IMM 기반 특징 보상 기법과 불확실성 디코딩의 결합

정회원 강 신 재*, 한 창 우*, 준회원 권 기 수*, 종신회원 김 남 수*°

Incorporation of IMM-based Feature Compensation and Uncertainty Decoding

Shin Jae Kang*, Chang Woo Han* *Regular Members*, Kisoo Kwon* *Associate Member*,
Nam Soo Kim*° *Lifelong Member*

요 약

본 논문은 잡음이 많이 존재할 경우 특징 보상 기법들의 불완전한 추정 방법으로 인하여 발생할 수 있는 불확실성 정보를 음성 인식의 디코딩에 반영해 줌으로써 좀 더 인식 성능을 향상시킬 수 있는 방법에 대한 연구이다. 기존의 특징 보상 기법들은 현재 시간에서의 깨끗한 특징 파라미터를 추정하는 단일점 추정 기법들이 대부분이다. 하지만 낮은 SNR 환경에서의 잘못된 추정 파라미터들이 음성 인식 엔진의 입력으로 사용될 경우 성능이 저하되기 때문에 추정된 파라미터의 불확실성 정보를 이용하여 디코딩을 해주면 추정 오류를 보완해줄 수 있다. 본 논문에서는 대표적인 Aurora-2 DB를 활용하여 적용된 기법의 성능 향상을 확인한다.

Key Words : 음성 인식, HMM, 특징 보상 기법, 상호 다중 모델, 불확실성 디코딩
speech recognition, HMM, feature compensation, interacting multiple model, uncertainty decoding

ABSTRACT

This paper presents a decoding technique for speech recognition using uncertainty information from feature compensation method to improve the speech recognition performance in the low SNR condition. Traditional feature compensation algorithms have difficulty in estimating clean feature parameters in adverse environment. Those algorithms focus on the point estimation of desired features. The point estimation of feature compensation method degrades speech recognition performance when incorrectly estimated features enter into the decoder of speech recognition. In this paper, we apply the uncertainty information from well-known feature compensation method, such as IMM, to the recognition engine. Applied technique shows better performance in the Aurora-2 DB.

Ⅰ. 서 론

음성 인식 성능을 하락시키는 요인들로는 주위 잡음 및 반향, 음향 모델을 학습할 때의 환경과 실제 테스트 환경의 불일치, 하드웨어 불일치 등 여러 가지가 있다. 이를 극복하기 위한 방법으로는 잡음으로 인해 왜곡된 음성 신호를 특징 영역에서 깨끗한 음성의 특징파라미터로 추정하는 특징 보상 기법과 모델을 주어진 환경에 맞게 변환시키는 모델 적응 기법이 있다. 일반적으로 특징 보상 기법은 온라인으로 빠르게 실

※ 본 논문은 우수제조기술연구 센서(ATC) 사업 (No. 10031489)의 지원을 받아 수행된 연구입니다.
* 서울대학교 전기·컴퓨터공학부 뉴미디어통신공동연구소 휴먼인터페이스 연구실(sjkang@.hi.snu.ac.kr) (° : 교신저자)
논문번호 : KICS2012-04-162, 접수일자 : 2012년 4월 2일, 최종논문접수일자 : 2012년 6월 5일

행이 가능한 반면 모델 적응 기법 만큼의 성능 향상을 가져올 순 없고 모델 적응 기법은 성능이 특징 보상 기법 보다는 높지만 모델 적응을 위해서는 변환하고자 하는 조건에 대한 음성 DB를 모아야 하는 단점이 있다.

본 논문에서는 수행속도가 빠르고 주어진 환경에 유연하게 적용 가능한 특징 보상 기법에 초점을 맞춘다. 특징 보상 기법은 깨끗한 음성의 특징 파라미터를 추정하는 것이 중요한데 낮은 SNR 환경에서는 불완전한 추정 기법들로 인해서 깨끗한 음성을 제대로 추정하지 못할 확률이 커지게 된다. 기존에 사용된 특징 보상 기법들은 현재 시간에서의 깨끗한 특징 파라미터를 추정하는 단일점 추정(point estimation) 기법들이 대부분이다. 이와 같이 잘못 추정된 파라미터들이 음성 인식 엔진의 입력으로 사용될 경우에는 음성 인식 성능을 저하시키는 원인으로 작용할 수 있기 때문에 추정된 파라미터의 불확실성 정보를 음성 인식 디코더에 반영해 준다면 성능을 보다 향상 시킬 수 있다.

Ⅱ. 상호 다중 모델

특징 보상 기법은 잡음, 반향, 채널 왜곡 등으로 변형되어 추출된 특징 파라미터를 깨끗한 음성의 특징 파라미터에 가깝게 보상해주는 기법이다[1]. 대표적인 기법으로 상호 다중 모델(interacting multiple model, IMM) 기반 특징 보상이 있다. IMM 알고리즘은 근사화된 선형 관계를 이용하여 각각의 혼합 요소에 해당하는 잡음의 통계적 특성을 추정하고 이를 종합하여 최종적인 잡음의 상태를 추정해서 잡음의 통계적 특성과 입력된 잡음 섞인 음성의 값을 이용하여 깨끗한 음성을 추정한 후 이를 음성 인식 시스템의 입력으로 이용하는 것이다. IMM 알고리즘은 로그-스펙트럼 영역에서의 깨끗한 음성의 사전 확률 분포를 가우시안 혼합 모델로 가정한다. 가우시안 혼합 모델은 확률분포가 정규분포의 선형결합으로 이루어지는 형태를 나타내며 최대 우도 학습 방법이나 벡터 양자화 알고리즘을 이용하여 코드북을 구성하여 이용할 수 있다. IMM 알고리즘에서는 각각의 혼합 요소 내에서 잡음과 깨끗한 음성의 결합이 부분적으로 선형이라는 가정을 한다. 실제로 로그-스펙트럼 영역에서 잡음과 깨끗한 음성이 결합되는 형태는 다음과 같은 매우 비선형적인 관계를 갖는다.

$$z^l = \log(\exp(x^l) + \exp(n^l)). \tag{1}$$

여기에서 x, n, z는 깨끗한 음성, 잡음, 잡음 섞인 음성을 나타내고 위첨자 l은 로그-스펙트럼을 나타낸다. 하지만 이 관계를 그대로 이용하여 잡음을 추정하는 것은 대단히 어려운 일이기 때문에 이를 선형관계로 근사화 시킨다.

$$z^l = A_k x^l + B_k n^l + C_k. \tag{2}$$

여기에서 $A_k(M\times M)$, $B_k(M\times M)$, $C_k(M\times 1)$는 $k(1 \le k \le K,\ K$는 가우시안 혼합 요소의 차수)번째 혼합 요소의 상수 행렬을 나타내고 M은 로그-스펙트럼의 차수를 나타낸다. IMM 알고리즘은 잡음이 로그-스펙트럼 영역에서 천천히 변하는 것을 가정하여 시간에 따라 변하는 잡음의 상태를 추정할 수 있다. IMM 알고리즘에서 잡음에 대한 상태방정식은 다음과 같이 구성된다.

$$n_t^l = n_{t-1}^l + w_t. \tag{3}$$

여기에서 w_t는 정규분포 확률과정를 따른다. 추정된 깨끗한 음성은 근사화나 추정 기법의 불완전성 때문에 실제 깨끗한 음성과 에러가 발생하게 된다. 이를 반영해주기 위해 깨끗한 음성의 불확실성도 같이 계산해서 이를 디코딩 단에 반영해준다.

$$\hat{\mu}_s^l(t) = \sum_{j=1}^{N} \gamma_j(t)\hat{\mu}_s^l(t|j) \tag{4}$$

$$\hat{\Sigma}_s^l(t) = \sum_{j=1}^{N} \gamma_j(t)[\hat{\Sigma}_s^l(t|j) + \left(\hat{\mu}_s^l(t|j) - \hat{\mu}_s^l(t)\right) \cdot \left(\hat{\mu}_s^l(t|j) - \hat{\mu}_s^l(t)\right)^T]. \tag{5}$$

여기에서 $\hat{\mu}_s^l(t)$, $\hat{\Sigma}_s^l(t)$는 시간 t에서 로그-스펙스럼 영역의 추정된 깨끗한 음성의 평균과 분산을 나타내고, $\hat{\mu}_s^l(t|j)$, $\hat{\Sigma}_s^l(t|j)$는 j번째 혼합 요소의 로그-스펙트럼 영역의 추정된 깨끗한 음성의 평균과 분산, $\gamma_j(t)$는 j번째 혼합 요소의 사후 확률을 나타낸다. 위첨자 T는 벡터나 행렬의 전치를 나타낸다. 실제 사용되는 특징 파라미터는 멜-주파수 켑스트럴 계수(MFCC)를 사용하기 위해서 위의 평균과 분산 값을 이산 코사인 변환을 이용하여 켑스트럼 영역으로 변환해준다. 이 때 분산값은 대각행렬로 가정한다[2].

Ⅲ. 특징 보상 기법의 불확실성 정보를 이용한 디코딩

왜곡된 특징 파라미터를 깨끗한 음성의 특징 파라미터에 가깝게 보상해주는 기법이 특징 보상 기법이다. 하지만 깨끗한 음성 추정 기법의 불완전성으로 인하여 추정 에러가 발생하게 되고 이 불완전성 정보를 음성 인식 디코딩에 반영해주면 음성 인식 성능을 향상시킬 수 있다[3,4,5]. 깨끗한 음성 데이터의 특징 벡터를 x, 추정된 깨끗한 음성의 특징 벡터를 $\hat{x}$이라고 할 때

$$x = \hat{x} + e \tag{6}$$

로 표현할 수 있다. 여기에서 e는 특징 보상 기법에서의 추정 에러를 나타낸다. 보통 e는 영평균 정규분포 확률변수로 가정한다.

$$e \sim N(e;\vec{0},\hat{\Sigma}_x). \tag{7}$$

특징 영역에서의 불확실성 정보를 고려한 디코딩 규칙은 아래와 같이 나타낼 수 있다.

$$\hat{W} = \arg max_W \left[\int_{x\in\Omega} p(x|\Lambda, W) p(x|\theta) dx \right] P(W). \tag{8}$$

여기에서 Λ는 고정된 음향 모델 파라미터, θ는 통계적 특징 추출 알고리즘에서 정해지는 음성 특징의 분포를 나타내는 파라미터, Ω는 특징 벡터열 x가 취할 수 있는 모든 값, W는 단어열, $\hat{W}$은 추정된 단어열을 나타낸다. 기존 디코딩 방식[6]과의 차이점은 $p(x|\Lambda, W)$항에 $p(x|\theta)$이 곱해져서 합해진 형태라는 것인데 이는 잡음 환경에서 단일점으로 추정된 깨끗한 특징 파라미터를 사용하여 확률값을 계산하는 것이 아니라 가능한 모든 특징 파라미터들의 확률을 고려하여 깨끗한 특징 파라미터의 확률값을 계산하겠다는 의미이다. 이와 같이 식 (8)의 $p(x|\theta)$항을 통해서 추정의 불완전성 효과를 반영할 수 있게 된다. Λ와 W는 서로 독립이라 가정하게 되면

$$\begin{aligned} \int_{x\in\Omega} p(x|\Lambda) p(x|\theta) dx &= \int_x p(x|\Lambda_s) p(x|\hat{\Sigma}_x) dx \\ &= \int_x N(x;\mu_s,\Sigma_s) N(x;\hat{x},\hat{\Sigma}_x) dx \\ &= N(\hat{x};\mu_s,\Sigma_s+\hat{\Sigma}_x) \end{aligned} \tag{9}$$

이 되고

$$\begin{aligned} p(x|s) &= \int_e p(x,e|s) de \\ &= \int_e p(x|e,s) p(e) de \\ &= \int_e N(e;\mu_s-\hat{x},\Sigma_s) N(e;0,\hat{\Sigma}_x) de \\ &= N(\hat{x},\mu_s,\Sigma_s+\hat{\Sigma}_x) \end{aligned} \tag{10}$$

이므로 식 (9)와 식 (10)을 통해

$$\int_{x\in\Omega} p(x|\Lambda) p(x|\theta) dx = p(x|s) \tag{11}$$

임을 알 수 있다. 따라서 식 (11)과 같이 불확실성 정보를 이용하여 확률값을 계산할 수 있다. 여기에서 s는 음성의 은닉 마르코프 모델의 상태(state)를 말한다.

Ⅳ. 실험 결과

불확실성을 고려한 IMM 알고리즘의 성능을 평가하기 위해 Aurora-2 DB를 사용하였다. Aurora-2 DB는 TI-DIGITS 데이터베이스에서 8kHz로 다운샘플링하고 2개의 표준 통신 채널 G.712 또는 MIRS 필터를 통과시켜서 얻어졌다. Aurora-2 DB는 3개의 테스트 세트로 구성되어 있고 여러 잡음을 녹음하여 깨끗한 음성에 SNR 0~20dB로 더하여 구성되어 있다. 특징 파라미터는 ETSI 표준 선처리기를 사용하여 13차 MFCC와 그의 delta, delta-delta 파라미터를 추출하여 총 39차를 사용하였다[7]. 본 논문에서 특징 보상 기법으로 사용하고 있는 IMM 알고리즘은 정적 특징 파라미터를 대상으로 수행되기 때문에 불확실성 정보를 디코딩에 반영해 줄 때에는 정적 특징 파라미터 변수에 해당하는 정보만을 고려하였다. 편의상 아무것도 처리하지 않은 기본 인식 성능과 IMM 기반 특징 보상 기법, IMM에서 추정된 불확실성 정보를 고려한 디코딩 기법들을 각각 Baseline, IMM, IMM+UD로 표기하였다. 음성 인식을 수행하기 위해 기존의 IMM에서는 HTK를 사용하였고, IMM+UD에서는 특징 보상 기법의 불확실성 정보를 반영해 주기 위해 HTK 코드를 수정하여 사용하였다[8].

표 1. Aurora-2 DB 평균 인식 성능 비교 (%)
Table 1. Aurora-2 DB average recognition results (%)

DB / 기법	set A	set B	set C
Baseline	65.58	61.75	72.48
IMM	85.30	86.60	80.56
IMM+UD	85.92	86.99	81.01

표 2. Aurora-2 DB 평균 RERR 성능 (%)
Table 2. Aurora-2 DB average RERR (%)

DB / 기법	set A	set B	set C
IMM	50.74	68.56	24.82
IMM+UD	52.20	69.14	26.81

표 3. 수행 속도 비교(×실시간)
Table 3. Processing time(× real-time)

기법 / 속성	IMM	IMM+UD
수행속도	0.0527	0.0602

표 1의 결과를 보면 IMM+UD 결과가 IMM보다 조금 성능이 향상되었음을 확인할 수 있다. 그리고 표 2를 통해 평균 relative error rate reduction (RERR)의 성능을 확인할 수 있다. 표 3은 CPU 코어 속도 3GHz, RAM 1GB의 데스크탑 환경에서 불확실성 정보를 적용하기 전과 후의 인식 수행 속도를 비교한 것이다. 디코딩과정에서 활성 은닉 마르코프 모델의 상태 수가 증가하여 응답속도가 증가할 수 있는 있지만 실험 결과에서는 비교적 큰 차이를 보이지 않고 있는데 이는 Aurora-2 DB가 숫자음 DB이고 음향 모델에 사용되는 상태의 수가 적기 때문에 발생한 결과라고 볼 수 있다.

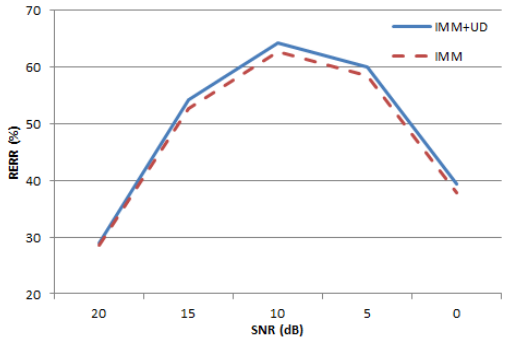

그림 1. SNR에 따른 RERR 비교
Fig. 1. Comparison of RERR according to SNR

그림 1은 SNR 변화에 따른 IMM과 IMM+UD와의 RERR을 비교한 것이다. 불확실성 정보를 활용하여 디코딩을 해줄 경우 좀 더 성능 개선이 있음을 확인할 수 있다.

V. 결 론

본 논문에서는 특징 보상 기법의 깨끗한 음성 추정 과정에서 얻을 수 있는 불확실성에 대한 정보를 디코딩에 반영해줌으로써 기존의 특징 보상 기법보다 더 향상된 인식 성능을 얻을 수 있었다. 제안된 기법은 특히 추정의 불완전함에 의해 불확실성이 크게 얻어지는 낮은 SNR 환경에서 좋은 성능을 보였다.

향후에는 동적 특징 파라미터의 영향이 음성 인식 성능에 매우 크게 작용하기 때문에 동적 특징 파라미터에서의 불확실성 정보도 추출하여 반영함으로써 성능을 더 향상시킬 수 있는 방법에 대한 연구를 진행할 것이다.

References

[1] N. S. Kim, “Feature domain compensation of nonstationary noise for robust speech recognition,” *Speech Communication*, Vol. 37, pp. 231-248, Jul. 2002.

[2] R. F. Astudillo and D. Kolossa, “Uncertainty propagation,” in *Robust Speech Recognition of Uncertain or Missing Data: Theory and Applications*, D. Kolossa and R. Haeb-Umbach, Eds. Springer, Jul. 2011.

[3] L. Deng, “Front-end, back-end, and hybrid techniques for noise-robust speech recognition,” in *Robust Speech Recognition of Uncertain or Missing Data: Theory and Applications*, D. Kolossa and R. Haeb-Umbach, Eds. Springer, Jul. 2011.

[4] L. Deng, J. Droppo and A. Acero, “Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion,” *IEEE Trans. Speech and Audio Processing*, Vol. 13, No. 3, pp. 412-421, May 2005.

[5] 강신재, 한창우, 권기수, 김남수, “IMM 기반 특징 보상 기법의 추정된 분산을 이용한 불확실성 디

코딩," 한국통신학회 동계종합학술발표회 논문집, 2012년 2월.

[6] Q. Hue and C. Lee, "A Bayesian predictive classification approach to robust speech recognition," *IEEE Trans. Speech and Audio Processing*, Vol. 8, No. 8, pp. 200-204, Nov. 2000.

[7] Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithm, *ETSI ES 201108 V1.1.3*, Sep. 2003, ETSI Std. Doc..

[8] S. Young, *The HTK Book*. Cambridge, U.K.: Eng. Dept. Cambridge Univ. 2006.

강 신 재 (Shin Jae Kang) 정회원

2008년 2월 충남대학교 전기·정보통신공학부 졸업
2010년 8월 서울대학교 전기·컴퓨터공학부 석사
2010년 9월~현재 서울대학교 전기·컴퓨터공학부 박사과정
<관심분야> 음성 인식, 음성 신호처리

한 창 우 (Chang Woo Han) 정회원

2006년 2월 서울대학교 전기공학부 졸업
2006년 3월~현재 서울대학교 전기·컴퓨터공학부 석박통합과정
<관심분야> 음성 인식, 음성 신호처리

권 기 수 (Kisoo Kwon) 준회원

2011년 2월 서울대학교 전기공학부 졸업
2011년 3월~현재 서울대학교 전기·컴퓨터공학부 석박통합과정
<관심분야> 음질 향상, 음성 신호처리

김 남 수 (Nam Soo Kim) 종신회원

1988년 2월 서울대학교 전자공학과 졸업
1990년 2월 한국과학기술원 전기 및 전자공학과 석사
1994년 8월 한국과학기술원 전기 및 전자공학과 박사
1998년 3월~현재 서울대학교 전기·정보공학부 교수
<관심분야> 음성 신호처리, 음성 인식, 통계적 신호처리, 패턴 인식, 휴먼 인터페이스