모바일 기기를 위한 음성인식의 사용자 적응형 후처리

(User Adaptive Post-Processing in Speech Recognition for Mobile Devices)

김 영 진 † 김 은 주 †
(YoungJin Kim) (EunJu Kim)

김 명 원 ††
(MyungWon Kim)

요 약 본 논문에서는 모바일 환경에서 고립단어 음성인식을 할 경우 화자종속 방법을 이용하여 성능을 높이는 사용자 적응형 후처리 방법을 제안한다. 이 방법은 인식기의 정확한 인식 결과를 위한 추가적인 처리들로 구성된다. 즉 인식기의 출력과 정확한 최종 결과들 간의 관계를 학습하여 이를 잘못된 인식기의 출력을 수정하는 데에 사용한다. 학습에는 패턴인식에 강인한 다층 퍼셉트론을 사용하며 학습 시간을 고려하여 모델을 세분화하고 동적으로 동작할 수 있도록 구현한다. 이 결과 인식기의 오류에 대해 41%를 수정하는 성과(오류 수정률: 41%)를 보였다.

키워드 : 신경망, 사용자 적응, 음성인식, 후처리

Abstract In this paper we propose a user adaptive post-processing method to improve the accuracy of speaker dependent, isolated word speech recognition, particularly for mobile devices. Our method considers the recognition result of the basic recognizer simply as a high-level speech feature and processes it further for correct recognition result. Our method learns correlation between the output of the basic recognizer and the correct final results and uses it to correct the erroneous output of the basic recognizer. A multi-layer perceptron model is built for each incorrectly recognized word with high frequency. As the result of experiments, we achieved a significant improvement of 41% in recognition accuracy (41% error correction rate).

Key words : Neural Network, User Adaptive, Speech Recognition, Post-Processing

· 본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음
· 이 논문은 2007 한국컴퓨터종합학술대회에서 '모바일 기기를 위한 사용자 적응형 음성인식 후처리'의 제목으로 발표된 논문을 확장한 것임
† 학생회원 : 숭실대학교 컴퓨터학과
liebulia@ssu.ac.kr
blue7786@ssu.ac.kr
†† 종신회원 : 숭실대학교 컴퓨터학과 교수
mkim@ssu.ac.kr
논문접수 : 2007년 9월 28일
심사완료 : 2007년 10월 23일

1. 서 론

최근 휴대형 모바일 기기의 다기능화, 소형화 추세로 인하여 기기를 편하게 사용하기 위한 사용자 친화적 인터페이스의 요구가 증대되고 있다. 특히 대표적인 사용자 친화적 인터페이스인 음성인식은 사용자가 대화를 나누듯 기능을 사용할 수 있다는 장점 때문에 많은 모바일 기기에서 사용 중이며, 잡음이 없는 환경에서 87~97%의 높은 인식률을 보이고 있다. 그러나 휴대형 모바일 기기의 특성상 다양한 잡음 환경에 노출되어 실생활에서의 인식률은 현저히 떨어지고 있다.

이러한 단점을 보완하기 위해서는 음성인식 결과를 효과적으로 보정할 수 있는 음성인식 후처리 방법이 필요하다. 음성인식 후처리 방법에는 오류 패턴(error pattern) 정보[1], 의미(semantic) 정보[2], 문맥(context) 정보[3], 발화 순차 패턴(sequence pattern)[4] 정보 등을 이용한 후처리 방법이 있다. 그러나 이러한 방법들은 단지 인식기의 최종 인식 단어만을 이용하므로 인식기와 독립적으로 후처리에 적용될 수는 있으나 주변 잡음에 의해 연속적으로 인식기의 오인식이 발생하는 경우 후처리 결과를 신뢰하기 어렵다.

본 논문에서는 주변 잡음 환경에서 인식기의 인식률을 높이기 위하여 사용자의 발화 특성 정보를 다층 퍼셉트론(multilayer perceptron)[5,6]으로 학습하고 오인식 단어를 보정하는 사용자 적응형 후처리 방법을 제안한다. 이 방법에서는 HMM(hidden Markov model)[7,8]기반 음성 인식기를 사용하여 오인식이 자주 발생되는 단어 모델을 생성하고 인식기 인식 결과에 영향을 미치는 모든 단어들의 인식 가능도인 likelihood를 이용한다. 이러한 정보로 만들어진 모델은 사용자의 사용에 따라 조금씩 적응되면서 사용자의 발화 특성을 최대한 반영하게 된다.

본 논문은 총 5절로 구성되어 있다. 2절에서는 국내외 주요 음성인식 후처리 방법에 대하여 소개하고 3절에서는 제안하는 사용자 적응형 후처리 방법에 대하여 기술한다. 4절에서는 제안한 방법에 대한 실험 결과를 기술하고 분석하여 타당성을 검증하며, 5절에서는 결론을 맺고 향후 연구 과제에 대해서 검토한다.

2. 관련 연구

2.1 오류 패턴 정보 후처리

[1]은 오인식에 일정한 유형이 있다고 가정하고 오인식이 자주 발생하는 단어와 단어의 오인식 형태를 쌍으로 묶어 오류패턴 사전을 구축하였다. 그리고 오인식을 수정하기 위한 후처리 방법으로 EPC (error-pattern correction)와 SSC(similar-string correction)를 제안하였다. EPC 방법은 인식기의 인식 결과와 미리 구축된 오류패턴 사전을 비교하여 오인식이 예상되는 부분을 검출한다. 그리고 예상된 오인식 단어를 사전상의 원래의 단어로 치환하여 오류를 수정한다. SSC 방법은 오류패턴의 단위를 EPC에서 처럼 오인식된 단어만 보는 것이 아닌 오인식 된 단어의 앞·뒤 부분까지를 하나로 묶어 오인식 블록으로 사용하는 방법이다.

이 방법은 매우 직관적인 방법으로 효과적인 오류 수정을 기대할 수는 있으나 오류 수정이 단순하게 단어, 블록 단위로 치환되기 때문에 올바른 단어가 잘못된 단어로 오인 될 위험도 함께 가지고 있다. 또한 음성인식 환경이 변하는 경우에는, 발생되는 잡음의 특성에 따라 인식기의 오류패턴이 다르므로 오류패턴 사전 자체를 재구축해야 한다는 문제가 있다.

2.2 어휘 의미 패턴 정보 후처리

[2]에서는 발화된 내용의 어휘 및 의미 정보를 나타내는 LSP(lexico-semantic pattern; 어휘의미패턴)을 정의하고 이를 이용한 후처리 방법을 제안하였다. LSP는 문장을 구성하는 단어들을 각 단어의 의미범주들로 재구성한 것을 말한다. 이를 위해 인식기로부터 인식 가능한 단어들을 비슷한 의미 범주로 묶고, 묶인 단어들을 어휘별로 구분지은 의미 범주 사전을 이용한다.

후처리에서는 의미 범주 사전을 사용자가 발화한 문장을 LSP로 변환된다. 이후 변환된 LSP와 학습데이타를 통해 미리 구성된 LSP들과 비교하여, 가장 유사한 LSP를 선택하고 이를 변화된 LSP와 바꾸어 의미적 오류를 수정한다. 마지막으로 LSP의 각 범주 정보들을 실제 단어로 바꾸는 어휘적 오류 수정을 통하여 최종 인식 결과를 도출한다. 이 방법은 음성인식의 오류를 수정하는데 발화된 내용의 의미정보를 고려하였다는 장점이 있으나, 학습데이타를 통해 구성될 LSP가 잡음 환경에 영향을 받아 신뢰성을 잃을 경우 후처리의 신뢰성 또한 보장할 수 없다는 단점이 있다.

2.3 신경망을 이용한 문맥 정보 후처리

[3]에서는 문맥 정보 기반 후처리를 제안하였다. 문맥 정보 후처리는 사용자가 발화한 단어들 간의 순차 패턴을 문맥으로 정의하고 이 패턴을 신경망으로 학습하여 최종 인식에 적용하는 방법이다. 이 방법으로 문맥 정보 후처리를 사용할 경우 주변 잡음에 강인한 인식 결과를 얻을 수 있고, 학습 후 적용 속도가 빠르다는 장점도 있다. 그러나 인식하고자 하는 단어의 수가 증가하거나 선행 단어의 수가 늘어나게 되면 이에 따른 학습 시간이 기하급수적으로 증가하므로 모바일 기기 사용자가 사용하고자 하는 다양한 잡음 환경을 학습시키는 데 많은 시간이 걸리게 된다. 또한 선행 단어의 수를 최적화하는 데 있어 적용하고자 하는 도메인마다 다르게 나타나므로 이를 결정하는 데 어려움이 있다.

3. 사용자 적응형 후처리

3.1 음성인식 시스템 구조

사용자 적응형 후처리를 이용한 음성인식 시스템은 그림 1과 같이 인식기와 사용자 적응형 후처리로 구성된다. 인식기는 사용자가 발화한 음성 신호에서 음성특징을 추출하고 HMM을 이용하여 일차적인 인식과정을 수행한다. 사용자 적응형 후처리는 인식기의 인식 결과를 받아 후처리 적용여부를 확인하고 후처리가 필요한 경우 적용하여 최종 결과를 도출한다.

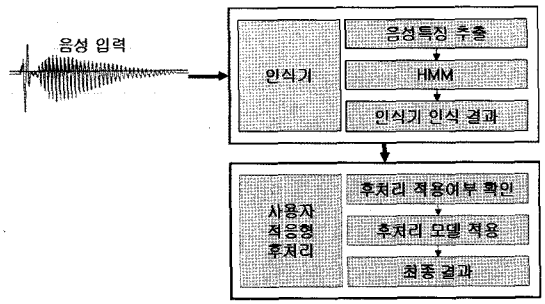


그림 1 음성인식 시스템 구조

3.2 사용자 적응형 후처리 모델

사용자 적응형 후처리 모델은 그림 2와 같이 다층 퍼셉트론 구조를 가진다. 입력층의 각 노드에는 인식기가 인식 가능한 단어들의 인식 가능도를 최소-최대 정규화(Min-max Normalization)한 likelihood가 입력되며, 출

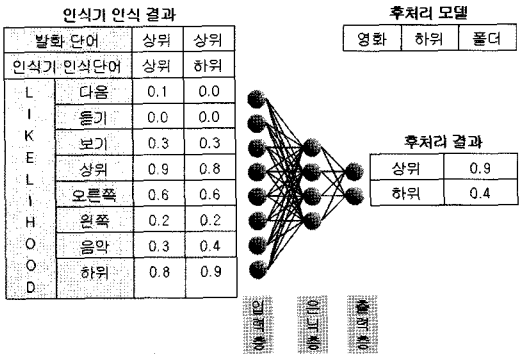


그림 2 사용자 적응형 후처리 적용 방법의 예

력층의 각 노드는 후처리에 의해 수정하고자하는 단어로 표현된다. 즉 사용자가 "상위"라고 발화한 음성이 "하위"라고 자주 오인식이 일어난다면 "하위"라는 이름의 모델을 만든다. 그리고 인식기의 인식단어로 "하위"가 인식될 경우 "하위"모델을 적용하여 진짜 "하위"인지 아니면 "상위"인지를 구분하도록 학습시킨다.

예를 들어 "영화", "하위", "폴더" 후처리 모델이 있을 때, 사용자가 "상위"를 발화하고 인식기가 이를 "상위"라고 인식한 경우 인식기가 인식 가능한 단어에 대한 likelihood 값을 그림 2와 같이 갖는다. 여기서 인식기의 인식 결과인 "상위"는 다른 단어들의 likelihood에 비해 가장 높은 값을 갖는 단어가 되어 인식 결과가 "상위"가 된다. 인식기가 "상위"를 "상위"로 인식한 경우 인식기의 인식 단어가 후처리 모델에 없으므로 인식기의 결과인 "상위"가 최종 결과가 된다.

만약 인식기가 "상위"를 "하위"로 오인식 하는 경우 인식기 결과 단어인 "하위"가 후처리 모델에 있으므로 "하위"라는 후처리 모델을 적용하게 된다. 따라서 likelihood가 가장 높아 "하위"로 오인식한 인식기 결과를 제안하는 후처리 방법을 적용하여 "상위"로 수정된다.

3.3 사용자 적응형 후처리 모델 생성

사용자 적응형 후처리 모델의 학습은 그림 3과 같이 5단계로 구성된다. 인식기 결과는 사용자에게 요청한 발화 음성을 인식한 인식기 인식 결과이다. 사용자 적응형 후처리 모델을 구성하는 데에는 모델 선정과 출력단어를 선정하는 두 단계로 이루어지며, 이는 인식기 결과의 인식 단어를 기준으로 생성된다. 모델 생성은 인식률이 낮은 단어가 어떠한 단어로 오인식 되는지를 분석하여 생성한다. 모델 생성 조건은 식 (1)과 같다.

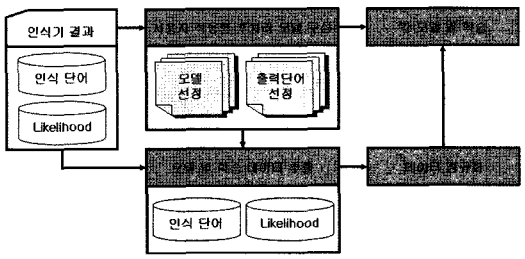


그림 3 사용자 적응형 후처리 모델 학습 진행도

$$P(X_r = M_i | X_t = W_t) > \alpha \quad (1)$$

식 (1)에서 X_r은 인식기 인식 단어를 의미하며 X_t는 발화 단어(목적 단어)를 의미한다. W_t는 t번째 발화 단어이고 M_i는 W_t의 i번째 인식 단어이며 α는 M_i의 오류율로 모델 선정 임계값이 된다. 즉 식 (1)은 사용자의 발화 단어 중 t번째의 단어에 대한 인식기 인식 단어의 비율을 의미하며 오류 비율이 α 이상인 경우 인식기 인식 단어를 모델로 선정하게 된다. 단, W_t와 M_i가 같은 단어일 경우 M_i는 모델로 선정하지 않는다.

표 1 모델 선정 및 출력 단어 선정 예

발화 순서	발화 단어	인식기 인식 단어	발화 순서	발화 단어	인식기 인식 단어
1	상위	상위	9	하위	하위
2	상위	상위	10	하위	하위
3	상위	하위	11	위	위
4	상위	하위	12	위	위
5	상위	상위	13	위	위
6	하위	하위	14	위	하위
7	하위	하위	15	위	위
8	하위	하위			

만약 표 1처럼 발화 단어가 "상위"와 "하위" 각각 다섯 개씩 존재하고 이에 따른 인식기 인식 단어가 다음과 같을 때, 모델 선정 임계값인 오류율을 0.3으로 하면 "하위"모델이 생성된다. 즉 발화 단어 중 "상위"에서 "하위"로 오인식 한 경우의 비율이 미리 선정한 오류율 이상이 된다. 같은 방법으로 "상위"모델의 생성 여부를 판단하며 오류율 0.3을 넘지 못하므로 "상위"모델은 생성되지 않는다.

출력 단어는 식 (2)에 의해 선정되며 식의 표기 방식은 식 (1)과 동일하다. 단, 출력 단어는 선정하고자 하는 i번째 모델 M_i에서 단어 W_t가 오류를 범하는 비율을 계산한다. 여기서 β는 출력 단어 선정을 위한 임계값으로 식 (1)의 α와 동일하게 적용되나 β 값이 너무 작은 경우 후처리 모델을 학습시키기 위한 데이타의 량이 적게 되어 좋은 성능의 모델 구축에는 어려움이 발생한다. 이 경우 β값은 α값 보다 큰 값을 적용한다.

$$P(X_t = W_t | X_r = M_i) > \beta \quad (2)$$

표 1에서 출력 단어 선정 오류율을 0.1로 정한다면 "상위"라고 발화한 5개의 데이타에서 "하위"로 인식한 것이 두 번이므로 오류율이 0.4가 되어 출력단어 선정 오류율인 0.1 보다 큰 값을 갖는다. 따라서 "상위"가 첫 번째 출력 단어로 선정되며 "하위"라고 발화한 5개의 데이타에서는 "하위"라고 인식한 것이 5개이므로 "하위" 역시 출력 단어로 선정된다. 마지막으로 "위"라고 발화한 5개의 데이타는 "하위"라고 인식한 경우가 한 번이므로 0.2의 오류율 갖게 되어 "하위" 모델에 "위"라고 하는 출력 단어를 추가하게 된다.

이렇게 생성된 후처리 모델을 학습시키기 위해서는 인식기의 인식 결과 중에서 필요한 데이타만 추출해야 한다. 이러한 데이타 추출은 모델명과 각 모델의 출력단어를 이용하게 된다. 마지막으로 신경망 학습을 위해 최대값 1, 최소값 0인 최소-최대 정규화를 사용한다.

4. 실험 결과 및 분석

4.1 실험 환경 및 데이타 수집

본 실험에서 사용된 음성 데이타는 16kHz, 16bit, 모노(mono)로 설정하였고 주변 잡음이 정제되지 않은 일반 잡음 환경에서 마이크로폰(microphone)을 이용하여 녹음하였다. 녹음된 웨이브(wave) 데이타는 인식기 학습을 위해 MFCC(Mel frequency cepstral coefficient) 특징추출 방법을 사용하여 음성의 특징을 추출하였고 인식기는 캠브리지(Cambridge) 대학에서 개발한 HTK (HMM toolkit)[9]를 이용하여 구축하였다.

HMM 기반 음성인식기인 HTK는 각 단어별 음성데이타로 학습된 음성인식 모델을 이용하여 발화된 단어의 인식 가능한 단어들의 인식 가능도을 log likelihood로 표현하고, 최대값을 가지는 단어를 최종 결과로 선정한다. 본 실험에서는 사용자의 발화가 다른 주변 단어에 어떠한 영향을 미치는지 알아야 함으로 최종 인식 결과만이 아닌 인식 가능한 모든 단어의 likelihood값을 사용한다.

4.2 실험 데이타 및 단어 선정

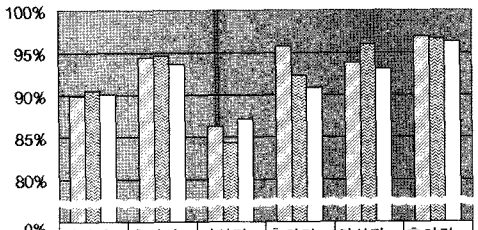
본 실험은 모바일 환경에서 사용되는 단어(명령어)들 중 많이 사용되는 단어 총 32개를 표 2와 같이 선별하였고 적응형 후처리 실험을 위해 3명의 사용자가 각각 단어 당 500회의 발화 데이타를 구축하였다. 이중 각 단어 당 150개의 데이타는 인식기를 학습하는데 사용하였으며 단어 당 50개의 데이타를 사용하여 전체 평가에 사용하였다. 그리고 나머지 데이타는 사용자 적응형 후처리 학습에 사용하였다.

표 2 실험에 사용된 32개 단어

No	단어	No	단어	No	단어	No	단어
1	다음	9	음악	17	시간	25	종료
2	듣기	10	정지	18	시작	26	지도
3	보기	11	찾기	19	아니오	27	축소
4	삭제	12	폴더	20	예	28	출발지
5	상위	13	하위	21	오른쪽	29	취소
6	아래	14	거리	22	왼쪽	30	크게
7	영화	15	도착지	23	위치	31	확대
8	위	16	소리	24	작게	32	확인

4.3 최종 인식률 평가

적응형 후처리를 위한 초기 모델을 생성하기 위하여 32개 단어에 대해 50회씩 발화한 데이타를 적용하였다. 이 후 인식률이 떨어지는 단어에 대하여 단어 당 50회씩 발화한 데이타를 추가하는 방식으로 총 4회 실시하였다. 그 결과 3명의 사용자에 대해 각각 15, 15, 16개의 모델이 생성되었다. 이때 모델과 출력 단어를 선정하는 임계값은 3%를 적용하였으며, 학습에 사용된 신경망



	인식기	후처리	인식기	후처리	인식기	후처리
	전체 단어		모델 선정 단어		비 선정 단어	
사용자 1	89.9%	94.4%	86.3%	95.7%	93.8%	96.9%
사용자 2	90.5%	94.6%	84.3%	92.3%	96.0%	96.6%
사용자 3	90.1%	93.6%	87.1%	90.9%	93.1%	96.3%
평 균	90.2%	94.2%	85.9%	93.0%	94.3%	96.6%

그림 4 사용자별 모델 인식률 비교

모델의 출력 노드 수는 모델 마다 가변적으로 적용하여 2~6개까지 생성되었고 입력 노드의 수는 32개로 고정하였다.

그림 4는 사용자 별 최종 인식결과 그래프이다. 여기서 인식기는 후처리를 거치지 않은 인식기만의 인식률을 나타낸 것이고 후처리는 인식기의 인식 결과를 가지고 후처리를 적용하여 나온 결과를 나타낸 것이다. 전체 단어는 사용자 별 전체 인식률을 나타낸 것으로 인식기의 평균 인식률은 90.2%이며 후처리를 적용하여 평균 94.2%로 성능을 약 4% 정도 향상시켰다.

실험결과 모델로 선정된 단어들인 경우 인식기 인식률이 평균 85.9%이고 적응형 후처리를 적용한 결과 평균 93.0%로 약 7.1%의 성능을 향상시켰으며, 비 선정 단어의 경우 94.3%를 96.6%로 약 2.3% 향상시켰다. 즉, 모델로 선정된 단어들의 인식률을 높임으로써 비 선정 단어의 인식률도 높아짐을 볼 수 있다.

4.4 후처리 성능 평가

본 절에서는 사용자 적응형 후처리 방법의 성능 평가를 위하여 다른 후처리 방법들과 비교하였다. 그러나 각 논문의 실험 방법이나 적용 환경에 차이로 인하여 인식률을 바로 비교하기 어려우므로 본 논문에서는 식 (3)과 같은 오류 수정률을 사용하여 비교하였다. 여기서 ERR_O는 후처리 적용전의 인식기의 오류율을 의미하며 ERR_+은 후처리 적용후의 오류율을 의미한다.

$$\text{오류 수정률} = \frac{ERR_O - ERR_+}{ERR_O} \quad (3)$$

표 3의 결과에서 볼 수 있듯이 본 논문에서 제안하는 사용자 적응형 후처리 방법이 41.0%의 오류 수정률을 보임으로써 다른 방법에 비해 음성인식기의 오인식을 효과적으로 보정하였음을 알 수 있다.

5. 결론 및 향후 연구

본 논문에서는 신경망을 이용한 사용자 적응형 음성

표 3 오류 수정률 비교

방법 / 평가	오류 수정률	비고
오류 패턴 정보 후처리[1]	8.5%	오류 수 (1,361개 ⇒ 1,245개)
어휘 의미 패턴 정보 후처리[2]	36.7%	인식률 (79.51% ⇒ 87.03%)
문맥 정보 후처리[3]	64.89%	인식률 (81.84% ⇒ 93.57%)
사용자 적응형 후처리	41.0%	인식률 (90.2% ⇒ 94.2%)

인식 후처리 방법을 제안하였으며 타당성 검증을 위한 실험을 하였다. 특히, 인식기의 오류를 분석하기 위해 인식기의 인식 결과 단어뿐만 아니라 인식기의 인식 가능도인 likelihood를 사용하였으며, 그 결과 41.0%의 오류 수정률을 보였다. 이와 같은 결과를 통하여 사용자의 발화 특성은 인식기의 인식 결과에 반영이 되고 이러한 인식 결과인 likelihood 패턴을 학습함으로써 인식기의 오류를 수정할 수 있음을 보였다. 또한 사용자 적응형 후처리는, 사용자의 지속적인 사용으로 오인식이 자주 발생되는 단어를 수정해 감으로써 점점 더 높은 인식률을 보일 수 있다는 장점을 가지고 있다.

그러나 사용자 적응형 후처리 모델을 학습시키기 위해 인식기의 오인식 데이타를 사용하므로, 인식기의 인식률이 비교적 높아 충분한 오인식 데이타를 얻지 못하는 경우 사용자의 발화를 많이 요구한다는 단점이 있다. 따라서 향후에는 사용자 적응형 후처리의 효율성을 높이기 위한 연구와 함께 후처리 모델을 학습시키기 위하여 데이타의 양을 줄이는 방법도 연구되어야 한다.

참 고 문 헌

[1] Satoshi Kaki, Eiichiro Sumita, and Hitoshi Iida, "A method for correcting speech recognition using the statistical features of character co-occurrence," International Conference On Computational Linguistics, vol. 1, pp.653-657, 1998.

[2] Minwoo Jeong, Byeongchang Kim, Gary Geunbae Lee, "Semantic-oriented error correction for spoken query processing," Automatic Speech Recognition and Understanding, IEEE, pp.156-161, 2003.

[3] Myung Won Kim, Joung Woo Ryu, Eun Ju Kim, "Speech Recognition with Multi-Modal Features Based on Neural Networks," International Conference on Neural Information Processing (ICONIP), LNCS 4233, pp.797-806, 2006.

[4] 송원문, 김명원, "문맥 및 사용 패턴 정보를 이용한 음성인식 후처리", 정보처리학회논문지 제13-B권 제5호, pp.553-560, 2006.

[5] Jiawei Han, Micheline Kamber, Data Mining concepts and techniques, pp.303-311, Morgan Kaufmann Publishers, 2001.

[6] Tom M. Mitchell, Machine learning, McGraw-hill international editions, pp.81-127, 1997.

[7] Deller, Hansen, Proakis, "Discrete-time processing of speech signals," IEEE PRESS, pp.677-744, 2000.

[8] M. Ostendorf, "From HMM's to segment models: a unified view of stochastic modeling for speech recognition," IEEE SPA. pp.360-378, 1996.

[9] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, Phil Woodland, The HTK book (for HTK version3.3), Cambridge University Engineering Department, 2005.