. 2003.

[2] S. Doh, R. Stern, "Inter-class MLLR for speaker adaptation", *in Proc. ICASSP*, vol. 3, pp. 1543-1546, June 2000.

[3] "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithms", Tech. Rep. Standard EST201 108, European Telecommunications Standards Institute (ETSI), Apr. 11 2000.

[4] 정성윤, 손종목 외, "한국어 숫자음 전화음성의 채널왜곡에 따른 특징파마리터의 변이 분석 및 인식실험", *대한음성학회 말소리*, 제 43호, pp. 179-188, 2002.

[5] M. J. F. Gales, S. J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination", *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 5, pp. 352-359, Sep. 1996.

[6] C. Lee, "Adaptive Compensation for Robust Speech Recognition", *Proc. Automatic Speech Recognition and Understanding*, pp. 14-17, Dec. 1997.

접수일자: 2005년 2월 10일
게재결정: 2005년 3월 15일

▶ 권석봉(Suk-bong Kwon)
주소: 305-732 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 음성인식기술연구실
전화: 042) 866-6221
FAX: 042) 866-6245
E-mail: sbkwon@icu.ac.kr

▶ 지미경(Mikyong Ji)
주소: 305-732 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 음성인식기술연구실
전화: 042) 866-6221
FAX: 042) 866-6245
E-mail: lindaji@icu.ac.kr

▶ 김회린(Hoi-Rin Kim)
주소: 305-732 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 음성인식기술연구실
전화: 042) 866-6139
FAX: 042) 866-6245
E-mail: hrkim@icu.ac.kr

▶ 이용주(Yong-Ju Lee)
주소: 570-749 전북 익산시 신용동 344-2 원광대학교
소속: 원광대학교 전기전자 및 정보공학부
전화: 063) 850-7451
FAX: 063) 850-7454
E-mail: yjlee@wonkwang.ac.kr