말소리와 음성과학 제1권 제4호 (2009)
pp. 119~125

차량 잡음 환경에서 인위적 왜곡 음성을 이용한 Eigenspace-based MLLR에 기반한 고속 화자 적응

Fast Speaker Adaptation Based on Eigenspace-based MLLR Using Artificially Distorted Speech in Car Noise Environment

송 화 전[1] · 전 형 배[2] · 김 형 순[1]
Song, Hwa Jeon · Jeon, HyungBae · Kim, Hyung Soon

ABSTRACT

This paper proposes fast speaker adaptation method using artificially distorted speech in telematics terminal under the car noise environment based on eigenspace-based maximum likelihood linear regression (ES-MLLR). The artificially distorted speech is built from adding the various car noise signals collected from a driving car to the speech signal collected from an idling car. Then, in every environment, the transformation matrix is estimated by ES-MLLR using the artificially distorted speech corresponding to the specific noise environment. In test mode, an online model is built by weighted sum of the environment transformation matrices depending on the driving condition. In 3k-word recognition task in the telematics terminal, we achieve a performance superior to ES-MLLR even using the adaptation data collected from the driving condition.

Keywords: speaker adaptation, eigenspace-based MLLR, environment selection

1. 서 론

최근 차량 내 설치된 텔레매틱스 단말기나 내비게이션 단말기에 주행 중 운전자의 안전한 단말기 작동과 편의를 제공하기 위해 음성인식/음성합성 기술 등이 적용되고 있다. 그러나 차량 내에서 음성인식의 경우 운전자가 실시간으로 계속 변하는 잡음 환경에 노출되어 있으므로 모든 환경에서 제대로 동작하기가 쉽지는 않다. 실제로 대부분의 음성인식 시스템은 훈련 환경과 테스트 환경이 다른 경우 심각한 성능 저하가 나타난다. 인식성능을 저하시키는 환경 불일치 요인으로 여러 가지가 존재하지만 특히 발성환경의 차이, 화자간의 차이 등이 대표적인 환경 불일치의 예이며, 이를 제거하기 위해 다양한 방법들이 제안되었다.

발성환경의 차이의 주요 요인으로 배경잡음이나 채널왜곡 등이 있으며 이에 대해서는 특징 공간(feature space)에서 보상하는 방법이 널리 사용된다. 특징 공간에서 채널왜곡 개선방법으로 간단하지만 효과적인 켑스트럼 평균 차감법(cepstral mean subtraction; CMS)이 있으며, 부가잡음에 대한 보상으로는 스펙트럼 차감법(spectral subtraction) 등이 대표적이다. 그러나 이들은 발성한 음성이 어떤 음소열로 구성되었는지에 대한 정보를 이용하지 않는다. 반면에 각각의 음소 모델이 잡음에 의해 왜곡되는 정도를 고려하여 달리 보상해 주는 통계적 매칭(stochastic matching; SM) 방법[1]도 제안되었다.

화자간의 차이를 보상하는 방법은 주로 모델 공간에서 이루어지며, 화자적응 방법[2]-[4]이 대표적이다. 이중 데이터가 상대적으로 충분한 경우 maximum likelihood linear regression (MLLR) 적응 방법은 화자 및 환경 불일치를 동시에 보상할 수 있는 효과적인 방법이지만, 적응 데이터가 아주 적은 경우 그 성능을 보장하지 못한다[2]. 이 방법의 대안으로 eigenvoice 방법[3]을 변환행렬에 적용한 eigenspace-based MLLR (ES-MLLR) 방법[4]은 고속 화자적응에 훨씬 유리하지만, 인식환경이 훈련

1) 부산대학교 kimhs@pusan.ac.kr 교신저자
2) 한국전자통신연구원 hbjeon@etri.re.kr
본 연구는 지식경제부 및 정보통신연구진흥원의 IT 성장동력기술개발사업의 일환으로 수행하였음.[2006-S-036-04, 신성장동력 산업용 대용량/대화형 분산/내장처리 음성인터페이스 기술 개발]

접수일자: 2009년 11월 10일
수정일자: 2009년 12월 5일
게재결정: 2009년 12월 7일

과 불일치가 존재할 때는 이를 보상하지 못하며, 이러한 문제를 해결하기 위한 방법으로 [5][6]에서 바이어스 보상과 eigenvoice (또는 ES-MLLR) 화자적응을 함께 적용한 방법을 개발하였다.

그러나 환경이 지속적으로 변하는 경우에는 기존의 방법으로 성능을 향상시키기에는 한계가 있으며, 이를 극복하기 위한 방법으로 다양한 환경에 대한 모델을 구성하여 인식 상황과 유사한 환경의 모델을 선택하는 방법도 제안되었다[7]. 이 방법의 경우는 적응 데이터에 다양한 인식 환경의 일부분을 포함하도록 하여 인식환경과의 차이를 최대한 줄이고자 하였다. 하지만, 차량 내 텔레매틱스 단말기 상에서의 화자 적응 시에는 운전자의 안전을 위해 정차 중에 화자 적응을 수행해야 하는 제약이 따르며, 주행 환경을 전혀 고려하지 않게 되므로 또 다른 왜곡을 야기하게 된다.

본 논문에서는 차량 내 텔레매틱스 단말기 상에서의 음성인식기의 성능을 향상시키기 위해 다양한 차량 주행 상황의 잡음 신호를 정차 중에 모은 적응 데이터에 인위적으로 부가한 후 ES-MLLR 방식에 기반을 두어 실시간 주행 환경 선택을 통한 잡음에 강인한 고속 화자 적응 방법을 제안하였다. 실제로 적응 데이터와 인식데이터에 존재하는 다양한 잡음에 대해 먼저 특징 공간에서 기본적인 잡음 처리를 수행하여 음질 개선을 한 후 특징 파라미터를 추출하지만 인식데이터와 인위적으로 잡음을 부가하여 생성한 적응데이터 사이에는 여전히 환경 불일치가 존재하므로 이를 실시간 화자 적응 방법을 통해 불일치를 최대한 줄이고자 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 간단히 ES-MLLR 기반 화자 적응 방법에 대하여 소개한다. 그리고 3장에서는 본 논문에서 제안한 인위적 잡음 데이터를 이용한 화자 적응 방식에 대해 소개한 후, 4장에서 제안한 방법의 성능 평가를 실시하고, 마지막으로 5장에서 결론을 맺는다.

2. Eigenspace-based MLLR 기반 화자 적응

새로운 화자에 대한 MLLR에 의한 화자 적응 방법은 다음과 같이 나타낼 수 있다.

$$\hat{\mu}_m = A\mu_m + b = W\xi_m \tag{1}$$

여기서, $\mu_m \in R^{D\times 1}$은 D 차원의 화자 독립 (Speaker Independent; SI) 은닉 마르코프 모델 (hidden Markov model; HMM)의 m번째 가우시안 믹스쳐(mixture)의 평균 벡터이며, $\hat{\mu}_m$은 새로운 화자의 믹스쳐 m의 평균 벡터이다. 또한, $W = [b\ A] \in R^{D\times(D+1)}$는 변환행렬 A와 바이어스 항 b를 포함한 affine 변환행렬이고, $\xi_m = [1\ \mu_m^T]^T \in R^{(D+1)\times 1}$은 확장 SI 평균 벡터를 뜻한다. 그러나 적응 데이터가 아주 적은 경우 MLLR 방법은 그 성능을 보장하지 못하며, MLLR 방식을 고속화자 적응에서 적용하기 위해 제안된 것이 ES-MLLR이다. 이는 MLLR 변환행렬을 eigenvoice 방법에 기반을 두어 추정하는 방식이며, 적응 데이터가 아주 적은 경우에도 신뢰성 있게 변환행렬을 추정할 수 있다. 따라서 ES-MLLR 방법에서 새로운 화자는 다음과 같이 K개의 기저 행렬(basis matrix)의 가중합으로 표현할 수 있다.

$$\hat{W} = E(0) + \sum_{k=1}^{K} \hat{w}(k)E(k) \tag{2}$$

여기서 $E(0) \in R^{D\times(D+1)}$는 훈련 DB에 포함된 R명의 화자 종속(Speaker Dependent; SD) 모델의 MLLR 변환행렬들의 평균 변환행렬을 의미하며, $E(k) \in R^{D\times(D+1)}$는 k-번째 기저행렬이고, 일반적으로 $K < R$ 이다. 물론 식(2)의 기저행렬은 ES-MLLR에서 R개의 변환행렬들의 벡터화로부터 얻은 eigenvector를 편의를 위해 다시 행렬 형태로 재구성한 것이다. 그리고 $\hat{w}(k)$는 적응데이터로부터 추정할 k-번째 기저 행렬의 가중치를 뜻한다. ES-MLLR 적응 방법에서 모델의 평균 벡터는 다음과 같이 적응된다.

$$\hat{\mu}_m = \hat{W}\xi_m = E(0)\xi_m + \sum_{k=1}^{K} \hat{w}(k)E(k)\xi_m \tag{3}$$

여기서 믹스쳐 m에 대해서 $e_m(k) = E(k)\xi_m$ 이라고 하면, 이는 ES-MLLR으로 만들어진 믹스쳐 m에 대응하는 eigenvoice를 의미하며, 식(3)은 $\hat{\mu}_m = e_m(0) + \Sigma_k \hat{w}(k)e_m(k)$과 같이 eigenvoice 방식의 수식과 동일하므로 가중치는 MLED 방법[3]으로 추정할 수 있다.

식(2)에 ES-MLLR를 통해 추정한 $\hat{W}$의 형태도 $[\hat{b}\ \ \hat{A}]$로 표현된다. 그러나 변환행렬 $\hat{W}$의 바이어스 항($\hat{b}$)은 훈련에 참여한 변환 행렬로부터 얻어진 훈련 환경을 나타내는 공간에서 차원이 감소한 eigenspace만을 나타내어 줄 뿐이다. 따라서 인식환경이 훈련 환경과 다른 경우에 바이어스 보상을 적용함으로써 eigenspace 공간의 위치를 인식 환경에 가깝도록 이동시킴으로써 성능 향상을 이룰 수 있다[6]. 식(2)에 바이어스 보상 성분을 추가하여 변환행렬을 다음과 같이 확장할 수 있다.

$$\hat{W}_c = [\hat{b} + \hat{b}_c\ \ \hat{A}] \tag{4}$$

여기서, $\hat{b}_c$는 훈련 환경과 다른 인식환경을 보상하기 위한 바이어스 벡터를 뜻한다. 따라서 $\hat{W}_c$는 훈련 환경에 대해 인식환경을 보상한 변환 행렬을 뜻한다. 식(4)의 $\hat{W}_c$를 식(3)의 $\hat{W}$

대신 대입하면 환경차이에 대한 보상벡터를 포함한 ES-MLLR에 기반을 둔 적응 모델은 다음과 같다.

$$\hat{\mu}_m = \hat{W}_c \xi_m \qquad (5)$$
$$= E(0)\xi_m + \sum_{k=1}^{K} \hat{w}(k)E(k)\xi_m + \sum_{d=1}^{D} \hat{b}_c(d)i(d)$$

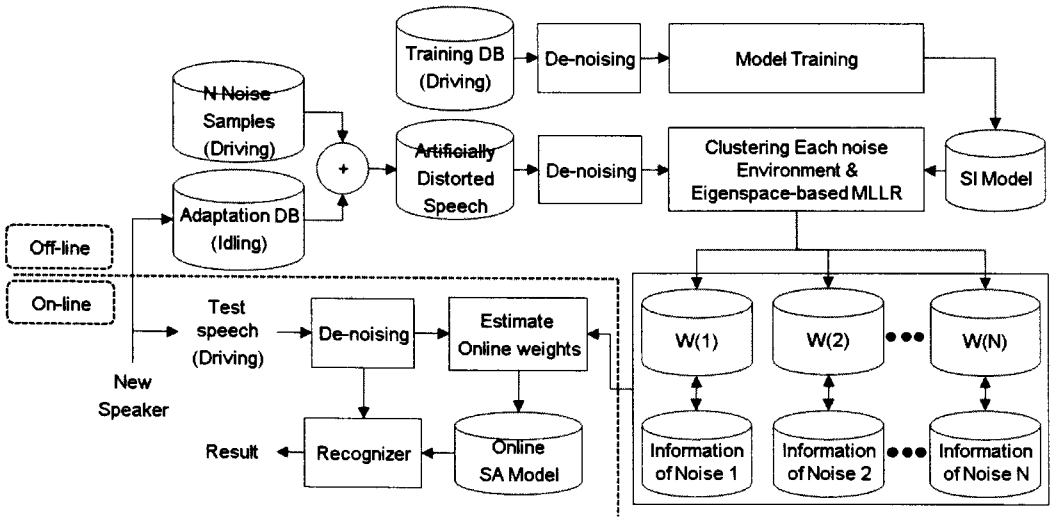

그림 1. 차량 잡음에서 인위적 적응 데이터를 이용한 고속 화자적응

Fig. 1. Fast speaker adaptation using artificially distorted adaptation data in car noise environment

여기서, $\hat{b}_c(d)$는 바이어스 보상벡터 $\hat{b}_c$의 d번째 차원의 값을 뜻하고, $i(d) = [\delta(d-1)\cdots\delta(d-D)]^T$이고, $\delta(x)$는 Kronecker delta 함수를 뜻한다. 식(5)에서는 단지 기저 벡터 수가 관측벡터의 차원인 D개만큼 증가하는 형태가 되며, 이들의 기저 행렬의 가중치와 보상벡터의 가중치는 MLED를 통하여 동시에 구할 수 있게 된다.

3. 인위적 잡음 음성을 이용한 eigenspace-based MLLR 기반 고속 화자 적응

훈련 환경과 인식 환경의 변화에 따른 보상을 위해 2절에서는 보상 벡터를 도입하였다. 그러나 인식기가 동작하는 환경이 시간에 따라 지속적으로 변한다면 적응 단계에서 사용한 환경 보상 부분이 제대로 동작하지 않게 된다. 특히, 본 논문에서 화자 적응 기술을 주행하는 차량 내에서 텔레매틱스 단말기나 또는 내비게이션 단말기에 탑재된 음성 인식기에 적용하여 성능을 향상시키기 위해서는 지속적으로 변하는 환경에 빠르게 대처해야 한다. 한 가지 방법으로는 비교사 방식(unsupervised mode)의 온라인(online) 화자 적응 방법이 있으나 차량의 주행 상태는 상대적으로 빠르게 변하며, 이전의 환경에 온라인 적응을 적용시킨 후 환경이 변한 부분으로 적응 속도를 높여 성능 향상을 달성하기가 쉽지 않다. 다른 방법으로는 [7]에서 제안한 방법을 사용할 수 있지만, 이 경우 다양한 잡음 환경에서 적응 데이터를 수집해야 하는 제약이 따르며, 또한 SI 모델을 구성하는 것도 깨끗한 DB를 사용해야 하는 제약이 따른다. 실제로 차량내 단말기에 탑재된 음성인식기의 경우 성능을 최대로 하기 위해 실제 주행상태의 차량에서 수집한 음성을 SI 모델 훈련에 사용하며, 또한 주행 중에 단말기를 통해 화자 적응을 수행하는 것도 운전자의 안전을 확보할 수가 없어 적용할 수가 없다. 따라서 차량 단말기에 화자 적응을 수행하기 위해서는 정차 중에 운전자가 화자적응을 수행하도록 해야 한다. 그러나 정차 중에 수집된 적응 데이터의 경우는 SI 모델 훈련에 사용한 데이터 및 주행 시의 환경과 다르므로 화자 적응을 통해서 인식 성능 향상을 기대하기 어렵다. 본 논문에서는 이와 같은 상황에서 화자적응 기술을 적용해 인식기의 성능을 향상시키는 방법을 제안한다.

먼저 그림 1에 차량 내에서의 화자적응의 성능을 향상시키기 위한 방법으로 본 논문에서 제안한 인위적 왜곡 음성을 이용한 화자적응 시스템의 전체 프로세스를 나타내었다. 그림 1에서 보면 우선 크게 오프라인(off-line)과 온라인(on-line) 두 가지 단계로 나눌 수 있다. 오프라인 단계에서는 주행 중 수집한 DB를 이용해 SI HMM을 만들며, 정차 중 수집한 음성에 인위적으로 다양한 잡음을 부가한 적응용 DB를 이용해 특징 공간에서 먼저 잡음 제거 기술을 적용한 후 화자적응을 거친 후 SA (Speaker Adapted) 모델을 만들게 된다. 여기서 각각의 주행 상황별 적응용 잡음 음성 DB에 대해 잡음 환경별 SA 모델을 만든다. 온라인 단계에서 실제 테스트 음성이 입력되면 잡음 제거 후 가장 유사한 환경을 선택하여 이와 관련된 모델을 이용하여 인식을 수행한다. 다음에 각각의 모듈에 대해 상세히 설명한다.

3.1 오프라인 단계

본 논문에서 제안한 방식은 [7]과 유사한 방식처럼 보이지만, 특히, 오프라인 단계에서의 개념이 완전히 다르다. 그로 인해 온라인 단계에서 모델 구성에 필요한 실시간 가중치를 구하기 위한 환경을 선택하는 부분이 달라진다. 그림에서 보는 바와 같이 차량 내에서 사용하기 위한 음성 인식기를 위해서는 훈련 데이터를 인식기의 성능을 최대로 하기 위해 실 주행 차량으로부터 음성을 수집한다. 그리고 테스터 데이터도 훈련 환경과 동일한 환경에서 운전자가 발성하게 된다. 따라서 훈련 환경과 테스터 환경의 불일치가 크지 않다. 그러나 훈련 환경과 인식 환경이 유사하지만, 적응을 수행하는 환경이 완전히 다르다. 즉, 운전자의 안전을 위해 정차 중에 적응을 수행해야 한다. 따라서 운전자에 대한 특성은 화자 적응에 의해 반영이 되지만 환경 특성은 훈련 및 주행 중의 환경과 완전히 다르므로 또 다른 왜곡을 유발하게 된다. 반면 [7]의 경우에서는 훈련 환경과 인식 환경 자체가 완전히 다르지만, 적응 환경과 인식 환경은 유사하다. 따라서 적응 환경을 인식 환경과 최대한 유사하게 해 주어야 인식기의 성능 하락을 방지할 수 있다. 이를 위해 본 논문에서는 미리 N개의 다양한 주행 환경에서의 차량 잡음을 수집하였고, 이를 정차 중에 수집된 음성에 부가하여 인위적으로 N 환경의 화자 적응 데이터를 구성하였다. 잡음을 부가하는 방법은 [8]에서 제안한 방법을 사용하였다. 그리고 훈련 및 인식 데이터, 그리고 적응 데이터의 불일치를 감소시키기 위한 잡음제거 단계로 기존의 다양한 잡음 제거 방식들이 "De-noising" 모듈에 사용 될 수 있다. 본 논문에서는 ETRI에서 개발한 Wiener filter 기반 잡음 제거방법[9]을 사용하였다.

[7]에서는 다양한 환경에서 수집된 적응 데이터에 대해 L개의 대표 환경으로 분리하기 위한 환경군집화 작업을 수행한다. 그러나 적응 데이터의 수가 적으므로 다양한 환경을 나타내기가 어렵다. 또한 적응 데이터 내에 얼마나 많은 환경이 포함되어 있는가에 대한 정확한 개수의 환경 정보를 알 수가 없으며, 또한 각각의 환경별로 몇 개의 적응 데이터가 수집되었는지도 알 수가 없다. 따라서 아주 적은 개수를 가진 적응 데이터는 자신의 환경을 제대로 표현할 수 없으므로 다른 근사한 환경과 함께 군집화가 이루어진다. 그러나 본 논문에서는 정차상태에서 수집한 적응데이터에 대해 다양한 차량 잡음을 인위적으로 부가하기 때문에 환경은 부가한 N개의 차량 잡음 환경으로 자동적으로 결정되며, 또한 각각의 환경마다 동일한 적응 데이터 개수를 확보하게 되므로 좀 더 자세한 환경 모델을 구성할 수 있다. 그 후 각각의 환경에 대해 벡터 양자화 (Vector Quantizer; VQ) 분류기를 사용하여 L 크기의 코드북 (codebook)을 구성한다. 물론 VQ 분류기 외에 다양한 방법의 군집화 기법을 도입할 수 있으나, 고속 화자 적응 방식이므로 적응 데이터의 수가 소량이며 또한 단말기 성능의 제약으로 인한 계산량 등을 고려하여 [6]과 같이 간단하지만 좋은 성능을 보이는 VQ 분류기를 선택하였다. 따라서 본 논문에서의 제안한 방식의 코드북의 크기는 $N \times L$이며, [6]에서는 L이 된다.

본 논문에서도 각각의 환경별로 코드북을 생성시키기 위해 환경에 변별력 있는 특징 벡터로 [7]에서 제안한 것과 같이 비음성 구간의 켑스트럼 평균을 사용해서 코드북을 생성하였다. 또한, N개의 환경에 대해 ES-MLLR을 이용하여 변환 행렬 $W(1), W(2), \cdots, W(N)$을 추정한다. 따라서 각각의 환경별로 변환행렬과 환경 선택을 위한 코드북이 구성된다. 그리고 각각의 환경에 대한 부가적인 정보를 포함할 수 있다. 예를 들면, 주행 속도, 오디오나 냉난방 팬의 동작 여부, 창문 개폐 여부, 탑승자 위치나 내비게이션 지도로부터 터널 내 주행 여부 등에 대한 정보 등이다. 이를 통해 좀 더 정확한 환경 정보를 얻을 수 있으며, 환경 선택 시 이러한 정보를 이용할 수 있다.

3.2 온라인 단계

온라인 단계는 먼저 입력 음성과 가까운 환경을 선택해야 하는 작업이 필요하다. 물론 오프라인 단계에서 구성한 코드북 대신 N개의 $W(n)$의 바이어스 항들을 사용하여 환경 선택을 할 수 있으며, 이를 위해서는 입력음성에 대해 1차 인식을 수행하여 SM 방법[1] 등을 통해 구할 수 있다. 이 경우 입력 음성의 바이어스 값을 제대로 추정하기 위해 1차 인식 성능이 중요하다. 또한 단말기의 성능의 제약으로 인해 인식 소요시간이 길어지게 되며 운전자의 불편도 야기하게 된다. 따라서 간단하지만 효과적인 환경선택과 가중치를 구하는 것이 중요하므로 본 논문에서는 [7]에서 제안한 fuzzy k-means 알고리즘 기반의 가중합 방식을 사용하였다.

그림 1의 오프라인 단계에서 각각의 환경의 적응 데이터를 이용해 ES-MLLR을 통해 N개의 변환행렬 및 환경 선택을 위한 코드북을 구하였다. 이것을 이용하여 테스트 음성의 온라인 변환행렬은 테스트 데이터와 오프라인에서 미리 구한 각각의 환경과의 거리를 구해서 다음과 같이 유사한 정도에 따라서 N개의 변환행렬의 가중합의 형태로 사용한다.

$$\hat{W}_{online} = \sum_{n=1}^{N} w_f(n)\, W(n) \tag{6}$$

여기서, $W(n)$는 n번째 환경의 변환행렬이고, $w_f(n)$은 가중치이며 fuzzy k-mean algorithm에 의해 다음과 같이 구한다.

$$w_f(n) = \frac{1}{\sum_{k=1}^{N} \left(\frac{\| x - C_k \|}{\| x - C_n \|} \right)^{\alpha}} \tag{7}$$

또한, C_k는 k번째 코드북에서 입력 벡터 x와 가장 가까운 중심값을 의미한다. 즉,

$$C_k = C(k, l^*) \tag{8}$$

여기서, $C(k,l)$은 k번째 환경의 코드북에서 l번째 중심 값을 뜻하며, $l^* = \arg\min_{1 \le l \le L} \| x - C(k,l) \|$ 이다. 따라서 입력 음성에 따라 각각의 환경을 나타내는 대표 중심 값이 지속적으로 변하게 된다. 반면에 [7]에서는 단순히 고정된 L개의 중심 값으로부터 가중치를 구한다. 최종단계에서 (6)에서 구한 변환 행렬을 이용해 다음과 같이 온라인 모델을 구성하게 된다.

$$\hat{\mu}_{m,online} = \hat{W}_{online} \xi_m \tag{9}$$

4. 실험 및 결과

4.1 실험환경

본 논문에서 제안한 방법의 성능을 비교 평가하기 위해 ETRI에서 제공한 차량 내에서의 실 주행 시 녹음한 DB를 사용하였다. 총 147명으로 이루어진 화자로부터 127명의 DB를 훈련에 사용하였고, 10명에 대해 적응 및 평가를 수행하였다. 또한 10명에 대해서는 평가만을 수행하였다. DB는 총 시내주행이 약 30%, 고속도로 주행이 60%, 시동을 켠 채 정차 중 상태에서 녹음한 음성이 약 10%정도로 구성되어 있다. 특징 추출 모듈 및 인식기는 ETRI에서 제공한 툴킷을 사용하였으며, 또한 ETRI에서 개발한 Wiener 잡음 처리[9]도 함께 수행하였다. 사용한 특징 파라미터는 39차 MFCC를 사용하였고, CMS을 수행한다. 그리고 실제 차량용 내비게이션 단말기에 탑재되기 위해 특징 파라미터가 fixed-point 연산에 의해 추출되었으나, 편의를 위해 화자 적응 및 인식실험은 floating-point 연산으로 수행하였다. SI HMM은 321개의 state-tying triphone 모델로 구성되어 있고, 상태당 믹스쳐 수는 16개이다. 이것을 기준으로 127명 훈련 화자에 대해 MLLR을 수행하여 변환행렬을 구한 후 ES-MLLR의 기저 행렬을 생성하였다. 본 논문에서의 모든 실험에서 ES-MLLR의 기저 행렬의 개수를 20개로 정하였다.

평가는 10명의 화자에 대해 3만 단어급 차량 AV 시스템의 제어와 관련된 명령어 및 지도 검색을 위한 제한된 지명에 대한 고립단어에 대해 이루어졌다. 각각의 화자마다 30개의 발화로 이루어진 두 가지 종류의 적응데이터가 제공되었다. 하나는 정차 중에 발성한 것이며 다른 하나는 동일한 단어를 차량 주행 중에 발성한 것이다. 이를 정차중의 적응데이터에 대해 본 논문에서 제안한 방식을 통해 화자 적응을 수행했을 때 얻을 수 있는 최대 성능의 기준을 주행 중의 적응 데이터를 사용한 경우의 결과로 삼을 수 있으며, 본 논문에서 제안한 방식이 정차중 방식의 결과와 주행 중 방식의 결과의 중간 이상의 결과를 얻기를 기대하였다. 또한 인식 데이터로는 각각의 화자에 대해 80 발화를 사용하였고 모두 주행 중에 수집한 것이다. SI 모델을 사용한 경우 인식율이 94.25%를 보였다. 또한 모든 실험에서 ES-MLLR 방법의 경우는 바이어스 항 보상 방법을 함께 사용한 것이다.

4.2 실험 결과

기존의 방법들에 대한 인식 실험 결과를 그림 2에 나타내었다. 정차 중 수집한 것과 주행 중 수집한 두 가지 경우의 적응데이터를 사용하였다. 예상한 것과 같이 주행 중 수집한 적응데이터를 이용한 경우의 MLLR(driving)과 ES-MLLR(driving)의 성능이 SI 경우보다 향상되는 것을 알 수 있다. MLLR의 경우는 20개 이상의 적응데이터를 사용해야 하지만 ES-MLLR의 경우는 10개 이하의 데이터에서도 좋은 성능을 보임을 알 수 있다. 또한 잡음 제거를 통해 데이터가 어느 정도 정규화 되었다고 가정한다면, 정차 상태의 적응데이터에 대해서도 ES-MLLR(idle)의 성능 하락이 발생하지는 않았음을 알 수 있다. 또한 MLLR(idle)의 경우에서도 적응데이터가 증가할수록 SI 성능보다 좋은 성능을 보임을 알 수 있다.

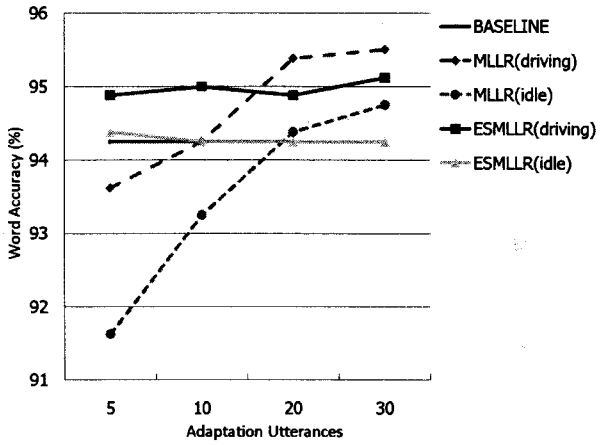

그림 2. 기존의 화자 적응 방법에 의한 실험 결과
Fig. 2. Results of conventional speaker adaptation

본 논문에서 제안한 방식의 유용성을 확인하기 위해 먼저 정차 중 적응데이터에 다양한 주행 잡음을 부가한 각각의 경우에 대해 화자 적응 성능을 살펴보았으며 그림 3에 그 결과를 나타내었다. 각각의 차량 잡음은 아스팔트 도로 주행(N1), 콘크리트 도로 주행(N2), 터널 주행(N3), 창문을 5cm 가량 연 상태에서 아스팔트 도로 주행(N4) 및 냉난방 팬을 중간정도로 튼 상태에서 아스팔트 도로 주행(N5) 상황에서 모았으며, 각각의 잡음 데이터를 정차 중에 모은 적응 데이터에 부가하여 5가지 환경의 인위적 왜곡 적응데이터 세트(set)를 구성하였다. 몇몇의 잡음의 경우는 주행 상태의 적응 데이터를 사용한 경우보다 높은 성능을 보이기도 한다. 따라서 환경 선택이 적절히 이루어진다면 높은 성능 향상을 얻을 수 있다. 또한 환경 선택을 통한 이상적인 최고 성능이 어느 정도인가를 알아보기 위해 각각의 화자별로 5가지 환경 실험 결과 중 적응 데이터 개수별로 제일 높은 성

능을 보인 것을 인위적으로 선택한 결과를 ES-MLLR(idle+Best)로 나타내었다. 이것은 만약 화자별로 각각의 적응 데이터에 대해 최적의 환경을 선택한다면 얻을 수 있는 최고의 인식성능의 한계를 나타낸다. 그러나 최적의 환경을 선택한다는 것은 현실적으로 불가능하므로 이보다는 주행 중 적응데이터를 사용한 경우인 ES-MLLR(driving)을 성능 비교의 기준으로 삼는 것이 적합하다. 그림 3의 결과 중 잡음을 인위적으로 부가한 몇 가지 실험에서는 주행 중에 수집한 적응 데이터를 사용한 결과를 상회하는 경우도 발생함을 알 수 있다.

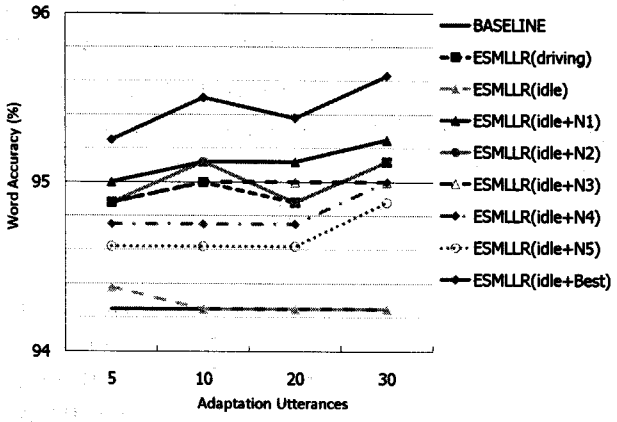

그림 3. 여러 가지 잡음을 부가한 경우의 ES-MLLR 방식의 인식 결과

Fig. 3. Results of ES-MLLR with the artificially distorted adaptation data by various noise types

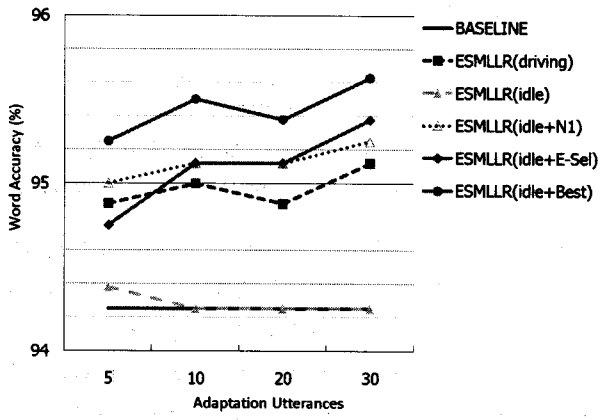

그림 4. 제안한 방식과 기존의 방식 인식 성능 비교

Fig. 4. The performance comparison of the proposed method and conventional methods

그림 4에 본 논문에서 제안한 방식의 결과를 나타내었다(ES-MLLR(idle+E-Sel)). 비록 ES-MLLR(idle+Best)의 성능에는 미치지 못하지만 ES-MLLR(driving) 보다는 높은 성능을 보인다. 즉, 제안한 방법의 결과가 정차 및 주행 중 적응 데이터를 사용한 결과들의 중간 정도에 위치할 것이라는 예상과는 달리 비교적 만족할 만한 성능을 보여줌을 알 수 있다. 물론 차량 주행 환경은 앞서 부가한 잡음보다 훨씬 다양한 형태로 존재하며, 이 모든 것을 다루는 것은 쉽지 않겠지만 대표적인 환경의 적절한 선택 및 또한 차량 텔레매틱스 단말기 등에서의 정보 공유 등을 통해 추가적인 성능 향상을 이룰 수 있을 것이다.

5. 결 론

본 논문에서는 다양한 주행 환경이 존재하는 차량 내에서 화자적응을 통해 인식성능 향상을 얻을 수 있는 방식에 대해 연구하였다. 운전자의 안전을 위해 정차 중에 적응 데이터를 얻은 후 인위적으로 다양한 주행 상태의 잡음을 부가하여 인위적 잡음 적응데이터를 구성하고 각각의 환경별로 변환 행렬을 구성하였다. 그러나 비록 입력 음성에 대해 먼저 특징 공간에서 잡음 처리 과정을 거쳤지만 여전히 존재하는 적응데이터와 인식데이터 사이의 불일치를 최대한 줄이기 위해 인식시 환경별 변환 행렬들의 가중합으로 실시간 환경 모델을 구성하였다. 실험결과 환경 선택시 테스트 데이터와 가장 유사한 하나의 환경을 이용하는 것보다 여러 환경들의 가중합을 이용한 방식이 가장 우수한 성능을 보였다.

참 고 문 헌

[1] Sanker, A. (1996). "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 3, pp. 190-202, May.

[2] Leggetter, C. J., Woodland, P. C. (1995). "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no.1, pp.171-185, Sep.

[3] Kuhn, R., Nguyen, P., Jungua, J. C., Goldwasser, L., Niedzielski, N., Finche, S., Field, K., Contolini, M. (1998) "Eigenvoices for speaker adaptation," *in Proc. ICSLP*, vol. 5, pp. 1771-1774.

[4] Chen, K. T., Liau, W. W., Wang, H. M., Lee, L. S. (2000). "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression", *in Proc. ICSLP*, pp. 742-745.

[5] Song, H. J., Kim, H. S. (2004). "Simultaneous Estimation of Weights of Eigenvoices and Bias Compensation Vector for Rapid Speaker Adaptation," *in Proc. ICSLP*, pp. 2945-2948, Oct.

[6] Song, H. J., Kim, H. S. (2006). "Fast Speaker Adaptation and Environment Compensation Based on Eigenspace-based MLLR," *Malsori*, vol. 58, pp. 35-44, Jun.
(송화전, 김형순, (2006). "잡음 환경에서의 Eigenspace-based MLLR에 기반한 고속 화자 적응," 말소리, 제58호, pp. 35-44)

[7] Kim, Y. K., Song, H. J., Kim, H. S. (2009). "Simultaneous Speaker and Environment Adaptation by Environment Clustering in Various Noise Environments," *Journal of Acoustical Society of Korea*, vol.28, No. 6, pp. 566-571, Aug.
(김영국, 송화전, 김형순, (2009) "다양한 잡음 환경하에서 환경 군집화를 통한 화자 및 환경 동시 적응," 한국음향학회 제

28권 제6호, pp. 566-571)

[8] ITU recommendation P.56. (1993). "Objective measurement of active speech level," Mar.

[9] Kang, B.-O., Jung, H.-Y., Lee, Y.-K. (2007). "Model Based Wiener Filter for Processing Dynamic Noise," *in proc. Conf. of The Korean Society of Phonetic Sciences and Speech Technology*, pp. 104-107, Nov.
(강병옥, 정호영, 이윤근, (2007) "동적 잡음 처리를 위한 모델 기반 Wiener 필터," 대한음성학회 가을학술대회 발표논문집, pp. 104-107)

• **송화전 (Song, Hwa Jeon)**
부산대학교 전자전기공학부
부산시 금정구 장전동 산32번지
Tel: 051-510-1704 Fax: 051-516-4279
Email: hwajeon@pusan.ac.kr
관심분야: 음성인식, 화자적응

• **전형배 (Jeon, HyungBae)**
한국전자통신연구원 음성언어정보연구부
대전시 유성구 가정로 138
Tel: 042-860-5788 Fax: 042-860-4889
Email: hbjeon@etri.re.kr
관심분야: 음성인식, 화자적응

• **김형순 (Kim, Hyung Soon)** 교신저자
부산대학교 전자전기공학부
부산시 금정구 장전동 산32번지
Tel: 051-510-2452 Fax: 051-515-5190
Email: kimhs@pusan.ac.kr
관심분야: 음성인식
현재 전자공학과 교수