공학박사 학위논문

마비말장애 화자의 음성 인터페이스 활용을 위한 어휘모델링 최적화

Optimizing Vocabulary Modeling for Dysarthric Voice User Interface

서울대학교 대학원

협동과정 인지과학 전공

나 민 수

마비말장애 화자의 음성 인터페이스 활용을 위한 어휘모델링 최적화

지도교수 정 민 화

이 논문을 공학박사 학위논문으로 제출함

2015년 10월

서울대학교 대학원
협동과정 인지과학 전공
나 민 수

나민수의 박사 학위논문을 인준함
2016 년 1 월

위 원 장 이 경 민 (인)

부위원장 정 민 화 (인)

위 원 신 효 필 (인)

위 원 김 회 린 (인)

위 원 김 선 희 (인)

국문 초록

마비말장애 화자의 음성 인터페이스 활용에서 빈번한 조음오류, 눌변, 빈번하고 불규칙적으로 발생하는 발화중단, 느린 발화속도 등은 오인식을 유발하는 요인이 된다. 선행연구에서는 장애발화의 음향 및 음운적 특성을 분석하고 이를 기반으로 장애발화 수정, 음향모델 적응, 발음변이 모델링, 문법 및 어휘 모델링 등을 통해 인식오류의 문제를 보완하였다. 본 논문에서는 장애발화의 특징을 반영하여 음향모델을 최적화했다. 또한 어휘모델의 구성에서 음소범주 기반의 조음특징과 인식오류와의 관계를 모형화하여 단어의 선택기준으로 도입했고 단어 간 유사도를 줄임으로써 장애화자의 음성 인터페이스 시 발생하는 인식오류를 줄였다. 마비말장애 화자를 위한 음향모델의 구축을 위해 첫째로 장애화자의 느린 발화속도에 따라 특징추출의 윈도우 크기와 HMM를 구성하는 state 개수를 조정하여 오류를 낮췄다. 둘째로 HMM의 출력확률 모델로서 GMM, Subspace GMM, DNN 등을 도입하여 인식오류를 비교했다. 셋째로 학습데이터 부족문제에 대한 대응방법으로 정상발화 도입의 효율성을 인식실험으로 확인했다. 조음특징과 인식오류율의 혼합선형모델 분석에서 자음범주 중 오류율과 유의수준 0.05 이하에서 상관관계를 보인 범주는 마찰음과 비음이고 모든 모음범주는 오류율과 유의수준 0.05 이하에서 상관관계를 보였다. 또한 혼합모델은 자음을 조음방법으로 범주화할 때가 조음위치로 범주화할 때에

비해 낮은 AIC를 보였고 모음을 혀의 위치로 범주화할 때가 낮은 AIC를 보였다. 음소를 조음방법과 혀의 위치로 범주화했을 때 마찰음은 단어에 포함될수록 인식오류를 높이는 결과를 보였고 비음은 인식정확도를 높였다. 폐쇄음, 파찰음, 유음 등은 인식결과에 대한 영향이 0에 가까웠다. 모든 모음 범주는 인식정확도를 높였다. 그 중 중설모음의 영향력이 가장 컸고 후설모음, 전설모음, 이중모음 순으로 작아졌다. 단어 간 인식오류의 유발 가능성을 Levenshtein 거리와 Cosine 거리 등 열 거리 기준의 단어 간 유사도로 모델링했고 유사도의 최소-최대 비교와 N-best 추정으로 정의된 단어간 유사도가 음성인식에 영향을 주는 요인이 될 수 있음을 확인했다. 장애발화 인식을 위한 어휘모델의 구성에서 먼저 조음특징 기반의 혼합모델로 단어의 조음점수를 계산하고 점수를 최대화하는 단어로 인식단어 리스트를 만들었다. 또한 단어 간 유사도를 최소화하도록 어휘모델을 수정하여 실험한 결과 기존 통화표 방식에 비해 절대적으로 5.7%, 상대적으로 34.6%의 인식오류가 감소하였다.

주제어: 음성인식, 마비말장애, 음향모델, 어휘모델, 조음오류, 단어간 유사도

학 번: 2006-30742

목차

그림 목차

표 목차

제 1 장 서론

최근 컴퓨터, 모바일 단말기, 웨어러블 디바이스, 임베디드 시스템 등 기기의 활용범위가 넓어지며 인간과 기계 사이의 상호작용의 편의성을 높이기 위해 스위치, 버튼, 키보드, 마우스 등 터치기반 입력방식과 음성, 모션 등 인식기반 입력방식 등 입력수단으로서의 사용자 인터페이스가 개발되고 있다. 음성인식은 화자의 말소리를 입력 매개체로 가정하는 음성 인터페이스의 핵심을 이루는 기술로 음성검색, 음성 딕테이션, 차량 인터페이스 등 상용 서비스에 성공적으로 적용되고 있다.

마비말장애는 신경계의 손상으로 인해 입술, 혀, 턱, 성대 등 말소리 산출에 관여하는 기관의 통제에 어려움을 겪는 장애를 지칭한다. 마비말장애 화자, 장애화자의 손상부위는 반드시 조음기관에만 한정되지 않아서 사지의 운동조절 능력 저하를 동반하기도 한다. 사지 마비를 동반하는 장애화자의 경우 세밀한 운동조절이 어렵기 때문에 터치를 통한 입력 정확도와 속도가 정상 사용자에 비해 현저히 떨어진다. 그에 따라 운동이 비교적 정확하게 조정되는 신체부위로 입력 편의성을 높이거나 보완·대체 의사소통 보조장치(Augmentative and Alternative Communication, AAC)의 사용, 장애발화를 인식하여 단어 또는 기능을 입력하는 방법 등이 대안으로 사용되기도 한다. 그러나 마비말장애를 가진 화자가 정상화자를 대상으로 개발된 음성 인식기의 사용하는 경우 빈번한

쥐어짜는 소리, 가래끓는 소리, 기식성, 비음성, 발화속도 감소/불규칙성, 조음오류 증가 등의 원인으로 인해 인식의 오류율이 크게 증가한다.

장애화자의 음성 인터페이스 사용에서 인식의 오류율을 낮추기 위해 입력된 장애발화의 음향적 특징을 변환하여 정상발화와 같이 수정, 음향모델의 학습 데이터로서 장애발화를 수집하고 음향모델을 학습 또는 적응, 장애화자의 발음변이를 관찰하고 발음사전에 추가하거나 음향모델의 위상을 변경하거나 확장, 인식단어의 크기를 제한하거나 입력발화의 구조를 단순화하여 탐색공간의 복잡도를 줄이는 등의 연구가 이루어졌다.

장애발화 인식을 위한 음향모델 학습에서의 문제점은 첫째로 발화속도가 음성인식에 영향을 주는 요인이 되고 (Fosler-Lussier & Morgan, 1999, Mirghafori et al., 1995, Siegler & Stern, 1995) 장애화자의 발화속도가 정상화자에 비해 느리다는 점이다. 둘째로 최근 기계학습 분야에서 딥러닝 아키텍쳐의 연구가 활발히 진행되고 있고 음성인식, 화상/모션 등의 패턴인식 분야에서 기존의 신경망을 사용하지 않는 접근방법에 비해 높은 성능을 보이고 있으나 깊은 층위를 가진 신경망의 학습은 대량의 데이터가 필요함을 전제로 한다. 그에 따라 장애발화인식의 학습 데이터의 크기가 제한적인 상황에서의 신경망 기반 모델을 포함한 HMM의 출력확률모델 간의 효율을 비교할 필요성이 있다.

장애화자의 음성 키보드, 음성 보완·대체 의사소통 보조장치 등의 활용에서 인식단어는 단어의 사전적 의미와 무관하게 특정 기능을

지칭하는 대용어로 사용되어 인식의 정확도를 높이는 방법들이 제안되었다. (Hawley et al., 2013, Hamidi et al., 2010, Kim et al., 2013) 특히 Hamidi et al.(2013)에서는 장애화자의 조음특성을 파악하고 이를 기초로 인식단어를 설정함으로써 음성인식의 정확도를 향상하였다. 장애화자의 음성인식 시 전문지식에 근거한 대용어의 설계가 음성인식의 오류율을 낮추는 방법이 될 수 있음을 보였지만 기준에 대한 구체적인 내용은 제시되지 않았다.

본 논문에서는 어휘모델의 검증에 앞서 장애발화의 인식에서 최소오류를 보이는 베이스라인 음향모델을 학습하고자 했다. 음향모델의 구축을 위해 첫째, 장애화자의 고립어 발화속도를 측정했고 정상화자 발화속도에 비해 느림을 확인했다. 그에 따라 특징추출 시 윈도우 크기와 HMM의 State 개수를 조정했다. 둘째, HMM의 출력확률 모델로서 GMM, Subspace GMM, DNN 등에 의한 인식오류율을 비교했다. 이로부터 분석 셋에 대한 파라미터 최적화 결과가 그와 분리된 테스트 셋에서도 적용됨을 확인할 수 있다. 출력모델을 같은 분량의 데이터로 학습한 비교실험에서 GMM에 비해 Subspace GMM과 DNN이 낮은 오류율을 보였고 SGMM과 DNN간의 차이는 크지 않았다. 셋째, 학습데이터 부족문제에 대한 대응방법으로 정상발화 도입의 효율성을 인식실험으로 확인했다. 어휘모델을 구성하는 단어의 선택을 위해 조음점수 최대화 기준을 선택기준으로서 제안했다. 음소범주의 인식오류에 대한 영향력을

선형혼합모델로 분석했고 이를 바탕으로 단어의 조음점수를 정의했다. 또한 인식단어를 구성하는 단어간 혼동에 의해 유발되는 인식오류를 줄이고자 열간 거리 정의에 기초하여 단어간 유사도를 정의했고 인식단어 리스트에서 단어간 유사도가 큰 단어를 교체하여 인식오류를 줄였다.

본 논문의 구성은 다음과 같다. 2장에서는 마비말장애 발화의 조음오류 특성 및 음성인식, 장애발화 인식의 모델링에 대한 관련연구를 요약한다. 3장에서는 음향모델의 학습 알고리즘과 파라미터를 최적화하여 장애발화 인식을 위한 베이스라인 음향모델을 구축했다. 4장에서는 음소범주가 인식 오류율에 미치는 영향을 통계적으로 분석했다. 5장에서는 열 간 거리를 기반으로 단어 간 유사도를 정의했다. 6장에서는 음성인터페이스에서의 인식단어 구성 방식과 단어리스트 구성방식에 따른 마비말장애 발화 인식실험 과정을 보이고 결과를 분석했고 7장에서 연구결과를 요약하고 향후 보완할 사항을 기술하였다.

제 2 장 관련연구

마비말장애 화자가 정상화자를 대상으로 개발된 음성 인식기를 사용하는 경우 인식률이 크게 하락하기 때문에 장애발화의 특성에 대한 연구와 특성을 반영하여 인식오류율을 낮추기 위한 연구가 진행되고 있다.

본 장에서는 마비말장애 발화의 조음오류분석에 대한 연구를 요약하고 음성인식의 문제정의와 세부모델에 대해 정리한다. 또한 장애발화 인식오류율을 낮추기 위한 연구와 장애발화 음성인터페이스의 개발사례를 소개한다.

제 1 절 마비말장애 발화 조음오류 분석

마비말장애 발화의 조음오류에 대한 연구결과는 실제 발화를 음소 단위로 전사하고 발화 시 의도했던 단어의 음소 열과 정렬하여 계산하는 혼동행렬을 기초로 음소간 조음 정확도와 대치, 삽입, 삭제 등 오류양상을 파악하는 분석방식과 조음오류가 빈번한 마찰음과 같이 한정된 현상에 초점을 두는 분석방식 등으로 분류할 수 있다.

Byrne(1959)에서는 강직형 뇌성마비 어린이 37명과 불수의 운동형 뇌성마비 어린이 37명의 발화로부터 자음과 모음의 혼동행렬을 계산했다. 자음 음소를 조음방법에 따라 구분했을 때 활음, 비음, 폐쇄음 등의 조음오류는 낮았고 유음, 마찰음, 파찰음의 오류는 높았다. 자음 음소를

조음위치에 따라 구분했을 때 양순음, 연구개음 등의 오류는 낮고 순치음, 후치음, 치경음, 경구개음 등의 오류는 높았다. 모음 음소는 평균적으로 자음 음소에 비해 오류가 낮았고 중모음과 저모음의 오류가 고모음에 비해 적었다.

Platt et al.(1980)에서는 17~55세 사이의 강직형 뇌성마비 남성 환자 32명과 불수의 운동형 뇌성마비 남성 환자 18명의 고립어 음성을 음소단위로 전사하고 자음 및 모음 음소의 혼동행렬을 작성했다. 발화된 단어는 22종의 단어의 시작에 위치한 자음, 18종의 단어의 끝에 위치한 자음, 9개의 모음 등 49종의 음소로 구성된다. 자음 음소를 조음방법에 따라 구분했을 때 활음, 비음, 폐쇄음 등의 조음오류는 낮고 마찰음, 파찰음의 오류는 높았다. 자음 음소를 조음위치에 따라 구분했을 때 양순음, 연구개음 등의 오류는 낮고 순치음, 후치음, 치경음 등의 오류는 높았다. 조음하고자 했던 음소와 실현된 음소의 쌍을 비교했을 때 대치오류가 발생한 경우, 두 음소의 조음방법은 서로 일치하지만 무성음화 되거나 조음위치에서 차이가 발생하는 경우가 서로 일치하지 않는 경우보다 많았다. 조음방법이 보존되는 대치오류의 비율은 단어의 마지막 자음에서 더 크게 발견됐다. 음소의 탈락은 단어의 시작보다 끝에서 더 빈번하게 발생했다. 모음 음소를 혀의 높이와 혀의 위치에 따라 모음 사각도로 표시했을 때 /i, æ, a/ 등 사각도의 끝단에 있는 음소의 조음 정확도가 낮았다.

Whitehill & Ciocca(2000)에서는 22명의 광동어 화자의 단음절 단어 발화를 음소단위로 전사하여 자음과 모음 음소의 혼동행렬을 작성했다. 자음 음소를 조음방법에 따라 구분했을 때 단어의 시작 자음에서는 비음, 폐쇄음, 유음 및 활음 등의 조음오류는 낮고 마찰음, 파찰음의 오류는 높았다. 자음 음소를 조음위치에 따라 구분했을 때 단어의 시작 자음에서는 경구개음, 연구개음, 양순음, 성문음 등의 오류가 낮고 순치음, 치경음 등의 오류가 높았다. 단어의 끝 자음에서는 양순음과 연구개음이 낮고 치경음이 높았다. 모음에서는 고모음과 저모음 사이, 그리고 전설모음과 후설모음 사이에서 통계적으로 유의한 조음오류의 차이가 발견되지 않았고 이는 영어권 분석 (Byrne, 1959, Platt et al., 1980)과 다른 결과이다. 저자는 이것을 광동어 모음이 영어 모음에 비해 수가 적기 때문에 모음 공간 상의 거리가 보다 가까워 차이가 두드러지지 않은 것으로 해석한다.

Liu, Tsao & Kuhl(2005)에서는 17~22세 사이의 북경어를 사용하는 남성 뇌성마비 환자 20명의 발화로부터 /i/, /a/, /u/ 등의 포먼트 공간크기와 모음 및 단어의 명료도를 계산했다. 장애화자의 평균적 공간크기가 정상화자에 비해 작은 것으로 관찰되었고 모음공간의 크기와 모음 명료도 간, 모음공간의 크기와 단어 명료도간 0.6이상의 상관관계를 관찰했다.

음성인식의 오류율이 조음 오류율과 비례한다고 가정할 때 조음오류에 대한 선행 연구결과로부터 마찰음과 파찰음 또는 순치음, 치경음보다는

활음, 비음, 폐쇄음 또는 양순음, 연구개음으로 단어를 구성하는 것이 인식오류를 낮출 수 있다는 가이드라인을 만들 수 있다.

그러나 이는 영어, 중국어 화자의 발화에 대한 연구결과로, 한국어와 음소체계가 다르기 때문에 연구결과가 그대로 적용될 것으로 볼 수 없다. 또한 발화의 단어가 단음절로 구성되어 다수의 음소가 단어에 섞여 있을 때 단어 오류율에 대한 개별 음소의 영향을 예측할 수 없고 음소와 음소 사이 동시조음에 대한 고려가 되지 않았다.

Throneburg, Yairi & Paden(1994), Dworzynski & Howell(2004) 등에서는 음소 또는 자음연쇄의 조음 복잡성과 음운발달순서, 음절 개수 등을 고려해 단어의 조음복잡성을 수량화하기 위한 지표를 만들었고 이은주, 한진순 & 심현섭(2004)에서는 한국어의 발달 및 조음 특성에 맞추어 개선하고 음소의 종류, 음절의 형태, 단어의 길이 및 자음연쇄 등으로 조음복잡성을 수치화했다.

표 1. 한국어 조음복잡성 지표 및 배점 기준, 이은주, 한진순 & 심현섭 (2004)

개별 지표	0점	1점
자음의 조음위치	양순음, 치경음, 성문음	치경경구개음, 연구개음
자음의 조음방법	폐쇄음, 비음	마찰음, 파찰음, 유음
모음의 종류	단모음	이중모음
음절의 형태	모음으로 끝남(개방형)	자음으로 끝남(폐쇄형)
어절의 길이	1~2음절	3음절 이상
인접자음의 출현여부	없음	있음
인접자음의 조음위치	같음	다름

따라서 조음 복잡성 지표를 통해 음소의 종류, 음소의 연쇄, 음절의 형태 등 조음 시 오류를 유발할 수 있는 환경에 대한 연구결과를 명시적으로 반영하여 조음오류의 가능성을 정량적으로 예측할 수 있다. 그렇지만 지표의 값이 이진체계로 구성되어 단어의 음절이 길어질수록 지표에 의한 예측값의 단어간 차이가 적어지고 지표의 개별 기준이 조음오류에 미치는 영향의 상대적인 차이가 고려되지 않았다.

제 2 절 음성인식 시스템의 구조와 모델

화자가 단어 열 W 를 염두에 두고 발화하여 음성 X 를 생성했을 때

음성인식은 음성 X를 입력으로 받고 이를 탐색공간에 전달하여 사후확률 P(W|X) 를 최대화하는 단어 열 $\widehat{W}$ 를 찾는 과제로 정의된다. 사후확률 P(W|X)은 베이즈 규칙에 의해 $\frac{P(X|W)P(W)}{P(X)}$ 로 표현될 수 있다.

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)} = \prod_{i=1}^{N} \frac{P(\vec{o_i}|W)P(W)}{P(\vec{o})} \qquad \text{식(1)}$$

전통적인 음성인식 시스템에서는 음향모델과 발음사전을 이용해 P(X|W)의 확률을 계산하고 언어모델과 단어 리스트를 이용해 P(W) 의 확률을 계산한다.

음성인식은 화자와 마이크간의 거리, 단어 리스트의 크기, 발화체의 유형, 발화의 구조, 입력언어의 구성 등에 따라 그 세부환경과 과제의 난이도를 분류할 수 있다(Deng & Yu, 2014). 먼저 화자와 마이크 사이의 거리에 따라 헤드 셋, 1m 이내 근거리 마이크, 1~3m 원거리 마이크 등으로 구분할 수 있다. 거리가 증가함에 따라 신호대잡음비 감소, 폐쇄된 공간에서 반향 발생, 롬바드 효과 발생 등의 원인에 의해 인식오류가 증가하므로 원거리 음성인식 분야에서는 잡음제거, 음성신호 보상, 반향/롬바드 효과 모델링 등의 신호수준의 보상기법이 연구되었다. 둘째로 인식과제는 단어 리스트의 크기에 따라 소어휘, 대어휘로 구분되었고 최근 인식오류의 감소와 연산효율의 증가로 인해 음성검색, 받아쓰기 등 분야에서 자연언어 수준의 단어 크기로 모델을 구축하고 있다. 셋째로 화자가 주어진 단어 또는 문장을 읽는 낭독체 발화와 제약이 없는 자유

발화 등 발화체에 따라 구분할 수 있다. 자유 발화에서는 간투어, 망설임, 발화수정, 반복, 발화의 파편화 등 눌변(disfluency)의 발생빈도와 발화의 속도변화가 비교적 높아 인식오류가 증가한다. 넷째로 입력되는 발화구조에 대한 가정에 따라 고립어 인식과 연결어 인식, 연속어 인식으로 구분할 수 있다. 고립어 인식에서는 하나의 발화가 하나의 단어로 구성되는 것으로 가정한다. 연결어 인식과 연속어 인식에서는 발화가 하나 이상의 단어로 구성되는 것으로 가정하고 연결어에서는 단어와 단어가 사전에 정의된 규칙에 따라 연결되는 것으로 가정한다. 따라서 연결어 인식에서는 규칙에 정의되지 않은 단어 열을 인식할 수 없다. 전통적으로 연속어 인식에서는 단어와 단어 사이의 연결을 학습 데이터로부터 학습하는 확률적 방식으로 모형화했다. 연속어 인식은 연결어 인식에 비해 단어와 단어 사이의 연결이 자유롭기 때문에 탐색공간의 복잡도가 크고 그에 따라 같은 발화에 대한 인식오류가 크다. 다섯째로 음성인식은 입력언어의 구성에 따라 단일 언어, 다국어, 혼합언어 인식 등으로 구분할 수 있다. 단일 언어인식에서는 시스템이 특정된 하나의 언어만을 인식할 수 있고 다국어 인식에서는 여러 언어를 인식할 수 있지만 입력된 발화의 언어 종류에 대한 정보를 주어야 한다.

음성인식 과제는 화자와 마이크 사이의 거리가 멀고 인식단어 리스트의 크기가 클수록 어려워지고 발화가 자유발화, 연속어, 혼합언어 환경일 때 난이도가 높아져 그에 따라 인식오류가 증가하게 된다.

2.1 특징추출

특징은 음성 X로부터 계산된 한정된 차원의 벡터 열, $\vec{O}$ 이다. 음성을 특징벡터 열로 변환하는 이유는 첫째로 음성으로부터 인식과정에 유용한 정보만을 추출하기 위함이고 둘째로 필요 없는 정보를 제거하여 계산 량을 줄이기 위함이다. 특징 코드의 종류는 크게 LPC와 MFCC 등으로 구분할 수 있다. 또한 fMLLR, STC, LVTLN 등과 같이 학습데이터에 대한 모델의 조건부 확률의 최대화를 목표로 한 특징 변환기법이 연구되었다.

2.2 음향모델

음향모델은 단어 W 가 특징벡터 열 $\vec{O}$ 를 생성할 때의 조건부 확률 $P(\vec{O}|W)$ 를 계산하는 역할을 한다. 전통적인 음향모델에서는 Hidden Markov Model을 음향모델의 표상으로 가정한다.

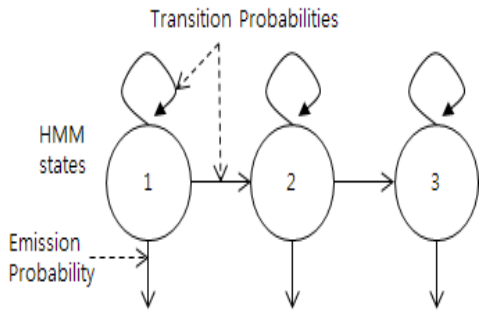


그림 1. Hidden Markov Model 구조

단어 W 는 음소 열 $\vec{P}$ 로 구성되고 각 음소 또는 선행 또는 후행하는 음소문맥의 종류마다 하나의 HMM이 정의된다. 그리고 하나의 HMM은

state의 연결로 구성된다. 각 state는 해당 state가 $\vec{o}$를 생성할 확률을 나타내는 출력확률(emission probability)과 state간의 연결확률(transition probability)로 구성된다.

음향모델의 매개변수(출력확률, 연결확률 등)의 학습에서는 forward-backward algorithm(Huang et al., 2001)이 사용된다. 이때 학습의 목적함수는 사전확률 $P(\vec{o}|W)$의 최대화 함수 또는 사후확률 $P(W|\vec{o})$의 최대화 함수 등이 제안되었다.

음향모델에 대한 연구는 첫째로 음향모델 학습의 목적함수가 조건부 확률을 최대화하는지 또는 사후확률을 최대화하는지에 따라 구분할 수 있다. 사전확률 최대화의 문제점은 학습과정에서 정답 외의 가설이 학습 발화를 생성하는 확률이 높아질 수 있고 이를 견제하지 못한다는 점이다. 이 문제를 보완하기 위해 사후확률을 베이즈 규칙에 의해 식(1)과 같이 표현하고 분모의 사전확률 $P(\vec{o})$를 $\sum_{W'} P(\vec{o}|W')P(W')$로 추정한다. W'는 단어 리스트에 포함된 모든 단어이다. 사후확률은 조건부 확률, 즉 정답 모델이 학습 발화를 생성하는 확률이 높을수록 그리고 오답 모델이 학습 발화를 생성하는 확률이 낮을수록 높아지므로 사후확률을 최대화함으로써 정답과 오답 사이의 변별력을 높인다. 사후확률 최대화 학습에서 목적함수는 MMI, MWE, MPE등이 제안되었다.

음향모델에 대한 연구는 둘째로 출력확률의 계산을 위한 표상의 유형에 따라 구분할 수 있다. Gaussian mixture model 은 랜덤변수 $\vec{x}$가 Gaussian

분포를 따른다고 가정하고 $\vec{x}$ 를 다수의 Gaussian으로 표현한다. 각 Gaussian의 출력확률은 식(2)와 같이 평균 벡터와 분산 행렬로 계산된다.

$$N(\vec{x};\vec{\mu},\sum) = \frac{1}{(2\pi)^{\frac{n}{2}}|\sum|^{\frac{1}{2}}} e^{-[\frac{1}{2}(\vec{x}-\vec{\mu})^t\Sigma^{-1}(\vec{x}-\vec{\mu})]} \quad \text{식(2)}$$

음성은 성별, 나이, 감정 등 다양한 변이로 표현되므로 하나의 state를 다수의 지역해로 표현하고 그에 따라 식(3)과 같이 Gaussian mixture로서 근사화하여 모델링한다.

$$f(\vec{x}) = \sum_{k=1}^{K} c_k N_k(\vec{x};\overrightarrow{\mu_k},\sum_k) \quad \text{식(3)}$$

또한 음성은 문맥적 요인에 의해 영향을 받으므로 GMM-HMM을 주변음소 문맥을 반영하는 단위로 모델링하기도 한다. 본 논문에서는 음소 자음 25개, 모음 21개, 묵음 2개 등 총 48개의 음소를 정의했다(부록 1). 트라이폰으로 음소문맥을 모델링할 경우 학습 데이터로부터 HMM의 평균, 분산, GMM의 가중치 등 약 1천만개의 파라미터 값을 추정해야 한다. 모든 트라이폰의 파라미터를 학습하는 경우 개별 트라이폰 문맥에 해당하는 음성 데이터가 학습 발화에서 고빈도로 발견되어야 하는 문제가 발생한다. 이와 같은 데이터 부족문제에 대처하기 위해서 매개변수를 줄이기 위한 연구로 Subspace GMM 표상이 제안되었다. Povey et al.(2011)에서는 GMM의 개별 state별로 정의되는 파라미터 셋을 그보다 적은 수의 GMM 파라미터와 State별 파라미터, 화자별 파라미터 등으로 분할하여 각 상태마다 정의되는 파라미터의 수를 줄임으로써 GMM의 데이터 부족

문제를 보완했다.

Deep neural network는 2개 이상의 은닉 층을 가지는 신경망이다. Hinton, Osindero & Teh(2006)에서는 pre-training과 fine-tuning의 2단계에 거쳐 은닉 층과 입출력 층으로 구성되는 다층 신경망의 학습 방법을 고안했다. 기존의 신경망 학습방법인 오류 역전파는 은닉 층을 거치며 오류가 점차 감소하는 문제점을 가지고 있어 효과적인 학습이 이루어지지 않았지만 ML 기준학습이 MAP 기준학습결과와 같은 RBM의 likelihood를 층별로 최대화하여 쌓는 방법으로 은닉 층의 파라미터를 초기화함으로써 다층의 은닉층에서의 학습 효율을 높였다. 학습된 신경망은 음소의 사후확률을 출력하므로 베이즈 규칙을 사용해 음소와 관측열의 결합확률을 먼저 추정하고 그로부터 다시 사전확률을 출력하여 HMM의 출력확률로 사용한다. 신경망은 사후확률을 출력하므로 DNN 학습은 변별적 학습(discriminative training)에 포함되지만 음성인식에서 DNN의 출력은 다시 베이즈 규칙에 의해 사전확률로 변환된다. 음성인식에서는 사후확률을 최대화하는 가설의 탐색 시 사전확률 $P(\vec{o})$를 추정하는 방식을 변별적 학습이라고 부르고 있으므로 본 논문에서는 DNN의 학습에서 $P(\vec{o})$의 정보를 추정하는지 여부에 따라 생성적 학습 또는 변별적 학습으로 구분하도록 한다.

또한 음향모델에 대한 연구는 목적함수가 조건부확률을 최대화하는지 또는 사후확률을 최대화하는지에 따라 생성적 학습과 변별적 학습으로 구분할

수 있다. 생성적 학습에서는 관측된 데이터 x의 확률분포함수와 은닉변수 y에 대한 x의 사전확률, 조건부 확률 p(x|y)를 학습한다. 그에 반해 변별적 학습에서는 데이터 x를 카테고리 y로의 분류를 목적으로 한다. 변별적 학습인 MMI 기준의 x와 y간 상호정보는 사후확률 p(y|x)과 같다(Bahl et al., 1986). 따라서 변별학습은 정답 가설의 사전확률을 최대화 하고 기타 가설의 사전확률을 최소화하는 방향으로 음향모델의 파라미터를 학습한다. Young et al.(2006)와 Povey et al.(2011a)에서는 음향모델의 변별적 학습을 위해 정답 사전확률은 학습 데이터에 대한 강제인식결과로부터 모델링하고 오답 사전확률은 학습 데이터에 대한 인식결과로부터 모델링하여 모델 파라미터를 추정한다.

$$F_{bMMI} = \sum_{u} log \frac{p(\overrightarrow{o_u}|s_u)^K p(W_u)}{\sum_W p(\overrightarrow{o_u}|s_u)^K p(W_u) e^{-bA(W,W_u)}} \qquad \text{식(4)}$$

또한 Boosted MMI에서는 정답 w와 오답 w_u간의 음소 또는 state단위 일치도를 나타내는 $A(w, w_u)$를 추가하여 모수 추정에서 정답과의 일치도에 의한 영향을 줄이도록 했다.

2.3 발음모델

단어의 발음은 표기로부터 확정될 수 없고 같은 표기라 하더라도 형태소의 종류, 주변 단어의 문맥 등에 따라 다르게 발음될 수 있다. 음성인식에서는 인식단어를 자소 열로 표기하기 때문에 하나의 단어가 하나 이상의 발음으로 실현될 수 있다. 예를 들면 명사 “감기”는 /감기/로 발음되지만

어간+어미의 조합인 “감+기”는 /감끼/로 발음된다. 따라서 음성인식의 발음사전은 개별 표제어에 대해 하나 이상의 발음변이를 기록한다.

발음변이는 CMU 발음사전, PRONLEX, BEEP 발음사전 (Weide, 2005, Kingsbury et al., 1997, Robinson, 1996) 등과 같이 전문가에 의해 정의되거나 자소-음소 자동변환 시스템에 의해 생성된다. 또한 전문가에 의해 생성된 발음사전에 포함되지 않는 단어의 발음변이만을 자소-음소 자동변환 시스템으로 생성하기도 한다.

발음변이의 자동 생성은 음운 및 형태론적 지식을 가진 전문가가 정의한 음소변동규칙에 기반하거나 데이터로부터 추상화된 지식에 기반한다. 발음변이의 학습 데이터는 음성 데이터를 전문가가 전사한 음소 열 또는 음성 데이터를 음소 단위로 인식하거나 강제 인식한 결과 등으로 구분할 수 있다. 발음변이의 생성에는 음성 데이터에서 관찰된 발음변이, 음소변동규칙, 결정트리, 신경망, 음소혼동행렬 등이 활용된다.

발음사전에 발음변이를 추가하는 것은 탐색공간의 구성에서 해당 음소 열을 공간에 추가하는 것과 같다. 화자가 해당 발음변이에 따라 발화할 때 추가된 공간에서의 인식 시 점수가 증가하여 변이가 추가된 단어의 인식 정확도가 높아질 수 있다. 그러나 변이가 추가된 단어가 오답인 상황에서는 확장된 탐색공간이 오답을 첫 번째 인식결과로 선택하도록 계산을 방해, 즉 혼동성을 높이는 역할을 하기도 한다. 따라서 발음사전에 추가할 발음변이의 선택에 따라 인식 정확도가 향상되거나 악화될 수 있다.

발음변이의 선택기준은 발음변이의 발견빈도, 적용된 음소변동규칙의 발견빈도, 특정 발음변이를 가정했을 때의 음향모델 사전확률, 단어의 발견빈도, 단어의 음소단위 엔트로피, 단어간 발음변이의 유사성, 단어의 빈도와 발음변이의 발견빈도 등이 제안되었다.

2.4 언어모델

언어모델은 식 (1)에서 단어연결 W가 발생할 확률인 P(W)를 추정하는 역할을 한다. 인식과제의 구분에서 고립어 및 연결어 인식에서는 P(W)를 균등분포로 가정하여 사후확률의 계산에서 소거할 수 있다. 연속어 인식에서는 텍스트 데이터로부터 단어 열의 발생확률을 추정하는 N-gram 방식의 언어모델을 사용한다. N-gram 언어모델에서 추정해야 할 매개변수의 개수는 인식단어 개수 V의 단어 열의 개수 N승, 즉 V^N개 이다. 그러나 텍스트 데이터에서 단어 열 간 발견빈도의 차이에 의해 V^N 단어 수준의 코퍼스를 갖추고 있더라도 단어 열이 한번도 발견되지 않는 데이터 부족 문제가 발생할 수 있다. 언어모델 학습에서 데이터 부족 문제를 해결하기 위해 백오프, 디스카운팅, 선형 보간법 등 방법이 제안되었다.

제 3 절 음성인식시스템의 활용분야

음성 인식기술을 기반으로 하는 음성 인터페이스는 대화 시스템, 차량용 네비게이션, 음성 검색, 가전제품 제어, 받아쓰기 시스템, 자동 통역기

등의 분야에서 상용화되어 쓰이고 있다.

마비말장애 화자를 위해서는 인식오류의 문제로 보다 기능의 종류와 발화의 구조, 인식단어 크기 등의 측면에서 보다 제한적인 음성 인터페이스가 개발되었다.

[Hawley 2013]에서는 음성인식기와 음성합성기를 연동하여 중도 마비말장애 화자를 대상으로 한 보완·대체 의사소통 보조장치 형식의 인터페이스를 개발했다.

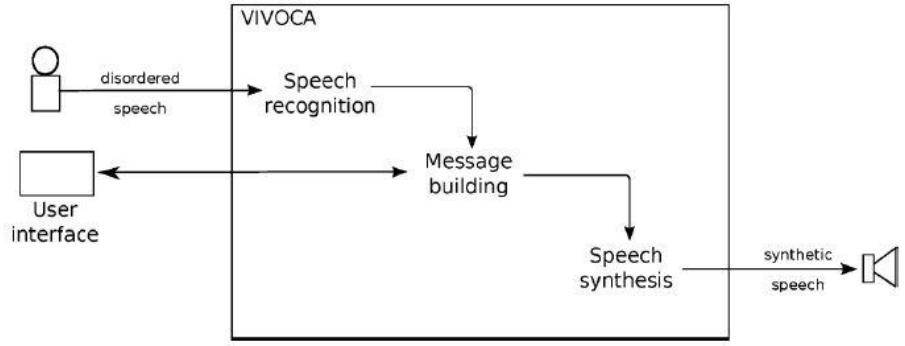


그림 2. VIVOCA 시스템 구조도, Hawley et al.(2013)

단어 인식, 통화표 단어를 인식의 대용어로 사용하는 스펠링 인식, 음성 보완·대체 의사소통 보조 등의 기능을 지원하고 약 50단어 이하의 소어휘 고립어 인식 테스트에서 평균 96% 수준의 인식 정확도를 보고했다.

Gemmeke et al.,(2013)에서는 사용자가 정의한 단어와 문장 셋으로부터 문법 구조를 추론하고 기기의 기능과 연동하여 사용할 수 있는 음성

인터페이스를 개발했다.

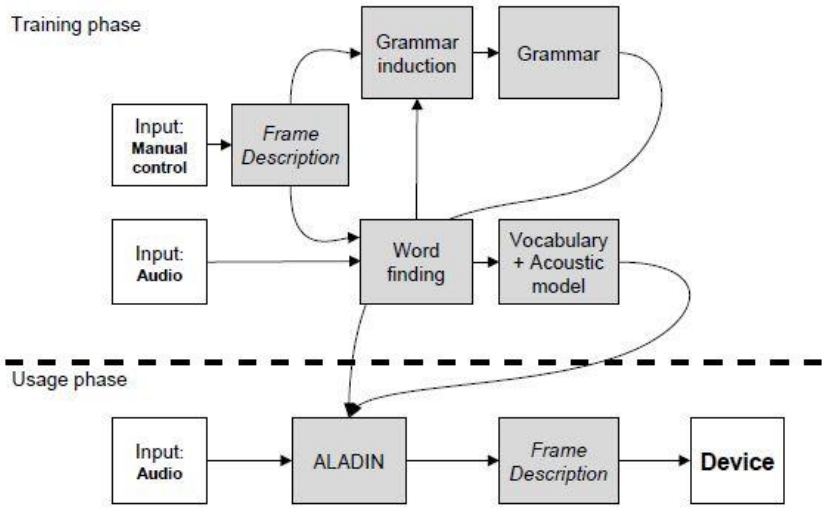


그림 3. ALADIN 시스템 구조도, Gemmeke et al.(2013)

또한 시스템은 장애를 가진 사용자에 의해 정의된 문장을 인식할 수 있도록 음향모델과 어휘/발음사전, 문법 등의 모델을 그에 맞추어 수정하고 홈 오토메이션, 카드 게임 등의 어플리케이션에 적용했다.

Hamidi et al.(2010)에서는 디지털 형식의 입력이 필요한 어플리케이션의 키 또는 스위치의 입력을 음성 입력으로 대체하는 방식의 장애화자 대상의 음성 인터페이스를 개발했다. 대용어를 사용한 스펠링 입력을 지원하고 47단어의 소어휘 고립어 방식으로 발화를 인식한다.

그림 4. CanSpeak 단어리스트(좌)와 Wepspeech 인터페이스(우) 외형, Hamidi et al.(2010)

사용자의 보호자와의 면담을 통해 개별 사용자의 발화특성을 파악하고 이를 바탕으로 대용어를 설계하여 음성인식의 정확도를 높일 수 있음을 보였다.

Kim et al.(2013)에서는 한국인 마비말장애 화자의 기기사용을 보조하기 위한 음성 키보드 기반의 사용자 인터페이스를 개발했다.

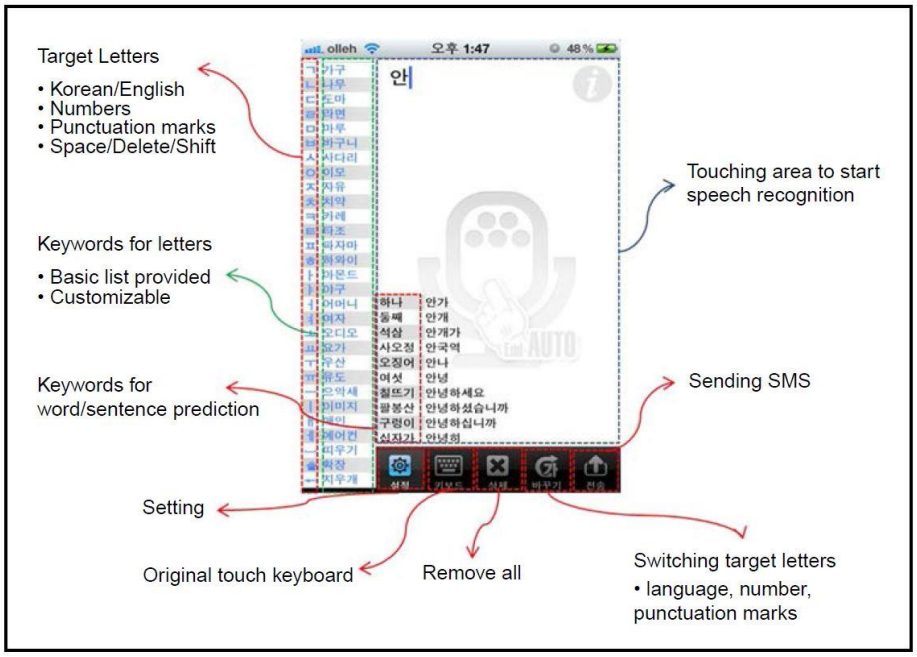


그림 5. 음성키보드 인터페이스 외형, Kim et al.(2013)

장애화자 그룹에 사용하고자 하는 기기와 기능, ASR 기술에 대한 선호도, 성능에 대한 요구사항 등의 항목에 대해 설문조사를 시행해 음성인식기술을 활용한 인터페이스의 필요성을 확인하고 활용기기와 응용을 결정했다.

제 4 절 마비말장애 발화 인식

마비말장애 발화 인식에서 오류를 유발하는 요인 중 첫째는 음향모델 학습의 데이터가 되는 장애발화의 음성특성이 정상화자의 특성과 차이를

보인다는 것이다. 음성인식의 핵심요소 중 하나인 음향모델은 구축을 위해 많은 분량의 음성 데이터를 필요로 한다. 장애발화 수집은 정상발화에 비해 물리적, 비용적 제한이 크기 때문에 음향모델 구축에 충분한 양의 장애음성을 확보하기가 어렵다. 또한 장애발화가 발성, 공명, 조음, 운율 등(Duffy, 2013) 측면에서 정상화자에 비해 큰 변이를 보이기 때문에 정상화자의 발화로 구축한 음향모델의 사용은 오인식의 원인이 된다. (Hux et al., 2000, Raghavendra, Rosengren & Hunnicutt, 2001, Christensen et al., 2012)

장애발화의 음향적 특성에 대한 연구결과를 기초로 정상발화와 가깝게 장애발화를 수정하는 연구가 있다. Rudzicz(2013)에서는 음소의 지속, 자음의 부적확한 조음, 모음 조음에서의 왜곡 등의 특성에 포커스를 두었다. 첫째로 무성자음의 발화에서 잘못된 유성음화 현상을 관찰했고 무성자음 발화를 고역필터에 통과시켜 250Hz 이하의 주파수 대역을 제거함으로써 잘못된 유성음화를 수정했다. 둘째로 장애화자의 공명음 발화속도가 정상발화에 비해 2배 가량 느리기 때문에 공명음 발화 구간의 연속성이 보존되도록 발화시간을 줄였다. 셋째로 장애발화의 모음 영역이 정상발화에 비해 좁기 때문에 선행연구의 포먼트 분석결과에 맞추어 모음의 포먼트를 수정하였다. Tolba & Torgoman(2009)에서는 음성의 떨림, 모음 조음 시의 왜곡 등의 특성에 포커스를 두었다. 장애발화 모음의 포먼트를 정상발화의 포먼트에 근접하게 변환하고 변환된 값에 따라

발화를 재합성하였다. 또한 고빈도 성대 떨림에 의한 에너지의 변이에 초점을 두었고 영위상 필터링으로 떨림을 감소시켰다. 이와 같은 지식 기반 접근방식은 무성자음에서의 유성음화, 공명음에서의 발화속도 감소 등 세부적인 현상을 모델에 반영할 수 있는 장점이 있지만 세부적인 만큼 구현이 어렵다는 단점이 있고 테스트 화자가 어떤 특성을 보일지 미리 알 수 없기 때문에 실제 테스트에서 나타나지 않을 현상에 대비한 수정이 과적용 될 수 있다는 위험이 있다.

적은 양의 장애 음성 데이터로 음향모델을 수정하기 위해 MLLR, MAP 등 모델의 화자 적응기법(Leggetter & Woodland, 1995, Gauvain & Lee, 1994)이 활용될 수 있다. 비교적 취득이 용이한 정상발화로 베이스라인 음향모델을 구축하고 취득이 어려운 장애화자 발화는 적응용 데이터로 사용하여 베이스라인 음향모델을 개별 장애 화자에 맞도록 적응하는 방법으로 장애발화 인식에 적용하여 인식률의 향상을 관찰했다.(Morales & Cox, 2009, Mengistu & Rudzicz, 2011, Christensen et al., 2013) 인식오류를 유발하는 두번째 요인은 빈번하게 발생하는 조음오류이다. 조음오류는 개별 단어의 기준 발음열과 해당 단어의 실제 발음을 비교할 때 음소가 대치, 왜곡, 삽입, 삭제되어 조음되는 오류를 지칭한다(Duffy, 2005). 조음오류에 대응하기 위해 조음오류를 발음변이로 표현할 수 있다. (Morales & Cox, 2009, Mengistu & Rudzicz, 2011, Christensen, Green & Hain, 2013)등에서 장애발화에서 발생하는 조음오류를 발음변이로서

발음사전에 추가하였다. Morales & Cox(2009), Matsumasa et al.(2009)에서는 장애 화자의 음소 삽입, 삭제, 대치 등 오류패턴을 음소혼동행렬 형태로 모델링했고 이 행렬을 음향모델과 혼합한 메타모델을 구성하여 인식률을 높였다. 장애발화에서는 조음오류 외에도 불규칙적인 묵음 삽입 또는 발화 중단 등이 빈번하게 발생한다. Polur & Miller(2006)는 10 state HMM과 신경망의 결합으로 단어 단위 음향모델을 구성하여 발화 내 묵음, 발화 중단 등 비정상적인 상황에 대한 대응력을 높였다.

인식오류를 유발하는 세번째 요인은 문장의 복잡한 구조이다. 인식기가 허용하는 인식단어의 연결 구조가 복잡하다면 탐색공간에 정답 가설과 경쟁하는 오답 가설이 증가하여 오인식이 늘어나게 될 수 있다. Hawley et al.(2007)에서는 TV, hi-fi 등 시스템의 음성 제어를 위해 문법 구조를 고립단어 구조로 제한하고 인식단어크기를 소어휘로 제한했다. Gemmeke et al.(2013)에서는 홈 오토메이션, 카드 게임 등의 음성 입력을 위해서 사용자가 정의한 단어 셋과 문법으로 표현되는 문장만을 인식하도록 제한하였다. Hamidi et al.(2010)에서는 음성 키보드를 위해서 알파벳 단어와 제어 명령어를 포함한 47개 단어를 인식할 수 있는 고립단어 인식기를 사용한다.

인식오류를 유발하는 네번째 요인은 가설간 혼동성이다. 발음열 모델링에서 발음변이의 추가는 해당 단어의 인식률을 높일 수 있지만 다른

단어와 혼동을 유발하여 오인식이 발생할 수 있다(Strik et al., 1999). Slobada & Waibel(1996) 에서는 음성 데이터를 근거로 음소혼동확률을 계산하고 혼동확률이 낮은 발음변이를 사전에 추가하여 인식률을 향상했다. Torre et al.(1997)에서는 음성 데이터에서 음소혼동확률을 계산하고 이를 기반으로 단어 간 혼동확률을 정의했다. Tsai, Chou & Lee(2007)는 같은 발음변이를 가지는 서로 다른 단어에 의한 혼동을 줄이기 위해 단어의 발견빈도를 기준으로 발음변이를 선택했다.

본 논문에서는 화자적응으로 장애발화의 특징을 음향모델에 반영하고 인식과제를 소어휘 고립어로 제한함으로써 인식오류의 문제에 대응했다. 또한 음향모델과 인식단어 리스트, 즉 어휘모델을 최적화함으로써 장애화자의 음성인터페이스 활용에서의 인식오류율을 낮추었다.

제 3 장 특징추출 및 음향모델 베이스라인 구축

음향모델의 학습 데이터와 테스트 데이터 간의 불일치는 인식률의 하락으로 연결된다. Hux et al.(2000)에서는 마비말장애 경도 화자 5명의 문장 발화를 Microsoft Dictation, Dragon NaturallySpeaking, VoicePad Platinum 등 범용 시스템으로 인식했고 정상발화보다 25% 이상 높은 오류율을 보였다. Hawley et al., 2007에서는 지체장애를 수반하는 중도 마비말장애 화자의 홈 오토메이션 시스템 활용을 돕기 위해 소어휘 음성 인터페이스를 도입했고 70단어 80% 수준의 인식 정확도를 보고했다. 동일한 장애화자의 발화를 청취에 의해 전사하고 정확도를 측정하고 같은 발화의 음성 인식률을 비교한 결과 음성 인식률이 청취 정확도보다 평균적으로 낮았고(Mengistu & Rudzicz, 2011) 청취 정확도와 음성 인식률 간의 비례 관계가 관찰되었다(Thomas-Stonell, 1998, Raghavendra, Rosengren & Hunnicutt, 2001, Sharma & Hasegawa-Johnson, 2013). 즉 장애화자의 음성인식률은 정상화자에 비해 낮았고, 장애정도가 클수록 인식률은 낮았다.

마비말장애 발화 인식의 효율을 높이기 위해서는 2장 2절에서 언급된 음향모델, 발음사전, 언어모델 또는 문법 등 수준에서의 장애발화 특성 반영이 필요하다. 본 장에서는 어휘모델링 최적화를 위한 선행연구로서 효과적인 음향모델 구축을 위해서 장애발화의 문제점을 분석하고 이를

해결하기 위한 음향모델 기법과 파라미터의 수치를 실험을 통해 결정하고 결과를 분석한다.

제 1 절 접근방법

마비말장애 발화 인식을 위한 음향모델의 구축과정에서의 문제점은 첫째로 장애화자의 조음속도가 정상화자에 비해 느리고 그에 따라 음소의 조음시간이 길다는 점이고 둘째로 특징추출과 음향모델의 학습절차와 파라미터의 값에 따라 테스트 시 오류율이 달라질 수 있다는 점이다. 셋째로 학습 데이터인 장애발화가 부족하다는 문제점이 있다.

장애화자의 긴 조음시간에 의한 인식 오류율을 줄이기 위해 먼저 장애화자와 정상화자의 고립어 발화속도와 조음시간을 비교했다. 시간의 차이에 따라 특징추출 시 푸리에 변환의 입력음성구간을 확장하고 HMM을 구성하는 state의 개수를 늘림으로써 음향모델의 학습과 인식 시 확장된 탐색공간을 할당하여 정상화자에 비해 긴 조음시간을 모델링했다.

학습절차와 파라미터의 범위에 따라 인식성능이 영향을 받을 수 있으므로 장애발화 셋에 대한 인식결과를 바탕으로 최적의 학습절차와 파라미터의 값을 탐색하고 분리된 테스트 셋에 대한 탐색결과의 유효성을 검토했다. 세부적으로 HMM의 출력확률 모델로서 GMM, Subspace GMM 및 DNN 각각의 파라미터를 조절하여 최소 인식오류를 보이는 지점을 탐색했다. 또한 GMM, Subspace GMM 및 DNN의 인식 오류율을 비교하여 가용한

학습 데이터가 적은 장애발화 인식에 적합한 모델을 결정했다.

제 2 절 개발 환경

2.1 음성 데이터

표 2과 같이 음성 데이터는 음향모델 학습 셋과 파라미터 최적화를 위한 분석 셋, 음향모델 성능 확인을 위한 테스트 셋 등 3개 셋으로 구분된다.

학습 셋은 600명의 정상화자가 100문장씩 발화한 음소균형 연속어 코퍼스 또는 30명의 장애화자가 500단어씩 발화한 음소균형 고립어 코퍼스로 구성된다. 500단어는 105개의 기기제어명령어와 108개의 통화표 단어, 287개의 음소균형 단어로 구성된다.

분석 셋은 15명의 장애화자가 173단어씩 발화한 고립어 코퍼스이고 발화 단어는 음향모델 학습 셋의 500단어와 중복되지 않는다. 173단어는 100개의 기기제어명령어와 36개의 통화표 단어, 37개의 음성분석용 단어로 구성된다.

테스트 셋은 15명의 장애화자가 173단어씩 발화한 고립어 코퍼스이고 발화 단어는 분석 셋의 173단어와 일치하고 발화 화자는 분석 셋의 15명 화자와 중복되지 않는다.

음성녹음 시 사용된 마이크는 SHURE SM12A 이고 음성은 16kHz, 16 bit PCM 형식으로 녹음되었다.

표 2. 음성 데이터

셋	발화구조	음소균형	#화자	#발화수 /1 화자	시간	장애정도별 화자 수
학습	연속어	PBS	600	100	120	정상
	고립어	PBW	30	500	11.4	17/10/3
분석	고립어	-	15	173	1.8	5/6/2/2
테스트	고립어	-	15	173	1.7	6/5/3/1

2.2 음성인식 환경 및 음향모델 학습절차

특징벡터의 윈도우의 크기는 25ms에서 45ms 사이에서 조정했다. 윈도우를 10ms씩 전진하여 MFCC 또는 PLP 계수를 추출했고 계수에 Cepstal mean and variance normalization, Splicing/Linear discriminant analysis, Semi-tied covariance, feature-space maximum likelihood linear regression 등의 변환 기법을 적용했다. 상태를 공유하는 트라이폰을 단위로 음향모델을 정의했다. 각 트라이폰은 시작과 끝 상태를 제외한 3~5개의 state로 구성된다. 발음사전은 1개의 단어마다 정상인 화자를 기준으로 한 1개의 기준 발음열을 포함하도록 구성했다. 하나의 발화가 하나의 단어로 구성되는 고립어 형식으로 총 673단어를 인식할 수 있는 네트워크를 정의했고 음향모델의 학습과 인식을 위해서 Kaldi toolkit을 사용했다.

특징벡터의 추출을 위해 먼저 개별 음성에 대한 MFCC 또는 PLP 계수를

계산하고 CMVN으로 정규화했다.

GMM 모델은 모노폰 단위의 모델로부터 상태를 공유하는 트라이폰 단위의 모델로 확장했다. 각 단계로부터 출력된 GMM으로 학습 데이터를 강제정렬하여 발화 별 그래프를 생성하고 이를 다음 단계의 GMM 학습과정에서 교사로서 입력한다. 또한 발화의 특징벡터와 주변문맥의 특징벡터를 결합하고 주요한 정보를 압축하는 LDA와 학습 발화의 조건확률을 높이는 방향으로 학습되는 STC 등 특징변환 기법을 학습 벡터열에 단계 별로 적용한다.

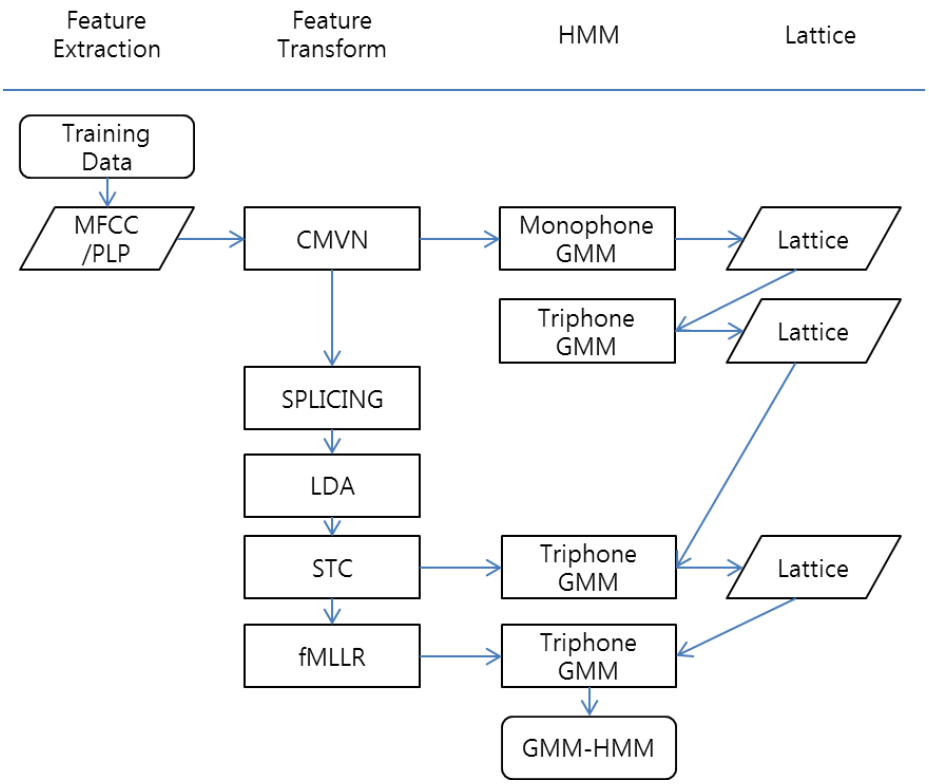


그림 6. GMM 학습절차

Subspace GMM과 DNN 모델의 학습과정은 그림 7, 그림 8과 같고 GMM 모델을 초기 그래프 생성에 활용하고 GMM 모델의 학습과정에서 생성된 변환을 학습 특징열에 그대로 적용한다.

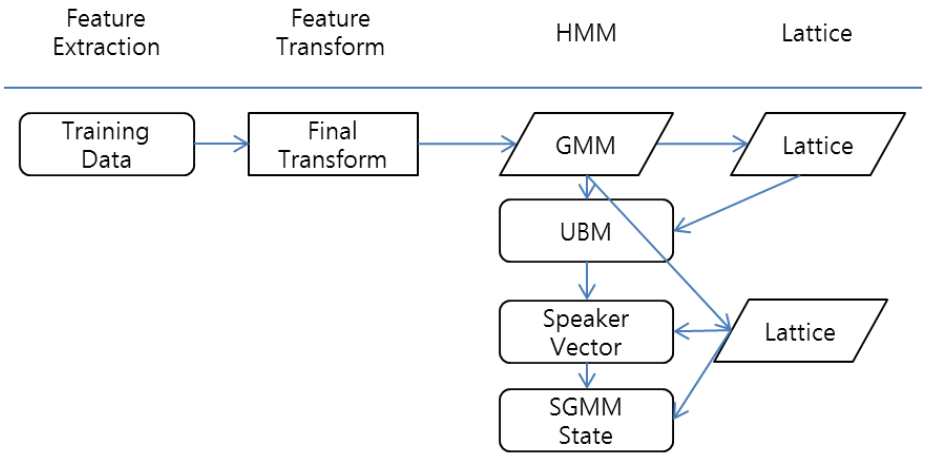


그림 7. Subspace GMM 학습절차

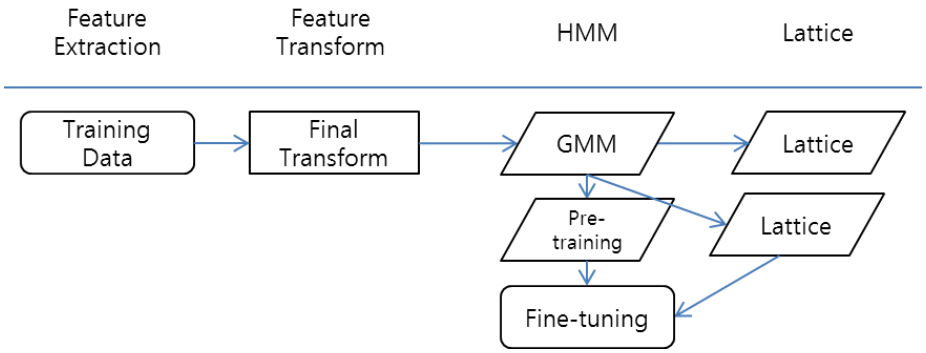


그림 8. Deep Neural Network 학습절차

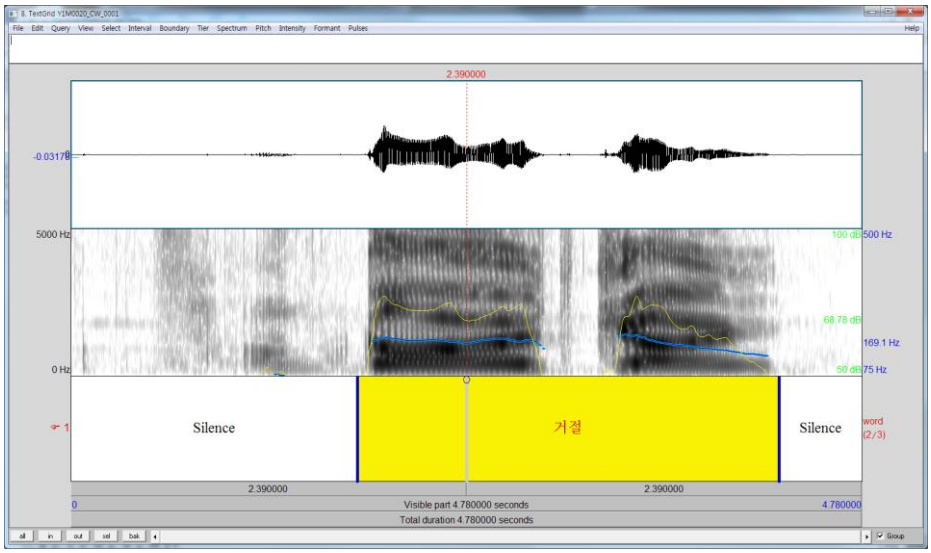
제 3 절 인식실험

3.1 발화속도 모델링

발화속도는 단위시간 동안 조음한 음소의 개수이고 음소당 평균 조음시간은 발화속도의 역수로 정의한다.

$$\text{발화속도, Phones Per Second} = \frac{\sum \textit{조음된 음소 개수}}{\sum(\textit{발화 끝 시간} - \textit{발화시작 시간})}$$

발화속도의 측정을 위해 필요한 음소 개수, 발화의 시작과 끝 시간 등의 정보는 해당음성에 대한 강제정렬의 전사결과로부터 검출했다.



"Y2M0020_CW_0001.lab"
0 17300000 sil
17300000 18200000 K+AX
18200000 28100000 K-AX+Z
28100000 33200000 JX-Z+JX
33200000 37900000 Z-AX+JO
37900000 42800000 AA-L
42800000 42800000 sp
42800000 47800000 sil
.

그림 9. 강제정렬 예제

강제정렬 예제는 발화 "Y2M0020_CW_0001"의 정렬결과, 0~1.73초 구간의 음소는 silence 이고 1.73~1.82초 구간의 음소는 초성 ㄱ임을 나타낸다. 그에 따라 조음된 음소의 개수는 초성 ㄱ, ㅓ, 초성 ㅈ, ㅓ, 종성 ㄹ 등 5개이고 발화의 시작시간은 1.73초, 발화 끝 시간은 4.28초임을 알 수 있고 초당 음소조음횟수를 5 / (4.28-1.73) = 2.0개로 계산한다.

장애화자와 정상화자의 발화속도를 비교하기 위해서 분석 셋에 포함된 15명의 발화속도와 같은 단어 리스트를 발화한 정상화자 15명의 발화속도를 측정했다.

표 3. 장애화자 및 정상화자의 발화속도 비교

장애정도	발화속도	평균속도	장애정도	발화속도	평균속도
정상	7.5	7.7	경도, 1	5.7	5.1
	7.1			4.1	
	7.3			4.9	
	7.2			5.2	
	7.1			5.6	
	8.5		경도-중도, 2/3	3.5	4.6
	6.5			4.4	
	8.3			5.7	
	8.0			4.9	
	7.7			3.4	
	7.7			4.4	
	7.9			4.6	
	8.2			5.6	
	7.4		중도, 4	4.8	4.3
	8.4			3.8	

장애화자의 고립어 173발화 평균 발화속도는 4.7로 정상화자의 평균 7.7에 비해 61% 수준으로 느리고 음소 별 조음시간은 1.6배 길다는 것을 알 수 있다.

특징벡터의 윈도우 크기는 푸리에 변환 시 입력되는 음성구간의 크기로 경험적으로 20~30ms 사이로 정의한다(Huang et al., 2001). 장애발화의 평균적인 음소 지속구간이 넓기 때문에 발화 내에 음소의 종류를 판단할 수 있는 정보가 그만큼 분산되어 나타날 수 있음을 고려한다면 윈도우의

크기를 증가시켜 한 프레임의 특징 계산에 반영되는 음성신호의 범위를 넓히는 것이 음향모델의 정확도를 높이는 방법이 될 수 있다. 정상발화 인식에서 25ms가 주로 사용되고 장애화자의 조음시간이 1.6배 길기 때문에 윈도우의 크기를 25ms~45ms 사이에서 5ms 간격으로 조정하고 결과를 비교했다.

추출된 특징벡터는 HMM의 state 간의 연결로 구성된 탐색공간에 전달되어 학습 또는 인식에 필요한 파라미터의 계산을 위해 입력되고 특징벡터는 현재 계산 중인 state와 직접 연결된 state들에만 전달될 수 있다. 따라서 탐색공간을 구성하는 state의 전체길이와 state간 연결확률은 발화의 지속시간 또는 속도에 영향을 주는 요인이 되고 하나의 HMM을 구성하는 state의 개수를 증가시켜 느린 장애발화 속도를 음향모델에 반영했다. 예를 들어 State의 개수를 5개로 정의했을 때 각 HMM의 구조는 그림 10과 같다.

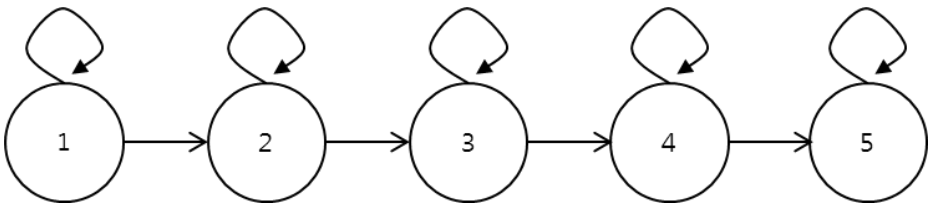


그림 10. 5개의 state로 구성된 HMM 구조

정상발화 인식에서 HMM은 시작과 끝의 state를 제외하면 3개의 state로 구성되므로 HMM 당 state의 개수를 3개~5개 사이에서 1개 간격으로

조정하고 결과를 비교했다.

음향모델의 학습 횟수가 최적화 대상이 되는 파라미터 종류의 수에 따라 기하급수적으로 증가하기 때문에 가능한 모든 조합으로 모델을 구축하고 최소 인식오류율을 보이는 파라미터를 찾는 전역해 탐색이 아닌 각 최적화 단계에서 최소 인식오류율을 보이는 파라미터를 찾는 지역해 탐색으로 최적화 파라미터로 선택했다.

표 4. 특징계수 종류 비교실험결과

Coding	Window Length	#State	#GMM	WER
MFCC	25	3	1	44.0%
PLP				44.5%

특징벡터의 계수로서 MFCC와 PLP를 비교했고 MFCC가 보다 낮은 오류율을 보이므로 이후 모델의 구축에서 MFCC 계수로 특징을 추출했다.

표 5. 윈도우 크기 비교실험결과

Coding	Window Length	#State	#GMM	WER
MFCC	25	3	1	44.0%
	30			43.8%
	35			43.8%
	40			42.9%
	45			44.0%

윈도우 크기를 조정한 실험에서는 40ms에서 가장 낮은 인식오류율을 보였기 때문에 이후 특징추출의 윈도우 크기를 40ms로 설정했다.

표 6. HMM당 State 개수 비교실험결과

Coding	Window Length	#State	#GMM	WER
MFCC	40	3	1	42.9%
		4		41.8%
		5		43.0%

HMM당 state의 개수를 조정한 실험에서는 state 4개에서 가장 낮은 인식오류율을 보였다.

3.2 출력확률 모델 파라미터 최적화

음향모델의 출력확률은 GMM, Subspace GMM, DNN 등으로 모델링하고 인식오류율을 비교했다.

GMM 모델에서는 모델에 대한 파라미터 크기의 영향을 관찰하기 위해서 음향모델의 Gaussian 및 state의 개수를 조정하여 인식오류율을 측정했다. 또한 데이터 부족문제를 보완하기 위해서 트라이폰의 상태공유를 위한 클러스터링 방법(Povey et al., 2011)을 사용한다. 그에 따라 GMM모델에서 조정할 파라미터는 트리구조 클러스터링의 단말과 노드의 개수이다. 단말 개수는 상태공유 후 중복을 제거한 상태 개수를 나타내고

노드 개수는 Gaussian의 개수를 나타낸다. 정상화자의 대어휘 발화과제인 Paul & Baker(1992), Pallett, Fiscus & Garofolo(1992) 등 인식설정에 맞추어 단말과 노드의 수를 정의하는 경우보다 인식과제의 규모가 작은 Tidigit, Leonard & Doddington(1993) 인식설정에 맞추었을 때 낮은 오류율을 보였다. 그에 따라 인식과제의 규모가 작은 Tidigit를 기준으로 단말과 노드의 개수를 정의하고 이들의 0.5배, 1~5배 등으로 크기를 조정하여 GMM모델을 구성한다.

SGMM의 학습에서도 상태공유 클러스터링 방법을 사용한다. SGMM에서 조정할 파라미터는 클러스터의 단말과 노드 개수이다. 노드 수는 음향모델을 구성하는 공통 Gaussian의 개수이고 단말 수는 파라미터 공유 후 중복을 제거한 상태의 개수, 하위상태 수는 state별 파라미터의 개수를 나타낸다. GMM과 같이 정상발화의 파라미터 수를 조정하고 0.5배, 1~5배 등으로 크기를 조정하여 모델을 구성한다.

Povey et al.(2011)에서는 영어권 정상발화인 Resource Management 코퍼스에서의 DNN 학습을 위해 층별 1,024개의 유닛으로 구성된 4층의 은닉층으로 신경망 구성했고 WSJ 코퍼스에서는 2,048개, 6층으로 구성했다. 이를 기준으로 유닛의 개수를 256개, 512개, 1,024개, 2,048개 등으로 늘려서 결과를 비교했다.

표 7. GMM 개수 비교실험결과

Coding	Window Length	#State	#GMM	WER
MFCC	40	4	WSJ	44.0%
			0.5	46.2%
			1	41.8%
			2	40.7%
			3	40.2%
			4	41.7%
			5	41.0%

GMM 모델의 state 및 GMM의 개수의 조정 실험에서 TIDIGIT 실험설정의 3배 크기에서 가장 낮은 인식 오류율을 보였다.

표 8. SGMM 파라미터 크기 비교

Coding	Window Length	#State	#GMM	#SGMM	boosting factor	WER, MLT	WER, DT
MFCC	40	4	3	0.5	0.1	43.6%	41.0%
				1		41.2%	39.8%
				2		39.9%	38.7%
				3		37.8%	38.0%
				4		38.4%	38.8%
				5		38.5%	39.0%

표 9. Boosting factor 비교실험결과

Coding	Window Length	#State	#GMM	#SGMM	boosting factor	WER, MLT	WER, DT
MFCC	40	4	3	3	0.1	37.8%	38.0%
					0.2	37.8%	37.8%
					0.3	37.8%	37.9%
					0.4	37.8%	37.8%
					0.5	37.8%	37.8%

Subspace GMM 모델의 state 및 GMM의 개수의 조정 실험에서 TIDIGIT 실험설정의 3배 크기에서 가장 낮은 인식 오류율을 보였다.

또한 변별적 학습에서 식의 Boosting factor를 0.1~0.5 사이에서 조정했을 때 조정된 실험값 사이에서 큰 차이가 발생하지는 않았다.

표 10. Hidden layer 층수 비교실험결과

Coding	Window Length	#State	#GMM	#Layer	#Unit	WER, MLT	WER, DT
MFCC	40	4	3	3	1024	40.6%	39.0%
				4		39.3%	39.4%
				5		39.5%	38.6%
				6		39.6%	39.3%

표 11. Hidden layer 당 unit 개수 비교실험결과

Coding	Window Length	#State	#GMM	#Layer	#Unit	WER, MLT	WER, DT
MFCC	40	4	3	5	256	40.3%	39.1%
					512	40.4%	38.4%
					1024	39.5%	38.6%
					2048	40.0%	39.1%

DNN 모델은 은닉 층이 5개, 층당 유닛 개수가 1,024개에서 가장 낮은 인식오류율을 보였고 변별적 학습으로 1~2% 가량의 향상을 보였다.

표 12. 발화속도 모델링 결과

Coding	Window Length	#State	#GMM	Dev.	Test
MFCC	25	3	1	44.0%	34.9%
MFCC	40	3	1	42.9%	34.3%
MFCC	**40**	**4**	**1**	**41.8%**	**33.3%**

발화속도 모델에서 윈도우 크기 조정에 의해 분석 셋에서 1.1%, state 개수 조정에 의해 1.1%의 향상을 보였고 테스트 셋에서도 각각 0.6%, 1%의

향상이 있음을 확인했다.

표 13. GMM 최적화 결과

Coding	Window Length	#State	#GMM	Dev.	Test
MFCC	40	4	1	41.8%	33.3%
MFCC	**40**	**4**	**3**	**40.2%**	**30.7%**

GMM 모델에서 파라미터 크기 조정을 통해 분석 셋에서 1.6%, 테스트 셋에서 2.6%의 향상을 확인했다.

표 14. SGMM 최적화 결과

Coding	Window Length	#State	#GMM	#SGMM	Boosting factor	Dev.	Test
MFCC	25	3	1	1	0.1	40.0%	30.0%
MFCC	**40**	**4**	**3**	**3**	**0.2**	**37.8%**	**28.3%**

SGMM 모델에서 파라미터 크기 조정을 통해 분석 셋에서 2.2%, 테스트 셋에서 1.7%의 향상을 확인했다.

표 15. DNN 최적화 결과

Coding	Window Length	#State	#GMM	#Layer	#Unit	Dev.	Test
MFCC	25	3	1	6	2048	42.6%	33.5%
MFCC	**40**	**4**	**3**	**5**	**1024**	**38.6%**	**28.0%**

DNN 모델에서 파라미터 크기 조정을 통해 분석 셋에서 4%, 테스트

셋에서 5.5%의 향상을 확인했다.

3.3 음향모델 학습 데이터의 구성

학습 데이터의 분량과 파라미터 크기의 변화에 따라 인식오류율을 관찰하고 이로부터 최소오류율을 보이는 파라미터 크기의 변화를 관찰했다.

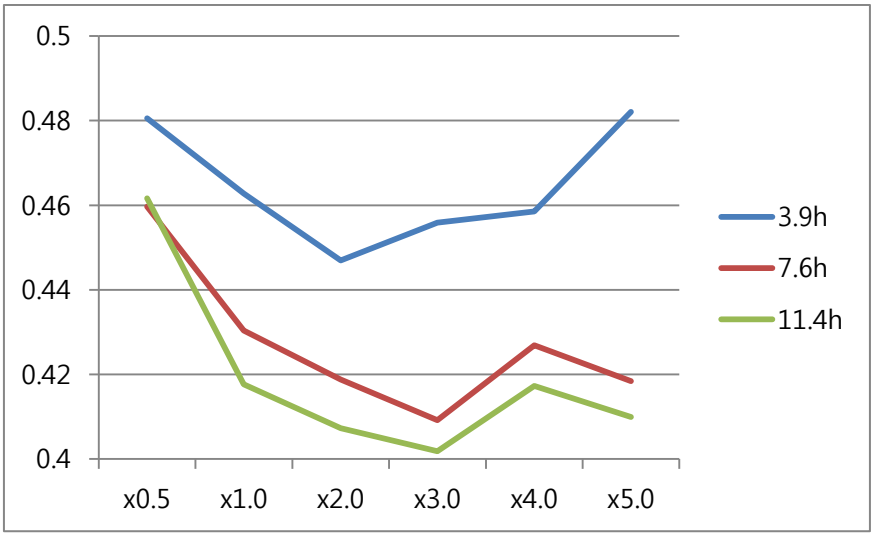


그림 11. 학습데이터 분량별 최적화 위치 변화 (분석 셋)

그림 11에서 학습데이터가 500단어 고립어 1셋, 5명 분량인 3.9h 셋일 때 GMM x2.0에 해당하는 파라미터 크기에서 최적화됨을 알 수 있다.

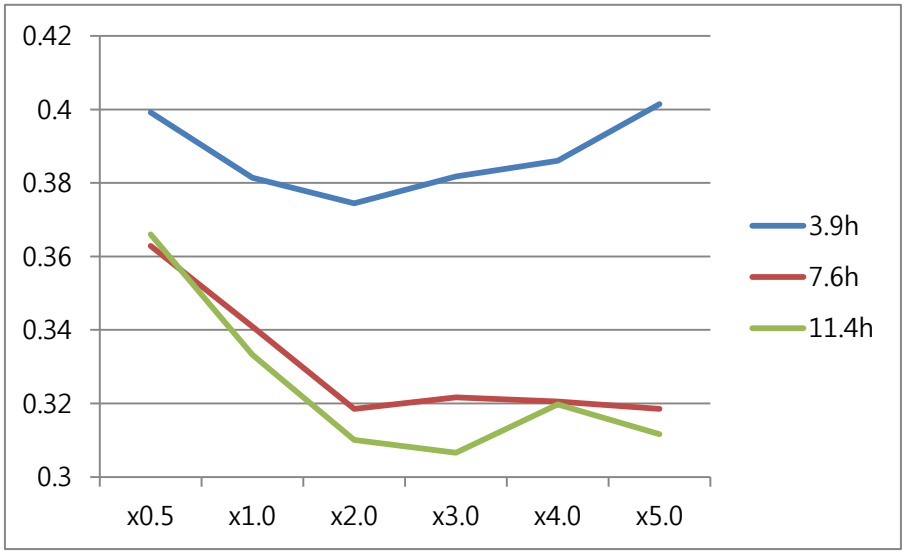


그림 12. 학습데이터 분량별 최적화 위치 변화 (테스트 셋)

표 16. 학습데이터 구성에 따른 인식오류율 비교: 장애발화와 정상발화

데이터 특성	화자 수	분량(시간)	x0.5	x1	x2	x3	x4	x5
장애	10	3.9	48.1%	46.3%	44.7%	45.6%	45.9%	48.2%
	20	7.6	46.0%	43.0%	41.9%	40.9%	42.7%	41.8%
	30	11.4	46.2%	41.8%	40.7%	40.2%	41.7%	41.0%
정상	600	120	68.2%					

표 16에서 첫째 학습데이터의 분량이 증가할수록 평균 오류율이 낮아짐을 확인할 수 있고 둘째 120시간의 정상발화로 구성한 음향모델의 인식오류율이 68.2%이고 3.9시간의 장애발화로 학습한 음향모델의

인식오류율이 44.7%로 정상발화 모델에 비해 20%이상 낮은 오류율을 보임에서 정상발화는 분량이 크더라도 장애발화 인식을 위한 학습데이터로 사용하기에 적합하지 않다는 것을 확인할 수 있다.

3.4 실험결과 분석 및 요약

표 4의 실험에서는 MFCC와 PLP 등 특징계수의 종류에 따른 인식오류율의 차이를 비교했다. 두 계수의 오류율은 통계적으로 유의한 차이를 보이지 않았으나 이는 Mengistu & Rudzicz(2011)의 실험결과와 일관성 있는 결과이다. 차이가 적지만 두 계수 중 보다 낮은 오류율을 보인 MFCC를 특징계수로서 사용했다.

다음으로 정상화자에 비교한 장애화자의 발화속도 비율에 따라 특징추출의 윈도우 크기(표 5) 및 HMM 당 state 개수(표 6)를 확장하여 실험했다. 윈도우 크기는 40ms에서 최소오류율을 보였고 HMM당 state 개수는 4개에서 오류율이 가장 낮았다. 정상화자와 장애화자의 음소당 발화시간의 비율은 1:1.6으로 윈도우 크기 조정실험에서 정상화자 기준을 25ms로 보았을 때 (Young et al., 2006, Povey et al., 2011) 이의 1.6배인 40ms에서 최적화되는 결과를 보였다. State 개수 조정실험에서는 HMM의 state개수 기준 3개의 1.6배인 5개보다 4개에서 더 낮은 오류를 보였는데 이는 state에 따라 학습되어야 할 파라미터의 개수가 증가하여 얻을 수 있는 이득이 줄어든 결과인 것으로 해석된다. Rudzicz(2013)에서는

장애화자의 공명음 조음에서 발화가 연장되는 것을 관찰하고 정상화자에 맞추어 조음시간을 단축시킴으로써 인식의 정확도를 향상했다. 이 방법은 장애발화에서 공명음 구간을 검출해야 하고 테스트 데이터를 직접 수정하기 때문에 검출오류발생의 위험이 있고 장애화자의 발화속도가 늦지 않을 때 부작용이 발생할 수 있다. 그에 비해 HMM의 구조를 변경하는 방법은 테스트 발화의 조음구간이 길 때와 짧을 때를 모두 가정하여 탐색공간을 구성하기 때문에 별도의 검출이 요구되지 않고 조음구간이 짧을 때, 즉 장애화자의 발화속도가 빠를 때 부작용이 보다 적다.

또한 출력확률모델로서 GMM, SGMM, DNN 등의 파라미터를 최적화하고 상대적인 차이를 비교했다. GMM(표 13)에서는 최적화 결과 state와 GMM의 개수를 Kaldi 툴킷 Tidigits 설정의 3배수로 지정했을 때 Tidigits 1배수에 비해 2.6% 낮은 오류율을 보였다. SGMM(표 14)에서는 GMM과 substate의 개수를 조정했고 Tidigits 설정의 3배수로 지정했을 때 베이스라인에 비해 1.7% 낮은 오류를 보였다. DNN(표 15)에서는 은닉층과 계층당 유닛 개수를 조정했고 5개 은닉층, 512 유닛에서 베이스라인에 비해 5.5% 낮은 오류율을 보였다. SGMM과 DNN은 GMM에 비해 낮은 오류율을 보였고 SGMM과 DNN 간의 차이는 크기 않았다. 그러나 화자적응 기법을 추가로 적용했을 때는 GMM이 가장 낮은 오류를 보였고 이는 학습 데이터가 부족한 음향모델의 학습에서 장애화자의 개인특성 반영이 우선적으로 강조되어야 함을 나타내는 결과이다.

120시간 분량의 정상발화를 음향모델의 학습 데이터로서 사용하여 모델링한 경우 소량인 4시간 분량의 장애발화에 비해 20% 이상 높은 오류율을 보였고 이로부터 학습 발화와 테스트 발화간의 특성차이가 크기 때문이고 정상발화를 학습 데이터로 사용하는 것은 부적합함을 확인했다.

제 4 절 결론

마비말장애 화자를 위한 음성인식기 구성에서 발생하는 사용자의 느린 발화속도, 가용한 학습 음성데이터 양의 제한, HMM 출력확률 모델간 차이 등에 초점을 두고 음향모델의 학습 알고리즘과 파라미터의 값을 조정하여 최적의 인식성능을 얻고자 했다. 발화속도를 반영한 모델에서 특징추출의 프레임 크기 조정, HMM의 state 개수 등의 조정을 통해 2.2% 가량의 오류를 낮췄다. 제한된 양의 학습 데이터를 최대한 활용하기 위해 학습 알고리즘과 파라미터를 조정한 결과 GMM에 비해 SGMM모델에서 4% 가량 낮은 오류를 보였다. 장애발화의 빈번한 조음오류에 의한 오류를 낮추기 위해 변별학습을 추가로 적용했다. 실험결과 주어진 테스트 셋에 대해 PLP, LDA, MLLT, VTLN, FMLLR을 적용하여 특징을 추출하고 SGMM으로 음향모델을 학습한 결과가 가장 높은 인식률을 보였다. 이후 이어지는 실험에서는 가장 낮은 오류율을 보이는 설정인 GMM을 HMM의 출력확률모델로 사용하고 fMLLR과 MAP로 적응하여 베이스라인 음향모델을 학습했다.

제 4 장 조음오류 특성 기반 인식단어 선택기준

빈번한 조음오류의 발생은 마비말장애 발화의 주요 특징 중 하나로 인식오류를 유발하는 요인이 될 수 있다. Byrne(1959), Platt et al.(1980), Whitehill & Ciocca(2000) 등 선행연구에서는 모음보다 자음에서의 조음오류의 발생비율이 높고 자음에서는 특히 마찰음과 파찰음, 모음에서는 모음 사각도의 외곽에 분포한 모음의 오류율이 상대적으로 높게 관찰되는 등 음소의 종류에 따른 조음오류의 차이를 분석했다. 그런데 음성인식에서 인식오류를 유발하는 잡음환경, 음소 내, 음소 간, 단어 수준의 변이로 각 수준에서 변이가 작은 경우 모델 또는 입력을 보상하여 오류를 줄이는 방법이 개발되었고 장애발화의 인식에서 특정 음소의 조음오류 빈도가 높더라도 같은 음소를 일관성 있게 다른 음소로 변형하여 발화한다면 음향모델 또는 발음모델에서 오류의 패턴을 보상하여 인식 오류를 줄일 수 있으므로 조음오류가 반드시 인식 오류로 이어지지는 않는다. 본 장에서는 인식오류의 최소화를 목표로 인식단어의 구성 기준을 파악하기 위해 단어를 음소 열로 정의하고 음소를 범주화하여 개별음소 범주의 조음에서의 인식오류 발생양상을 관찰하고 일반화 선형혼합모델(Generalized linear mixed model)로 모델링했다.

제 1 절 한국어 음소 정의와 조음 특징에 의한 범주화

한국어의 자음은 조음위치와 조음방법에 따라 표 17과 같이 구분된다. (신지영, 2011)

표 17. 한국어 자음 구분, 신지영(2011)

		조음위치				
		양순음	치조음	경구개음	연구개음	성문음
조음 방법	파열음	p, p^h, $p͈$	t, t^h, $t͈$		k, k^h, $k͈$	
	파찰음			tɕ, $tɕ^h$, $tɕ͈$		
	마찰음		s, $s͈$			h
	비음	m	n		ŋ	
	유음		l, ɾ			

한국어 단모음은 혀의 위치와 높이에 따라

표 18 과 같이 구분된다. 또한 이중모음은 활음과 단모음의 조합으로 나타낼 수 있고 활음의 종류에 따라 /y/-계 이중모음, /w/-계 이중모음, /ɯ/-계 이중모음으로 분할된다 (표 19).

표 18. 한국어 단모음 구분, 신지영(2011)

	전설	중설	후설
고	i		ɯ, u
중고	e		o
중저	ɛ		ʌ
저		a	

표 19. 한국어 이중모음 구분, 신지영(2011)

	/y/-계	/w/-계	/ɯ/-계
이중모음	je, ㅖ	ɥi, ㅟ	ɯi, ㅢ
	jɛ, ㅒ	we, ㅞ	
	jɐ, ㅑ	wɛ, ㅙ, ㅚ	
	jo, ㅛ	wɐ, ㅘ	
	ju, ㅠ	wʌ, ㅝ	
	jʌ, ㅕ		

분석을 위한 음성 데이터에 대한 실험결과에서 단어를 구성하는 음소의 개수가 인식오류율에 큰 영향을 주는 것으로 관찰되었다. 따라서 단어가 인식에 영향을 주는 특성을 코딩하기 위해서 특징이 있고 없음을 나타내는 이진 코드보다는 그 특징에 해당하는 음소의 개수를 나타내는 이산 코드를 사용했다.

인식단어의 선택기준은 대상 단어 셋이 변동되더라도 적용될 수 있어야 하므로 선택기준을 단어보다 작은 단위인 음소에 대해 정의했다. 또한

단어의 자음은 조음위치 또는 조음방법으로 범주화할 수 있고 모음은 혀의 위치 또는 혀의 높이에 따라 분류할 수 있으므로 자음 2종류 및 모음 2종류의 조합에 따른 4종류의 코드가 각각 하나의 기준적용의 단위가 될 수 있다.

제 2 절 연구 목표

마비말장애 화자의 음성인터페이스 사용에서 인식오류율이 낮은 단어의 선택기준을 파악하기 위해 고려할 사항은 다음과 같다. 먼저 단어 인식오류의 예측을 위한 일반화 선형혼합모델 구축 시 독립변수로서 자음 특성 또는 모음 특성을 단독으로 사용하여 성능을 비교한다. 다음으로 4종류의 음소 범주화로 혼합모델을 학습하고 결과를 비교한다. 마지막으로 혼합모델 구축 결과를 통해 음소 범주와 인식률 간의 관계를 확인한다.

제 3 절 분석 데이터

3.1 음성 데이터

13명의 마비말장애 화자의 고립어 발화 셋을 분석용 코퍼스로 사용한다(Choi et al., 2011, Choi et al., 2012). 단어 셋은 100개의 기기제어명령어, 36개의 통화표 단어, 37개의 음성분석용 단어 등 총 173개의 단어로 구성된다(부록 2, 173 단어 리스트).

매 화자의 173개 단어 발화 2세트 중 1세트는 음향모델의 적응을 위해

사용하고 다른 1세트는 인식을 위해 사용했다. 음성녹음 시 사용된 마이크는 SHURE SM12A 이고 음성은 16kHz, 16 bit pcm 형식으로 녹음되었다.

3.2 음소의 범주 및 통계

분석용 데이터를 구성하는 단어의 표준 발음으로부터 각 단어의 음소열을 출력하고 출력된 각각의 음소를 먼저 자음 또는 모음으로 구분하고 자음인 경우 조음방법과 조음위치에 따라 범주화했다(표 17). 또한 모음인 경우 혀의 위치와 혀의 높이에 따라 범주화했다(표 18). 2.1의 분석 데이터 173개의 단어는 총 575개의 자음과 총 458개의 모음으로 구성된다. 분류된 자음과 모음의 발견빈도를 각각 자음합계와 모음합계로 나눈 비율은 다음 표와 같다.

표 20. 173단어 리스트의 자음 범주 비율

자모	분류기준	음소범주	비율	합계
자음	조음방법	파열음	33.7%	100%
		마찰음	13.2%	
		파찰음	12.7%	
		비음	30.1%	
		유음	10.3%	
	조음위치	양순음	19.1%	100%
		치조음	43.0%	
		경구개음	12.7%	
		연구개음	22.4%	
		성문음	2.8%	

표 21. 173단어 리스트의 모음 범주 비율

자모	분류기준	음소범주	비율	합계
모음	혀의 위치	이중모음	11.4%	100%
		전설모음	33.2%	
		중설모음	18.8%	
		후설모음	36.7%	
	혀의 높이	이중모음	11.4%	100%
		고모음	31.2%	
		중고모음	15.9%	
		중저모음	22.7%	
		저모음	18.8%	

3.3 음성인식

특징추출과 음향모델은 3장의 최적화된 학습절차에 따라 생성했다. 발음사전은 1개의 단어마다 정상인 화자를 기준으로 한 1개의 기준 발음열을 포함하도록 구성했다. 하나의 발화가 하나의 단어로 구성되는 고립어 형식으로 총 173단어를 인식할 수 있는 네트워크를 정의했고 음향모델의 학습과 인식을 위해서 Kaldi toolkit(Povey et al., 2011)을 사용했다. 단어의 개수가 5개 이상인 분할의 평균 인식률은 그림 13에서와 같고 단어를 구성하는 음소의 개수가 커질수록 인식률이 높아지는 경향을 확인할 수 있다.

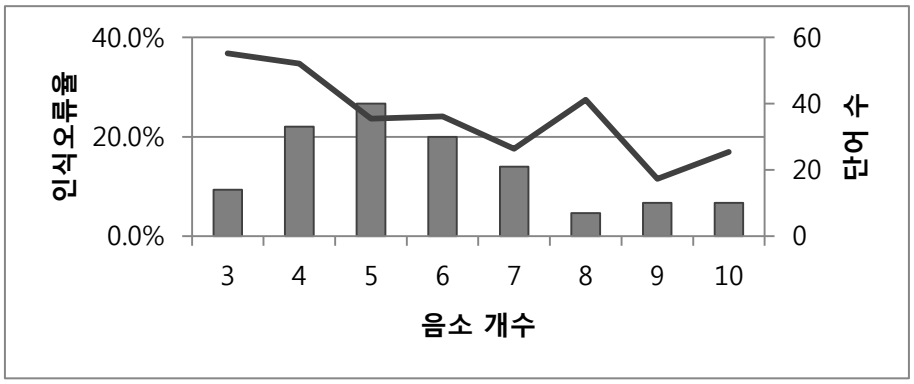


그림 13. 단어의 구성 음소개수별 인식오류율

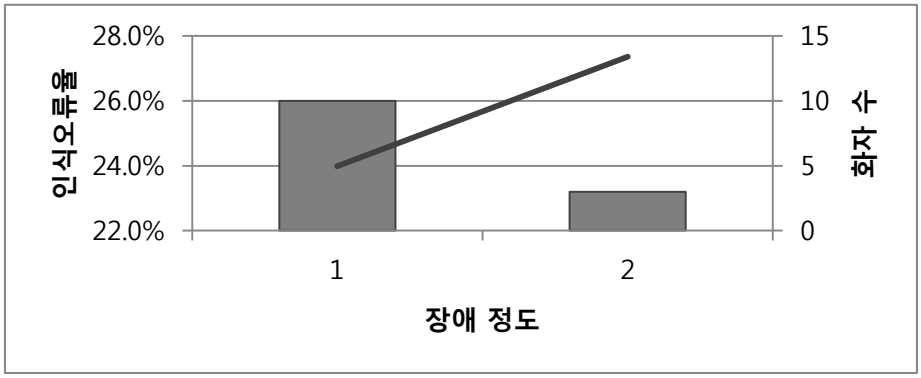


그림 14. 장애정도별 인식오류율

3.4 데이터 분석

단어의 인식률에 미치는 영향의 분석을 위해서 인식된 각 발화의 기준 발음열로부터 각 음소 범주에 해당하는 음소 개수를 출력했고 해당 요인과 발화의 인식결과와의 관계를 R의 lme4 package에 포함된 glmer function을 이용해 일반화 선형혼합모델로 모형화했다. 그림 13에서 전체음소개수가 클수록 인식률이 높아지는 결과를 보였지만 이것은 이미 독립변수로 정의된 음소 개수 또는 음소 범주의 개수의 합과 같기 때문에 공선성 제약을 지키기 위해서 고정변수로 포함하지 않았다. 그림 14에서 장애정도가 클수록 인식률이 낮아지는 관계를 보였기 때문에 화자의 장애정도 평가결과를 모델의 고정변수로 추가했다. 단어의 종류와 화자에 독립적인 단어선택기준을 찾는 것이 목표이므로 단어 종류와 화자 아이디를 모델의 임의변수로 추가했다.

표 22. 혼합모델 정의

모델구분	번호	모델 수식
자음특징 모델	1	correctness ~ (1|word_id) + (1|spk_id) + severity + labial + dental + palatal + velar + glottal
	2	correctness ~ (1|word_id) + (1|spk_id) + severity + plosive + affricate + fricative + nasal + lateral
모음특징 모델	3	correctness ~ (1|word_id) + (1|spk_id) + severity + high + midhigh + midlow + low + diphthong
	4	correctness ~ (1|word_id) + (1|spk_id) + severity + frontal + central + back + diphthong
자모특징 모델	5	correctness ~ (1|word_id) + (1|spk_id) + severity + labial + dental + palatal + velar + glottal + high + midhigh + midlow + low + diphthong
	6	correctness ~ (1|word_id) + (1|spk_id) + severity + labial + dental + palatal + velar + glottal + frontal + central + back + diphthong
	7	correctness ~ (1|word_id) + (1|spk_id) + severity + plosive + affricate + fricative + nasal + lateral + high + midhigh + midlow + low + diphthong
	8	correctness ~ (1|word_id) + (1|spk_id) + severity + plosive + affricate + fricative + nasal + lateral + frontal + central + back + diphthong

예를 들어 M0013 화자가 “포도”(AP0001)라는 단어를 발화하여 “포도”라는 단어가 인식됐을 때, 발화하고자 한 단어로부터 표준 발음열을 찾고 표준 발음열의 음소 셋으로부터 자모 특징을 검출한다.

의도한 단어 :	포도
↓	↓
기준 발음열 :	[p^h o t o]
↓	↓
자모특징변환 :	Recognized as = “포도”, correctness = 1, word_id = AP0001, spk_id = M0013, severity = 2, phone1 = 1, phone2 = 1, phone3 = 2, labial = 1, dental = 1, plosive = 3, high = 2, back = 2

제 4 절 분석 결과

1) 모델의 예측에서 자음과 모음 특징을 비교하기 위해서 자음의 조음위치 범주만으로 변수를 구성하여 혼합모델(모델 1)을 만들었고 같은 방식으로 조음방법(모델 2), 혀의 위치(모델 3), 혀의 높이(모델 4) 범주로 각각 혼합모델을 만들었다. 혼합모델은 각 모델의 Akaike information criterion을 기준으로 비교하였다. AIC 비교결과 자음특징만을 사용하여 혼합모델을 생성하는 경우(모델 1, 2)보다 모음특징만으로 생성한 혼합모델(모델 3, 4)이 더 낮은 AIC를 보였다. 또한 자음특징은 조음방법, 모음특징은 혀의 위치 범주로 구성한 혼합모델이 인식오류율을 더 잘 설명했다.

표 23. 혼합모델의 AIC 평가결과

모델구분		모델번호	AIC
자음특징 모델	조음위치	1	2307.47
	조음방법	2	2303.80
모음특징 모델	혀의 높이	3	2272.51
	혀의 위치	4	2271.65
자모범주 단위 모델	조음위치/혀의 높이	5	2276.77
	조음위치/혀의 위치	6	2275.83
	조음방법/혀의 높이	7	2269.03
	조음방법/혀의 위치	8	2268.07

2) 자음특징과 모음특징으로 혼합모델을 생성했을 때, 조음방법을 기준으로 정의할 때(모델 8, 9)가 조음위치를 기준으로 정의할 때(모델 5, 6)보다 낮은 AIC를 보였고 혀의 위치 기준(모델 6, 8)이 혀의 높이 기준(모델 5, 7)에 비해 낮은 AIC를 보였다. 4종류의 음소범주 중 자음은 조음방법, 모음은 혀의 위치로 범주화했을 때 가장 낮은 AIC를 보였다.

3) 자모특징을 모두 사용한 혼합모델 중 낮은 AIC를 보인 모델 8의 세부모델은 표 24와 같다.

표 24. 혼합모델8의 파라미터

Factor	Estimate	Std. Error	z value	Pr(>|z|)	Sign.
Intercept	3.9602	0.4725	8.3810	0.0000	***
Plosive	0.0416	0.0799	0.5200	0.6029	
Fricative	-0.2586	0.1103	-2.3450	0.0190	*
Affricate	0.0328	0.1139	0.2880	0.7734	
Nasal	0.2007	0.0821	2.4450	0.0145	*
Lateral	-0.0379	0.1376	-0.2750	0.7832	
Diphthong	0.0829	0.1316	0.6290	0.5291	
Front	0.3563	0.0949	3.7530	0.0002	***
Central	0.7971	0.1274	6.2580	0.0000	***
Back	0.4375	0.1126	3.8850	0.0001	***
Severity	-1.5684	0.2137	-7.3410	0.0000	***

Significance level: *** p<0.001, ** p<0.01, * p<0.05

고정변수 중 마찰음, 비음, 전설모음, 중설모음, 후설모음, 장애정도 등이 통계적으로 유의했다. 장애정도, 마찰음, 유음 등의 개수는 증가할수록 인식오류가 증가했다. 비음, 폐쇄음, 파찰음, 이중모음 등은 증가할수록 인식오류가 감소했지만 비음을 제외한 변수의 인식오류에 대한 영향은 0에 가까웠다.

제 5 절 결론

모델 9에서 자음범주 중 인식오류와 유의수준 0.05 이하에서 상관관계를 보인 자음 범주는 마찰음과 비음이고 모음 범주는 단모음 전체이다. 단어에서 마찰음, 유음 등의 개수가 증가하면 그에 따라 인식오류가 증가하는 경향을 보였고 비음과 모음은 추가할수록 인식의 정확도가 증가하는 경향을 보였다. 모음 중 이중모음의 영향력이 가장 작았고 중설모음이 정확도를 가장 크게 증가시켰고 전설모음과 후설모음 순으로 영향력의 크기가 감소했다. 파찰음과 폐쇄음은 추가할수록 인식의 정확도가 증가하기는 하지만 증가의 크기가 0에 가깝다.

개별범주의 영향력과 조음오류에 대한 선행연구(Byrne,1959, Platt et al.,1980, Whitehill & Ciocca, 2000)의 비교에서 마찰음, 유음, 이중모음, 전설/중설/후설모음, 비음 등은 일치하는 결과로 해석할 수 있다.

파열음 발화 시 조음의 복잡함이 조음오류로 이어져 음향모델의 적응 이후에도 인식오류의 원인으로 작용하였을 것으로 추정할 수 있고 그에 따라 단어의 선택에서 파열음이 포함되지 않도록 기준을 세워야 한다. 마찬가지로 비음은 단어의 선택 시 포함되도록 기준을 세워야 한다. 또한 모델 8의 모든 모음범주는 유의수준 0.001 이하에서 양의 상관관계를 보였다. 중설모음의 계수가 가장 크고 후설모음, 전설모음 순서로 작아졌고 따라서 단모음 중에서도 중설모음이 단어 선택 시 가장 우선시 되어야 한다.

제 5 장 단어 간 유사도 최소화 기준 인식단어 최적화

고립단어 인식에서 입력발화는 인식 시 사용자가 어떤 단어를 발화했는지 찾기 위한 단서가 된다. 인식 시 “칠번” 또는 “일번”과 같이 발음열이 유사한 단어의 쌍이 단어 리스트에 있음을 가정할 수 있는데 이런 경우 발화 칠번의 입력구간 중 초성 ㅊ 이후의 구간은 단어 칠번과 일번의 변별에 도움을 줄 수 없으므로 차이가 나는 음소 1개 구간의 계산결과에 따라 단어의 인식결과가 결정된다. 마찬가지로 단어간 발음열의 유사도가 높을수록 단어의 변별에서 단서가 줄어들고 그에 따라 인식오류율이 높아질 수 있다. 특히 변별의 단서를 줄이는 문제는 마비말장애 발화가 정상발화에 비해 빈번한 조음오류를 포함하기 때문에 정상발화의 인식에서보다 더 많은 인식오류를 유발할 수 있다. 발음변이 모델링에 대한 연구에서 발음변이의 추가 시 추가된 발음변이와 유사한 발음을 가진 단어의 인식 정확도를 하락시키는 부작용이 있음을 보고했다.

Tsai, Chou & Lee(2007)에서는 단어 w_i 중에서 발음변이 v_j 가 발견된 확률, $\mathrm{pf} = \mathrm{P}(v_j|w_i)$ 과 발음변이 v_j 를 포함하는 단어의 역비율인 $\mathrm{iwf} = \log\frac{|\#words|}{|\#w_j|}$의 곱으로 $\mathrm{pf} - \mathrm{iwf\ score}$를 계산하고 발음변이의 선택에 적용했다. 단어 w_i에서 발음변이 v_j의 실현빈도가 낮을수록 pf가 낮아지고 v_j가 여러 단어에서 중복되어 포함될수록 iwf 가 감소하여 낮은 $\mathrm{pf} - \mathrm{iwf\ score}$ 를

가지게 되고 그에 따라 발음사전에 포함되기 어려워진다. 이 방법은 발음변이 선택 시 변이추가에 의해 다른 단어에서 발생하는 혼동성의 증가를 반영하였다는 장점이 있지만 발음변이 실현확률, pf 의 계산 시 학습 발화의 강제인식결과를 바탕으로 확률을 계산한다. 즉 학습 데이터가 충분히 많은 단어를 포함하고 테스트 데이터의 미등록 단어의 비율이 낮은 상황을 가정하고 있다. 그러나 본 논문에서는 마비말장애 화자의 음성 인터페이스 사용으로 데이터의 단어 종류와 양에 제한이 있으므로 그러한 가정을 따를 수 없다.

Fung & Liu(2005)에서는 같은 단어의 다른 발음변이를 추가함에 따라 발생하는 혼동을 음향적 혼동과 음운적 혼동 등으로 경우를 나누고 각각 음향모델 공유와 발음변이 추가에 반영하여 인식률을 향상하였다. 그러나 같은 단어 내 두 가설의 음소 대 음소 혼동만을 가정하였기 때문에 다른 단어의 유사한 음소 열에 의한 혼동 발생에 대해서는 대응할 수 없다. 따라서 혼동성 예측에 같은 단어와 다른 단어의 발음열에 대한 음소열 대 음소열 혼동의 반영이 필요하다.

본 장에서는 Levenshtein 거리와 Cosine 거리를 기준으로 단어의 음소 열간 거리와 단어간 유사도를 정의하고 최소/최대 유사도 기준으로 인식단어 리스트를 구성하여 인식률을 비교한다. 또한 N-best 가설 예측으로 정의된 단어간 유사도를 통해 인식 시 혼동을 유발하는 단어를 추정할 수 있음을 확인한다.

제 1 절 열 거리 기반의 단어 간 유사도

인식의 탐색공간에서 단어는 음소의 열로 표현되므로 단어의 거리는 음소열간의 거리로 측정할 수 있다. 측정방식은 Levenshtein, Damerau-Levenshtein distance, Needleman-Wunsch algorithm, Smith-Waterman algorithm, Jaro/Jaro-Winkler distance 등 동적 프로그래밍 기반 측정방식과 q-gram, cosine, jaccard, dice, simple matching 등과 같은 특징 기반 측정방식으로 구분할 수 있다(Van der Loo, 2014).

동적 프로그래밍 기반 측정방식에서는 대치, 삽입, 삭제, 교환 등 수정함수를 정의하고 한 단어를 수정하여 다른 단어와 같게 만들기 위해 소모되는 최소비용을 두 단어 간 거리로 정의한다. 단어간 거리를 정의하기 위해 먼저 각 단어를 발음사전에 정의된 음소 열로 보았고 두 음소 열을 동적 프로그래밍으로 정렬하여 최소편집거리를 계산했다. 동적 프로그래밍의 수정함수의 페널티는 삽입 1, 삭제 1 등으로 정의했고 대치는 대응될 음소 1개가 삭제되고 음소 1개가 새로 삽입되는 것으로 보아 페널티를 2로 정의했다. 비교될 단어에 다중 발음변이가 존재하는 경우 정의된 두 단어의 모든 발음변이 쌍마다 최소편집거리를 계산하고 거리의 최소값을 대푯값으로 가정했다.

표 25. 동적 프로그래밍 기반 단어간 거리 계산 예제

	단어	발음열	음소표기	수정내용	최소편집거리	
단어 1	삼번	삼번	ㅅ ㅏ ㅁ ㅂ ㅓ ㄴ			
단어 2	사번	사번	ㅅ ㅏ - ㅂ ㅓ ㄴ	삭제 1	1	1
단어 1	이번	이번	- ㅣ - ㅂ ㅓ ㄴ			
단어 2	이전	이전	- ㅣ - ㅈ ㅓ ㄴ	대치 1	2	2
단어 1	모자	모자	ㅁ ㅗ - ㅈ ㅏ			
단어 2	문자	문짜	ㅁ ㅜ ㄴ ㅉ ㅏ	대치 2 삽입 1	5	5
단어 1	우산	우산	- ㅜ - ㅅ ㅏ ㄴ			
단어 2	우편	우편	- ㅜ - ㅍ ㅕ ㄴ	대치 2	4	4
		웁편	- ㅜ ㅂ ㅍ ㅕ ㄴ	대치 2 삽입 1	5	

전체단어내의 각 단어 쌍마다 최소편집거리의 최대값을 계산하고 거리값을 0~1사이로 정규화하기 위해 각 단어간 거리를 거리의 최대값으로 나누었다.

$$\text{Distance}(w_1, w_2) = \frac{Distance(w_1,w_2)}{Maximum\ Distance}$$

$$\text{Similarity}(w_1, w_2) = 1 - \frac{Distance\ (w_1,w_2)}{Maximum\ Distance}$$

특징 셋 기반 측정방식에서는 단어를 대표할 수 있는 특징 셋을 정의하고 두 단어의 특징 셋 간의 거리를 단어 간 거리로 정의한다. 단어를 특징으로 표현하기 위해서 발음열의 개별음소를 하나의 Q-gram으로 가정했고 두 단어의 음소열로부터 계산된 Q-gram vector f_1 , f_2 간의 cosine 거리를 계산했다.

제 2 절 인식률에 대한 단어 간 유사도의 영향

열간 거리로부터 계산된 단어간 유사도가 인식단어 구성 시 인식오류율에 미치는 영향을 검토하기 위해서 3장의 분석 셋에 대한 인식단어 구성실험을 수행했다. 인식오류에 대한 영향에서 음소 개수의 영향을 배제하기 위해 단어 선택 시 전체단어 셋을 음소의 개수 별로 4개, 5개, 6개, 7개 등으로 분할했고 분할된 각 셋 별로 단어구성 실험을 수행했다. 각 셋에 포함된 모든 단어에 대해서 해당 단어 w_a과 유사도가 가장 큰 9개의 단어 셋(Max)과 단어 w_a 과 유사도가 가장 낮은 9개의 단어 셋(Min)을 구성했고 인식단어 셋이 Max일 때 단어 w_a의 인식률과 Min일 때 단어 w_a의 인식오류율 간의 차이를 비교했다.

표 26. 최대유사도 단어 및 최소유사도 단어간 인식오류율 차이

		최대유사도	최소유사도	차이
Levenshtein Distance	음소개수, 4	14.5	5.6	8.9
	5	7.9	2.1	5.8
	6	10.1	1.9	8.2
	7	6.2	4.0	2.2
Cosine Distance	4	14.7	5.6	9.1
	5	6.9	2.3	4.7
	6	9.3	2.1	7.2
	7	6.6	3.3	3.3

비교결과 음소개수가 4일 때의 평균 인식률 차이는 Levenshtein 거리로 유사도를 정의했을 때 8.9%이고 cosine 거리로 정의했을 때 9.1%를 보였다. 모든 경우에서 유사도가 낮은 단어 셋에서 보다 낮은 인식오류율을 보이므로 단어간 유사도가 인식률에 영향을 주는 요인임을 확인할 수 있다.

제 3 절 N-best 추정

화자가 단어 w_1을 의도하고 발화한 음성입력 $\vec{x}$에 대한 단어 w_1의 조건부 확률은 $p(\vec{x}|w_1)$ 이고 이 값의 계산과정에서 인식단어 리스트 내의 다른 단어의 정보는 사용되지 않는다. 그렇지만 단어 리스트 내에 $p(\vec{x}|w_1)$ 보다 높은 확률, $p(\vec{x}|w_i)$를 가지는 단어 w_i가 존재하는 경우 오인식이 발생한다. 즉 단어 리스트의 구성에 따라 인식결과가 달라 질 수 있다.

특정 단어를 발화한 음성 셋에 대해 각각 인식을 수행하고 N-best 인식결과를 관찰할 때 정답과 다른 단어가 높은 빈도로 N-best 인식결과에 포함된다면 다음 인식 시 해당 오답단어가 다시 오인식을 유발할 것으로 예측할 수 있으므로 단어 리스트 구성 시 그 오답단어를 제외함으로써 오류를 줄일 수 있다. 다음은 단어 “켜기”의 인식 시 단어 리스트의 구성에 따라 인식률이 변함을 보이는 인식예제이다.

표 27. 혼동가설에 의한 인식결과 차이 비교

화자	173 단어 인식결과			173 – n 단어 인식결과		
	가설 1	가설 2	가설 3	가설 1	가설 2	가설 3
F0013	끄기	붙이기	전화걸기	**켜기**	이메일쓰기	문자쓰기
F0024	끄기	거북이	**켜기**	**켜기**	크게	바가지
F0025	선택	우산	우편	단어선택	곡선택	이전
F0027	거북이	정지	화장실	한글	카메라	호랑이
F0036	칠	**켜기**	키다리	**켜기**	느리게	이메일
M0019	거절	거북이	**켜기**	**켜기**	토끼	크게
M0020	기러기	**켜기**	전화걸기	**켜기**	토끼	문자보내기
M0022	딸기	다음칸	**켜기**	**켜기**	도라지	단추
M0023	확대	거북이	다음칸	문자보기	올라가요	켜기
M0028	**켜기**	기러기	전화걸기	**켜기**	토끼	한칸띄우기
M0029	바퀴	아버지	전화받기	토끼	켜기	바가지
M0031	거북이	전화받기	아버지	한칸띄우기	바가지	켜기
M0032	거북이	**켜기**	아버지	**켜기**	토끼	문자보기

오류유발 단어 셋 = {거북이, 거절, 기러기, 끄기, 다음칸, 딸기, 바퀴, 붙이기, 선택, 아버지, 우산, 우편, 전화걸기, 전화받기, 정지, 칠, 켜기, 키다리, 화장실, 확대}

인식단어 리스트가 전체 단어 (173 단어)일 때 "켜기"의 인식오류율은 12/13*100=92.3(%) 이고 상위 3개의 가설 중 정답이 포함될 때 정인식으로 판단할 때의 오류율은 6/13*100=46.2(%)이다. F0024의 켜기 발화의 인식에서 끄기, 거북이 등의 단어가 “켜기”에 비해 높은 점수, $p(\vec{x}|w_i)$를 얻었고 F0036에서는 “칠”이 “켜기”에 비해 높은 점수를 얻었다.

이처럼 3-best 인식결과에 포함된 오답 단어 20개를 오류유발 단어로 가정하고 인식단어에서 제외한 후 2번째 인식을 수행했을 때 F0024 화자의 인식결과에서 "켜기" 단어의 조건부 확률은 첫번째 같은 단어를 인식했을 때와 동일하지만 "켜기"보다 높은 조건부 확률을 보였던 "끄기", "거북이" 등의 단어가 두번째 인식단어에서 제외되었기 때문에 "켜기"가 가장 높은 점수를 보이는 단어로 출력되었다.

오류유발 단어 셋에 포함된 20개의 단어를 인식단어 리스트에서 제외했을 때 "켜기"의 인식오류율은 각각 5/13*100=38.5(%), 2/13*100=15.4(%)로 향상됨을 확인할 수 있다. 분석 셋에 포함된 173 단어로 인식단어 리스트를 구성하고 13명의 각 화자마다 173 단어를 테스트한 결과 1-best 인식오류율은 24.1%, 3-best 인식오류율은 10.4% 이었고 이 첫번째 인식결과를 바탕으로 각 단어마다 오류유발 단어 셋을 정의하고 인식단어에서 제외한 후 다시 인식했을 때 1-best 인식오류율은 7.8%, 3-best 인식오류율은 5.1%로 향상되었고 이로부터 인식결과로서 발견되는 단어를 리스트에서 제외하는 것이 인식결과 향상으로 이어짐을 확인할 수 있다. 그러나 실제 인식 시 예와 같이 오인식 여부를 수동으로 판단하여 시스템에 전달할 수는 없기 때문에 n-best 인식결과에서 나타날 단어의 사전 예측이 필요하다.

정답과 발음이 유사한 단어는 탐색공간 상에 정답과 유사한 HMM 열로 가설이 구성된다. 그에 따라 인식과정에서 조건부 확률 총합의 계산 시

정답과 거의 같은 점수를 출력하게 되고 그 중 정답에 비해 높은 점수를 얻는 경우 오인식을 유발하게 된다. 발화자가 정답을 의도하여 발화하는 경우라면 정답과 유사한 단어가 N-best 인식결과로 출력될 것이므로, 특정 열간 거리로 계산한 유사도로 n-best 가설을 예측하고, 실제 n-best 인식결과와 비교하여 예측의 정확도를 확인하는 것이 유사도를 평가하는 방법이 된다.

3장의 분석 셋에 포함된 모든 단어 쌍의 유사도를 계산하고 각 단어마다 계산된 유사도의 산술평균을 해당 단어의 대표 유사도로 계산했다. 각 단어는 13명이 한번씩 발화했기 때문에 총 13개의 발화로 구성되고 각 발화마다 인식결과로 최대 3-best 가설을 출력했다. 산술평균의 역치를 설정하고, 해당 역치보다 높은 유사도를 가지는 단어는 3-best 가설에 3번 이상 포함되는 것으로 예측했을 때, 예측의 recall과 precision, F-measure를 다음과 같이 계산했다.

$$\text{Recall} = \frac{\#\,Words\ Predicted\ as\ confusing\ \&\ Observed\ in\ N-best\ list}{\#\,Words\ Observed\ in\ N-best\ list}$$

$$\text{Precision} = \frac{\#\,Words\ Predicted\ as\ confusing\ \&\ Observed\ in\ N-best\ list}{\#\,Words\ Predicted\ as\ confusing}$$

$$\text{F}-\text{measure} = \frac{2 * Precision * Recall}{Precision + Recall}$$

다음 그래프는 산술평균의 역치(x축)를 변화시켰을 때의 F-measure (y축)의 변화를 나타낸다.

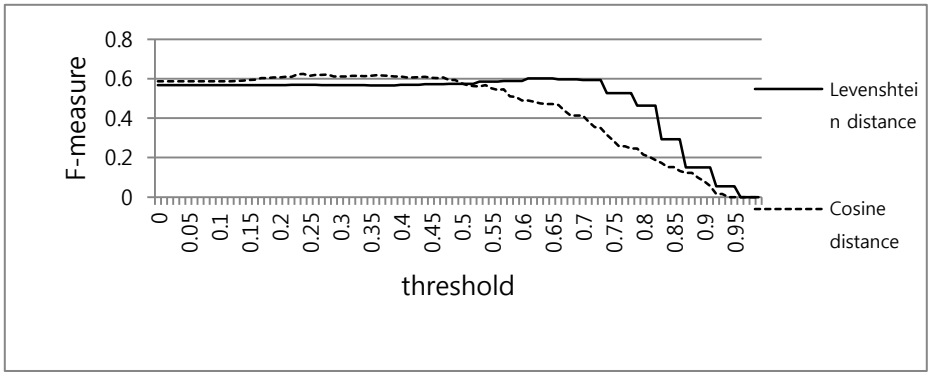


그림 15. N-best 추정 결과

Levenshtein similarity는 역치가 0.61~0.65사이에서 가장 높은 f1 score, 0.602를 보이고 Cosine similarity는 역치가 0.24에서 가장 높은 f1 score, 0.624를 보인다. 이 결과는 table 11에서와 같이 두 기준을 적용했을 때 성능의 차이가 크게 나타나지 않는 것으로 해석할 수 있다.

제 6 장 인식실험

제 1 절 단어 리스트 구성

본 논문에서 마비말장애 화자가 사용하는 음성 키보드에는 자음과 모음 등 총 26개의 자소가 키로서 배치되어 있다. 사용자가 키보드 상의 개별 키를 활성화하기 위해서 해당 키의 자소로 시작하는 단어를 발화해야 한다. 예를 들어 사용자가 자소 'ㄱ'이 표시된 키를 누르기 위해서 단어 '기러기'를 발화해야 하고 자소 'ㅏ'의 키를 누르기 위해서 단어 '아버지'를 발화해야 한다. 또한 '띄어쓰기/줄바꿈/지우기/취소' 등과 같은 기능제어를 위한 21개의 단어 리스트에 추가했다. 따라서 단어 셋의 선택에 대한 테스트에서 전체 단어는 크게 자소 후보 단어 셋과 기능제어 후보 단어 셋으로 구분된다.

테스트를 위한 전체 단어 중 기능제어 후보 단어 셋은 각 기능과 의미적으로 연관된 5개의 단어 셋으로 구성되어 있다. 각 기능마다 5개의 단어를 후보 셋으로서 할당하였고 21개의 기능을 정의하였으므로 기능제어 후보 단어 셋은 총 105개의 단어로 구성된다. 전체 단어 중 기능제어 후보 단어 셋에 포함되지 않는 단어는 그 단어의 첫번째 자소의 종류에 따라 분류된다. 자소의 종류에 따라 분할된 단어 셋을 해당 자소의 후보 단어 셋으로 할당하였고 자소 후보 단어 셋은 총 354개의 단어로 구성된다. 인식단어 선택과제는 자소 후보 단어 셋과 기능제어 후보 단어

셋 등의 각 후보 셋으로부터 한 단어씩을 선택하여 총 47 단어를 인식단어 리스트로 구성하는 과제이다.

표 28. 베이스라인 통화표 단어 리스트

자모구분	음소	단어	자모구분	음소	단어
자음	ㄱ	기러기	자음	ㅎ	한강
	ㄴ	나폴리	모음	ㅏ	아버지
	ㄷ	도라지		ㅑ	야자수
	ㄹ	로마		ㅓ	어머니
	ㅁ	미나리		ㅕ	연못
	ㅂ	바가지		ㅗ	오징어
	ㅅ	서울		ㅛ	요지경
	ㅇ	잉어		ㅜ	우편
	ㅈ	지게		ㅠ	유달산
	ㅊ	치마		ㅡ	은방울
	ㅋ	키다리		ㅣ	이순신
	ㅌ	통신		ㅐ	앵무새
	ㅍ	파고다		ㅔ	엑스레이

표 29. 베이스라인 기기제어명령 단어 리스트

기능	인식단어	기능	인식단어
전화	전화	위로	아이
메시지	메시지	아래로	두더지
일정	일정	왼쪽	좌측
날씨	날씨	오른쪽	우측
뉴스	뉴스	확인	확인
위치알림	위치알림	취소	다시
네이버	네이버	이전	요전
다음	한메일	다음	이후
구글	구글	보내기	보내기
탐색기	탐색기	종료	끝내기
계산기	계산기		

1.1 베이스라인 인식 단어 리스트

통화표 단어들은 조음의 용이성을 기준으로 선정되었고 500단어 셋에는 3세트의 통화표 단어 리스트가 포함되어 있다. 각 통화표 리스트의 단어를 자소의 대표 단어로 선택했고 기능제어 후보 단어 셋 중 1-best 단어를 해당 기능의 대표 단어로 선택하여 총 3종류의 베이스라인 단어 리스트를 만들었다. 3장의 분석 셋에 대한 인식결과로부터 단어를 구성하는 음소의 개수와 인식률 간의 연관성을 확인하였으므로 제안하는 모델이 단어의 음소 개수에 의한 영향보다 정확한 예측을 할 수 있음을 보이기 위해서 음소개수의 최대화 및 최소화를 기준으로 각 자소와 기능을 대표하는 단어의 선택하여 베이스라인 단어 리스트를 만들었다. 한 자소 후보 단어

셋 중에서 음소개수가 가장 큰 단어가 둘 이상 존재하는 경우 그 중 한 단어를 임의로 선택하여 단어 리스트를 구성했고 인식을 총 100회 반복하여 평균 인식률을 계산했다. 또한 임의 선택으로 베이스라인 인식단어 리스트를 만들었고 총 100회 반복하여 평균 인식률을 계산했다.

1.2 조음점수 최대화 단어 리스트

4장의 혼합모델 중 가장 낮은 AIC를 보인 조음방법/혀의 위치 범주 혼합모델의 가중치와 단어의 발음열에서 계산된 음소범주에 속하는 음소개수의 가중합으로 개별단어의 조음점수를 계산한다.

$$\text{Articulatory Score}(\text{word}_i) = \sum_{\text{phone class j}} \text{weight}_j * (\text{number of phones}_j)$$

조음점수 최대화 단어리스트 구성방식에서는 47개의 자소와 기능 후보 단어 셋마다 조음점수가 가장 높은 단어를 인식 단어로 선택했다.

1.3 단어 간 유사도 최소화 단어 리스트

단어간 유사도가 높은 단어를 교체함으로써 인식률을 높일 수 있음을 보이기 위해서 조음점수 최대화를 기준으로 만들어진 단어 리스트 내의 단어 쌍의 유사도를 계산했고 기준점보다 높은 유사도를 가지는 단어를 후보 단어 셋에 포함된 다른 단어 중 조음점수와 유사도간 차이가 가장 높은 단어로 교체함으로써 단어 리스트를 구성하고 인식오류율을 관찰했다.

```
For word i in vocabulary,
  Compute maximum similarity of word i
  If maximum similarity > threshold,
    For word j in grapheme bin of word i,
      Find maximum of (articulatory score - similarity)
    Replace word i by the word with the maximum value
```

그림 16. 단어선택 알고리즘

인식단어의 혼동유무를 판단하는 기준이 되는 기준점은 0~1 사이의 값을 가진다. 기준점을 0에서부터 0.02 간격으로 점진적으로 증가시켜 인식 단어 리스트를 만들었다. 0.00~0.20 구간에서 동일한 단어 리스트가 출력됐고 0.40~1.00 구간에서 동일한 리스트를 출력됐기 때문에 0.44~0.78 구간의 인식률만을 표시했다.

제 2 절 기초 실험

인식률에 대한 조음점수와 유사도 요인의 영향을 실험으로 확인하기에 앞서 인식단어 리스트의 크기, 적응 및 테스트 단어 개수 등 실험환경을 설정하기 위해서 분석용 음성 코퍼스에 대해 기초실험을 수행했다.

표 30. 기초실험 인식결과

인식단어 크기	적응 발화수	장애정도				
		1	2	3	4	2/3
173	173	5.1	24.0	27.4	64.7	24.8
47, 임의선택*100 회		2.9	15.2	15.5	51.4	15.3
차이	-	2.2	8.8	11.9	13.3	9.5
47, 임의선택*100 회	47	5.2	21.7	22	61.9	21.8
	47*2	4	16.1	15	52.5	15.8
	173	2.9	15.2	15.5	51.4	15.3
47, 최소음소, 100 회	47*2	5.3	16.7	11.2	46.6	15.4
47, 최대음소, 100 회	47*2	3.6	13.3	22.1	53.7	15.2

표 30의 번호 1과 2의 결과로부터 인식단어 리스트의 크기가 47에서 173으로 늘어났을 때 장애정도에 따라 인식오류율이 2.16% ~ 13.29% 증가함을 알 수 있다. 또한 장애정도가 클수록 단어 리스트의 크기증가에 의한 오류율 증가의 폭이 커짐을 관찰했다.

경도(장애정도 1) 그룹의 173 단어 인식오류율은 5.1% 이고 47 단어 랜덤 셋 오류율은 2.9% 이다. 이 그룹에서 단어의 크기가 상대적으로 큰 173 단어 인식오류율이 약 5%로 이미 높은 정확도를 보이고 있다. 따라서 인식단어의 개수를 줄임으로써 얻을 수 있는 향상이 클 것으로 기대하기 어렵다. 그에 반해 경도-중등도, 중등도-중도(장애정도 2, 3) 그룹에서 173 단어의 인식오류율이 24.8%이고 인식단어 크기를 47로 줄였을 때 오류율이 15.3%로 상대적으로 크게 감소했다. 따라서 본 논문에서는 장애정도 2, 3의 화자에 대해 단어리스트 구성실험을 수행했다.

실험에서 음향모델 적응 데이터의 규모를 결정하기 위해서 인식단어 47개 단어를 1회씩 발화한 음성, 2회씩 발화한 음성, 173 단어를 1회씩 발화한 음성 등을 적응 데이터로 사용하여 인식실험을 수행했다. 47 단어의 2회 발화 적응 모델은 1회 발화 적응 모델에 비해 1.23~9.42% 향상을 보였고 173 단어 적응 모델과 비교할 때 오류율의 차이는 1% 내외이다. 173 단어의 발화와 47 단어의 2회 발화를 입력하기 위해 소요되는 노력의 차이는 적지만 테스트 셋인 500 단어 발화와 47 단어의 2회 발화의 차이는 크기 때문에 47 단어의 2회 발화를 적응 데이터로 사용했다.

그림 13에서 단어를 구성하는 음소의 개수가 늘어날수록 오류율이 낮아지는 경향이 관찰된다. 특히 4 또는 5음절로 구성된 단어의 경우 평균적으로 낮은 오류율을 보였지만 사용의 편의성을 고려하여 테스트 셋의 500단어 중 4, 5음절로 구성된 118 단어를 선택대상에서 제외하고 나머지 382개의 2, 3음절 단어 중에서 인식단어를 선택했다.

음소개수가 인식률에 주는 영향은 장애정도에 따라 다르게 나타났다. 장애가 비교적 적은 장애정도 1, 2 그룹의 화자들은 음소개수를 최대화 하는 경우의 오류가 최소화 하는 경우의 오류율 보다 낮았다. 그에 반해 장애가 보다 심한 장애정도 3, 4 화자들은 음소개수를 최대화 하는 경우의 오류율이 최소화 하는 경우보다 높음을 확인할 수 있다. 따라서 장애가 큰 화자에 대해서는 단어를 구성하는 음소개수에 대한 제한이 필요함을 알 수 있다.

제 3 절 실험 환경

3.1 음성 코퍼스 구성

실험을 위해서 13명의 장애화자가 500 단어를 3번씩 발화한 코퍼스를 사용했다. 테스트 셋의 단어는 105개의 기기제어명령어, 108 통화표 단어, 287 PBW 단어 등으로 구성된다. (부록 3)

음향모델은 3장의 절차에 따라 학습했다. 학습 셋과 테스트 셋에 포함되는 화자가 서로 겹치지 않도록 2분할하여 화자독립 음향모델을 만들었고

개별화자의 음성으로 적응했다. 화자적응을 위해 각 장애 화자가 인식단어를 발화한 음성 2세트가 데이터로 사용됐다. 실험에서 인식단어는 47개의 단어로 구성되므로 화자적응 발화의 개수는 94개이다.

음성의 특징추출과 음향모델, 발음사전 등 구성은 3장의 실험에서와 같다. 500 단어의 표준 발음을 먼저 자음과 모음으로 구분하고 자음인 경우 조음방법과 조음위치에 따라 범주화했고 모음인 경우 혀의 위치와 혀의 높이에 따라 구분했다. 500개의 단어는 총 1,808개의 자음과 총 1,472개의 모음으로 구성된다. 분류된 자음과 모음의 발견빈도를 각각 자음합계와 모음합계로 나눈 비율은 다음 표와 같다.

표 31. 500단어 리스트의 자음 범주 비율

자모	분류기준	음소범주	비율	합계
자음	조음방법	파열음	39.7%	100%
		마찰음	13.3%	
		파찰음	10.3%	
		비음	25.1%	
		유음	11.7%	
	조음위치	양순음	15.9%	100%
		치조음	43.6%	
		경구개음	10.3%	
		연구개음	25.2%	
		성문음	5.0%	

표 32. 500단어 리스트의 모음 범주 비율

자모	분류기준	음소범주	비율	합계
모음	혀의 위치	이중모음	20.3%	100%
		전설모음	28.0%	
		중설모음	18.2%	
		후설모음	33.5%	
	혀의 높이	이중모음	20.3%	100%
		고모음	32.6%	
		중고모음	13.9%	
		중저모음	14.9%	
		저모음	18.2%	

제 4 절 인식 결과

4.1 베이스라인 모델

테스트 단어 중 중 임의로 선택한 47 단어로 인식단어를 구성할 때 100회 반복시행에서 오류율은 평균적으로 17.2%였다. 통화표 셋으로 인식단어를 구성할 때 오류율은 16.5~18.2%, 음소길이 최대화 모델의 평균 오류율은 16.8%이다.

4.2 조음점수 최대화 모델

표 33. 조음점수 최대화 모델 실험결과, 적응발화 크기 47*2

구분			장애정도			향상
			2	3	2/3	
베이스라인	통화표 방식	셋 1	19.0	15.6	18.2	-
		셋 2	16.6	16.3	16.5	
		셋 3	16.4	22.7	17.8	
	음소길이 최소		21.4	19.8	21.1	
	음소길이 최대		15.1	22.3	16.8	
	랜덤 통화표		16.8	18.5	17.2	
조음점수 최대화 모델			13.7	14.2	13.8	2.7

조음점수 최대화 모델의 인식오류율은 13.8%로 베이스라인에 비해 최소 2.7%의 향상을 보였고 베이스라인 오류율과 향상된 모델의 오류율을 분산분석으로 각각 비교할 때 인식률의 향상이 통계적으로 유의함을

확인했다.

표 34. 조음점수 최대화 모델 실험결과, 적응발화 크기 500

구분			장애정도			향상
			2	3	2/3	
베이스라인	통화표 방식	셋 1	14.29	15.6	14.59	-
		셋 2	16.6	15.6	16.37	
		셋 3	17.23	22.7	18.49	
	음소길이 최소		19.81	18.92	19.6	
	음소길이 최대		13.94	17.24	14.7	
	랜덤 통화표		14.59	17.13	15.17	
조음점수 최대화 모델			10.9	9.22	10.51	4.08

표 34의 음향모델의 적응 데이터의 분량을 500 발화로 확장한 실험에서도 조음점수 최대화 모델은 베이스라인에 비해 유의한 향상을 보였다.

4.3 단어 간 유사도 최소화 모델

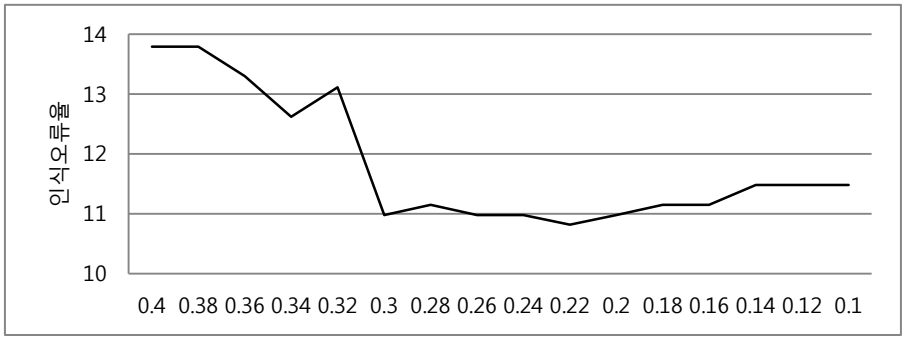


그림 17. 단어 간 코사인 유사도 최소화 모델 실험결과, 인식단어 크기 173

그림 17은 조음점수 최대화 단어 리스트를 유사도 기준으로 최적화했을 때의 인식오류율이다. 그래프의 x축은 단어 교체를 판단하는 역치값이고 y축은 인식오류율을 나타낸다. 역치값이 작아질수록 단어 리스트에서 더 많은 단어들이 교체된다. Cosine similarity를 기준으로 최적화했을 때 최소 인식오류율은 10.8%였다. 베이스라인 인식률에 비해 5.7%, 조음점수 최대화 단어리스트에 비해 3% 향상된 결과이다. 인식률의 최대값은 역치 0.22에서 관찰됐고 N-best 추정에서의 최대값이 역치 0.24에서 관찰된 것과 일관성 있는 결과이다.

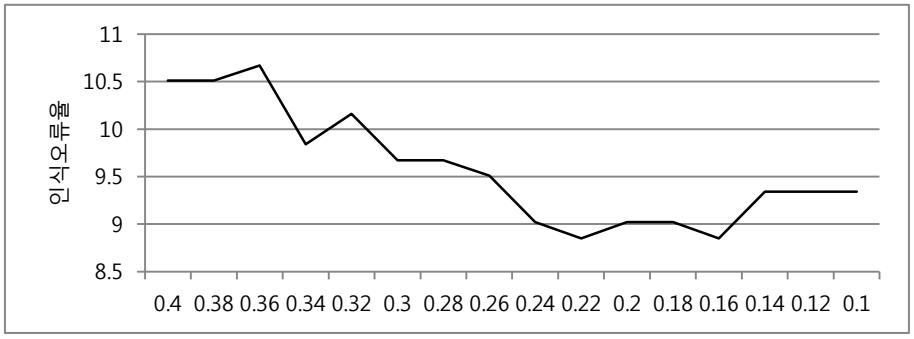


그림 18. 단어 간 코사인 유사도 최소화 모델 실험결과, 인식단어 크기 500

그림 18은 적응 데이터의 분량을 47*2 발화에서 500*1 발화로 늘렸을 때의 인식결과를 나타낸다. 적응 데이터의 분량을 늘렸을 때 통화표 단어 셋 3을 제외한 베이스라인, 조음점수, 유사도 최적화 실험결과의 인식률이 향상되었다. Cosine similarity 기준의 최적화 시 인식오류율은 최소 8.9%로 베이스라인에 비해 1.8% 낮았다. 조음점수 및 유사도 최적화의 결과가 적응 데이터의 분량을 늘렸을 때 보다 크게 향상됨을 확인할 수 있다.

제 7 장 결론

제 1 절 연구결과 요약 및 평가

본 논문에서는 마비말장애 화자가 음성 인터페이스를 사용함에 있어 발생하는 인식오류의 문제를 음향모델과 어휘모델의 층위에서 개선하고자 했다. 먼저 장애화자의 언어병리학적 연구결과를 바탕으로 발화속도를 반영하여 음향모델의 구조를 개선했다. 또한 어휘모델로서 최적의 인식단어를 선택하기 위해 음운론을 기초로 음소를 범주화하여 조음특징을 정의했고 어휘모델의 최적화로 장애발화의 인식오류를 줄일 수 있음을 인식실험을 통해서 확인했다. 따라서 본 논문은 언어학, 언어병리학의 연구결과가 음성공학의 개발로 이어지는 학제적 연구로서 의미를 가진다.

장애발화에 대한 어휘모델의 실험을 통한 검증에 앞서 인식모델을 최적화하기 위해 장애화자의 발화속도, 학습 데이터의 제한, HMM 출력확률 모델간 차이 등에 초점을 두고 HMM의 구조, 학습 알고리즘과 파라미터 값 등을 조정하여 최소오류율을 보이는 베이스라인 음향모델을 구축했다.

마비말장애 발화 인식의 단어 선택기준으로서 조음특징 기반의 혼합모델을 제안했고 제안된 기준에 의해 선택된 인식단어를 사용함으로써 인식성능이 향상됨을 확인했다. 단어의 발음열로부터 음소범주기반의 조음특징을 정의하고 조음특징과 인식오류 간의 관계를 일반화 선형혼합모델을

사용하여 모형화했고 자음특징과 모음특징을 단독으로 사용하여 모델을 각각 만들었을 때 자음특징에 비해 모음특징이 더 낮은 AIC를 보였다. 음소의 범주화 시 자음에 대해서 조음위치보다 조음방식을 기준으로 범주화 하는 모델이 더 낮은 AIC를 보였다. 모음에서는 혀의 위치가 혀의 높이 특징보다 낮은 AIC를 보였다. 조음위치와 혀의 위치를 변수로 사용한 모델구축결과에서 마찰음, 비음, 전설모음, 전설모음, 후설모음 등의 변수가 통계적으로 유의했다. 단어를 구성하는 음소 중 마찰음, 유음 등의 개수가 증가하면 인식오류가 증가하고 비음, 모음 등이 증가하면 인식 정확도가 증가하는 경향을 보였다. 모음 중에서 이중모음이 정확도의 향상정도가 가장 작고 중설모음이 가장 컸다. 파찰음과 폐쇄음은 추가할수록 인식 정확도가 증가하지만 증가량이 0에 가까웠다. 장애발화의 조음오류에 대한 선행연구(Byrne, 1959, Platt et al., 1980, Whitehill & Ciocca, 2000)에서는 마찰음, 유음 등의 조음오류는 크고 파열음, 비음과 모음 등의 오류는 상대적으로 적었다. 4장의 음소범주와 인식오류간의 관계와 비교하면 마찰음, 유음, 비음, 모음 등이 조음오류 연구와 일치하는 결과이다. 파열음의 인식 정확도에 대한 영향이 적은 것은 한국어의 파열음에는 경음이 포함되어 있기 때문으로 추정된다. 구축된 혼합모델을 이용해서 테스트 단어의 조음점수를 계산했고 조음점수를 기준으로 단어를 선택했다. 선택된 단어 리스트로 인식했을 때 베이스라인 중 최소오류율을 보인 결과에 비해 오류율이 2.74% 감소했다.

인식단어 리스트 내의 혼동을 유발하는 가설을 예측하기 위해 Levenshtein 거리와 Cosine 거리를 기준으로 단어 간 거리와 유사도를 정의했다. 단어간 유사도를 최소화 및 최대화하여 두 리스트간 인식오류율의 차이를 관찰했고 유사도를 이용해 인식 시 단어 리스트 내의 다른 단어에 의해 발생하는 인식오류를 예측할 수 있음을 확인했다. 단어간 유사도 최적화로 조음점수 기준 단어 리스트에 비해 최대 3%의 인식오류를 감소시켰다. 또한 음향모델 적응 데이터를 증가시켰을 때 두 기준 모두에서 인식률의 향상폭이 더 커짐을 확인했다.

제 2 절 기여도 요약

본 논문에서는 첫째로 장애화자의 발화속도를 모델링하고 HMM의 출력확률모델의 종류와 파라미터를 조정하여 주어진 음성 데이터에 대해 최적의 성능을 보이는 인식환경을 탐색함으로써 베이스라인 음향모델을 구축했다. 정상화자와 장애발화의 평균 발화속도를 측정하여 비교했고 비율에 따라 특징 추출 시 윈도우 크기를 조정하여 테스트 셋에 대해 약 0.6%, HMM을 구성하는 state 개수를 조정하여 약 1%의 인식오류를 줄였다. 출력확률모델로 GMM, SGMM, DNN 등을 비교했고 파라미터의 크기를 조정하여 각각 3.6%, 1.7%, 5.5%의 인식오류율을 감소시켰다. 둘째로 자모 음소를 범주화하고 각 범주가 인식오류에 미치는 영향력을 통계적으로 모델링했다. 또한 이를 바탕으로 단어에 적용되는 조음점수를

정의하고 높은 점수를 가지는 단어로 인식단어 리스트를 구성하여 인식오류를 낮출 수 있음을 실험으로 확인했다. 장애발화의 인식결과에 대한 범주의 영향력은 일반화된 선형혼합모델로 모형화했다. 자음보다 모음이 높은 영향력을 보였고 자음은 조음방법, 모음은 혀의 위치로 범주화했을 때 가장 낮은 AIC값을 보였다. 또한 장애화자의 조음오류에 대한 선행연구와 인식오류에 대한 연구결과의 공통점과 차이점을 비교했다. 셋째로 열 간 거리 기준으로 단어 간 유사도를 정의했고 유사도의 높고 낮음이 인식결과에 영향을 주는 요인이 됨을 확인했다. 또한 이를 기반으로 단어 리스트 최적화하여 인식오류를 보다 낮출 수 있음을 확인했다. 조음점수를 최대화하고 단어 간 유사도를 최소화함으로써 인식단어 리스트를 최적화 했을 때 기존 베이스라인 단어 리스트에 비해 절대적으로 5.7%, 상대적으로 34.5%의 낮은 인식오류를 보였다.

제 3 절 향후 연구

마비말장애 발화특성에 대한 선행연구(Duffy, 2013)에서는 신경계의 손상 부위와 발화의 특성에 따라 마비말장애의 유형을 강직형, 이완형, 실조형, 과다운동형, 과소운동형, 혼합형 등으로 분류했고 같은 유형에 속하는 화자간에도 손상 위치와 정도에 따라 발화의 특성에 차이가 발생한다.

따라서 단어 리스트 구성 시 장애유형과 개별화자의 특성을 고려하는 맞춤형 접근방식이 인식오류율을 보다 낮추기 위한 방법이 될 것이다.

음성인식에서 오류는 단어 외에도 단어의 발음변이, HMM state, GMM 등 층위에서의 영향을 받아 발생할 수 있다. 본 논문에서는 어휘모델링의 최적화에 초점을 두어 단어간 유사도를 통해 가설의 혼동 유발을 추정했으나 단어 내외 발음변이의 문맥적 요인에 의한 인식오류 유발, HMM state 또는 GMM 과 같은 출력확률 모델 수준에서의 인식오류 유발을 반영한다면 보다 정밀한 모델을 통해 추정의 정확도를 높일 수 있을 것이다. 또한 현재 특정 단어의 유사도를 정의하기 위해 단어 리스트에 존재하는 모든 단어와의 유사도를 계산하고 최대값, 즉 유사도가 가장 높은 단어와의 유사도를 대푯값으로 근사화했는데 정보이론에서의 혼잡도(perplexity) 개념을 도입한다면 특정 단어와 경쟁하는 다수의 단어와의 관계를 반영할 수 있을 것이다.

음향모델 학습 데이터의 부족 문제를 보완하기 위해 대량의 정상발화 데이터를 장애발화의 특성에 맞추어 왜곡시켜 학습 데이터로서 활용하는 방법을 제안했다. 데이터의 특성 변환에 의해 상대적인 인식오류 감소의 가능성을 확인할 수 있었지만 발화 왜곡 시 장애발화의 속도에만 초점을 두어 장애발화의 특성차이를 충분히 반영하지 못한 제한적인 모델링의 결과였다. 향후 데이터 변환과정에 발화속도 차이 외에도 장애발화의

음향/음성적 특징에 대한 선행연구에서 밝혀진 요인을 반영함으로써 보다 현상에 가까운 모델을 구축할 수 있을 것이다.

4 장에서 단어의 조음특징과 인식률의 관계를 혼합선형 회귀모델로 설명했다. 자음 및 모음의 군집, 장애정도 등을 모델의 고정변수로, 단어와 화자 등을 임의변수로 가정했으나 장애화자의 조음 능력을 잘 반영할 것으로 추측되는 음소의 연쇄관계는 변수로 반영하지 않았다. 변수로 추가 시 AIC 가 오히려 증가했기 때문인데 이것은 단어의 구성이 500 단어로 제한되어 있고 문맥에 독립적인 모노폰 수준에서 음소의 균형을 맞추도록 설계되어 단어 내 음소문맥의 발견빈도가 충분하지 않기 때문으로 추정된다. 다양한 음소문맥을 포함하는 단어의 발화 데이터 수집과 인식결과를 추가하여 통계모델에 음소문맥의 영향을 반영함으로써 단어 선택 모델이 인식오류를 보다 정교하게 예측할 수 있을 것이다.

참고 문헌

Allen, J., Hunnicutt, M., Klatt, D., Armstrong, R. & Pisoni, D. (1987). From Text to Speech: The MITalk System. *Cambridge University Press*.

Bahl, L., Brown, P., Souza, P., & Mercer, R. (1986). Maximum mutual information estimation of hidden Markov model parameters for speech recognition. *Proceedings of ICASSP,* 49-52.

Bisani, M., & Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication, 50*(5), 434-451.

Byrne, M. (1959). Speech and Language Development of Athetoid and Spastic Children. *Journal of Speech and Hearing Disorders,* 231-231.

Choi, D., Kim, B., Lee, Y., Um, Y., & Chung, M. (2011). Design and creation of Dysarthric Speech Database for development of QoLT software technology. *Proceedings of International Conference on Speech Database and Assessments,* 47-50.

Choi, D., Kim, B., Kim, Y., Lee, Y., Um, Y., & Chung, M. (2012). Dysarthric Speech Database for Development of QoLT Software Technology. *Proceedings of LREC,* 3378-3381.

Christensen, H., Cunningham, S., Fox, C., Green, P., & Hain, T. (2012). A comparative study of adaptive, automatic recognition of disordered speech. *Proceedings of INTERSPEECH,* 1776-1779.

Christensen, H., Aniol, M., Bell, P., Green, P., Hain, T., King, S., & Swietojanski, P. (2013). Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech. *Proceedings of INTERSPEECH,* 3642-3645.

Christensen, H., Green, P., & Hain, T. (2013). Learning speaker-specific pronunciations of disordered speech. *Proceedings of INTERSPEECH,* 1159-1163.

Duffy, J. (2013). *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Mosby.

Dworzynski, K., & Howell, P. (2004). Predicting stuttering from phonetic complexity in German. *Journal of Fluency Disorders, 29*(2), 149-173.

Federico, M., Bertoldi, N., & Cettolo, M. (2008). IRSTLM: an open source toolkit for handling large scale language models. *Proceedings of INTERSPEECH,* 1618-1621.

Fung, P., & Liu, Y. (2005). Effects and modeling of phonetic and acoustic confusions in accented speech. *The Journal of the Acoustical Society of America, 118*(5), 3279-3293.

Gauvain, J., & Lee, C. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing, 2*(2), 291-298.

Gemmeke, J., Ons, B., Tessema, N., Van hamme, H., van de Loo, J., Pauw, G., Daelemans, W., Huyghe, J., Derboven, J., Vuegen, L., Van Den Broeck, B., Karsmakers, P., & Vanrumste B. (2013). Self-taught assistive vocal interfaces: An overview of the ALADIN project. *Proceedings of INTERSPEECH,* 2038-2043.

Hamidi, F., Baljko, M., Livingston, N., & Spalteholz, L. (2010). CanSpeak: A Customizable Speech Interface for People with Dysarthric Speech. *Computers Helping People with Special Needs, 6179*, 605-612.

Hawley, M., Enderby, P., Green, P., Cunningham, S., Brownsell, S., Carmichael, J., . . . Palmer, R. (2007). A speech-controlled environmental control system for people with severe dysarthria. *Medical Engineering & Physics, 29*(5), 586-593.

Hawley, M., Cunningham, S., Green, P., Enderby, P., Palmer, R., Sehgal, S., & O'neill, P. (2013). A Voice-Input Voice-Output Communication Aid for People With Severe Speech Impairment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 21*(1), 23-31.

Hinton, G., Osindero, S., & Teh, Y. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation, 18*(7), 1527-1554.

Hux, K., Rankin-Erickson, J., Manasse, N., & Lauritzen, E. (2000). Accuracy of three speech recognition systems: Case study of dysarthric speech.*Augmentative and Alternative Communication, 16*(3), 186-196.

Kent, R. D. (1992). The biology of phonological development. *Phonological development: Models, research, implications,* 65-90.

Kim, H., Martin, K., Hasegawa-Johnson, M., & Perlman, A. (2010). Frequency of consonant articulation errors in dysarthric speech. *Clinical Linguistics & Phonetics, 24*(10), 759-770.

Kim, S., Hwang, Y., Shin, D., Yang, C. Y., Lee, S. Y., Kim, J., ... & Chung, M. (2013). VUI development for Korean people with dysarthria. Journal of Assistive Technologies, 7(3), 188-200.

Kingsbury, P., Strassel, S., McLemore, C., & MacIntyre, R. (1997). CALLHOME american english lexicon (PRONLEX). Linguistic Data Consortium, Philadelphia.

Lee, E. J., Han, J. S., & Sim, H. S., The Effects of the Phonetic Complexity on the Disfluencies and the Articulation Errors of People Who stutter., 9(3), 139-156.
(이은주, 한진순, & 심현섭. (2004). 조음복잡성이 비유창성과 조음오류에 미치는 영향. 언어청각 장애연구, 9(3), 139-156.)

Leggetter, C., & Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language,* 171-185.

Leonard, R., & Doddington, G. (1993). Tidigits speech corpus. *Texas Instruments, Inc*.

Liu, Y., & Fung, P. (2005). Acoustic and phonetic confusions in accented speech recognition. *Proceedings of INTERSPEECH,* 3033-3036.

Liu, H., Tsao, F., & Kuhl, P. (2005). The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy. *The Journal of the Acoustical Society of America, 117*(6), 3879-3889.

Matsumasa, H., Takiguchi, T., Ariki, Y., Li, I., & Nakabayashi, T. (2009). Integration of Metamodel and Acoustic Model for Dysarthric Speech Recognition. *Journal of Multimedia, 4*(4), 254-261.

Mengistu, K., & Rudzicz, F. (2011a). Adapting acoustic and lexical models to dysarthric speech. *Proceedings of ICASSP,* 4924-4927.

Mengistu, K., & Rudzicz, F. (2011b). Comparing Humans and Automatic Speech Recognition Systems in Recognizing Dysarthric Speech.*Advances in Artificial Intelligence,* 291-300.

Morales, S., & Cox, S. (2009). Modelling Errors in Automatic Speech Recognition for Dysarthric Speakers. *EURASIP Journal on Advances in Signal Processing,* 308-340.

Nakashika, T., Yoshioka, T., Takiguchi, T., Ariki, Y., Duffner, S., & Garcia, C. (2014). Dysarthric speech recognition using a convolutive bottleneck network. *Proceedings of 2014 12th International Conference on Signal Processing (ICSP),* 505-509.

Pallett, D., Fiscus, J., & Garofolo, J. (1992). Resource management corpus: September 1992 test set benchmark test results. *Proceedings of ARPA Microelectronics Technology Office Continuous Speech Recognition Workshop*.

Paul, D., & Baker, J. (1992). The design for the Wall Street Journal-based CSR corpus. *Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics*.

Platt, L., Andrews, G., & Howie, P. (1980). Dysarthria of Adult Cerebral Palsy: II. Phonemic analysis of articulation errors. *Journal of Speech Language and Hearing Research, 23*, 41-55.

Polur, P., & Miller, G. (2006). Investigation of an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals. *Medical Engineering & Physics,28*(8), 741-748.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., Glembek, O., Goel, N., Karafiat, M., Rastrow, A., Rose, R. C., Schwarz, P., & Thomas, S. (2011). The subspace Gaussian mixture model—A structured model for speech recognition. *Computer Speech & Language, 25*(2), 404-439.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesel, K. (2011). The Kaldi speech recognition toolkit.

Raghavendra, P., Rosengren, E., & Hunnicutt, S. (2001). An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems. *Augmentative and Alternative Communication, 17*(4), 265-275.

Robinson, A. The british english example pronunciation (beep) dictionary. URL: http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html.

Rudzicz, F. (2013). Adjusting dysarthric speech signals to be more intelligible. *Computer Speech & Language, 27*(6), 1163-1177.

Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lööf, J., Schlüter, R., & Ney, H. (2009). The RWTH aachen university open source speech recognition system. *Proceedings of INTERSPEECH,* 2111-2114.

Sharma, H., & Hasegawa-Johnson, M. (2013). Acoustic model adaptation using in-domain background models for dysarthric speech recognition.*Computer Speech & Language, 27*(6), 1147-1162.

Slobada, T., & Waibel, A. (1996). Dictionary learning for spontaneous speech recognition. *Proceeding of Fourth International Conference on Spoken Language Processing, 4*, 2328-2331.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit." *Proceedings of INTERSPEECH*.

Strik, H., & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication, 29*(2), 225-246.

Thomas-Stonell, N., Kotler, A., Leeper, H., & Doyle, P. (1998). Computerized speech recognition: Influence of intelligibility and perceptual consistency on recognition accuracy. *Augmentative and Alternative Communication, 14*(1), 51-56.

Throneburg, R., Yairi, E., & Paden, E. (1994). Relation Between Phonologic Difficulty and the Occurrence of Disfluencies in the Early Stage of Stuttering. *Journal of Speech, Language, and Hearing Research,37*(3), 504-509.

Tolba, H., & El_Torgoman, A. (2009). Towards the improvement of automatic recognition of dysarthric speech. *Proceedings of 2009 2nd IEEE International Conference on Computer Science and Information Technology,* 277-281.

Torre, D., Villarrubia, L., Hernandez, L., & Elvira, J. (1997). Automatic alternative transcription generation and vocabulary selection for

flexible word recognizers. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2*, 1463-1466.

Tsai, M., Chou, F., & Lee, L. (2007). Pronunciation Modeling With Reduced Confusion for Mandarin Chinese Using a Three-Stage Framework. *IEEE Transactions on Audio, Speech and Language Processing, 15*(2), 661-675.

Van der Loo, M. (2014). The stringdist package for approximate string matching. *The R*.

Vesely, K., Ghoshal, A., Burget, L., & Povey, D. (2013). Sequence-discriminative training of deep neural networks. *Proceedings of INTERSPEECH*, 2345-2349.

Weide, R. (2005). The Carnegie mellon pronouncing dictionary [cmudict. 0.6]. Carnegie Mellon University: http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Whitehill, T., & Ciocca, V. (2000). Speech errors in Cantonese speaking adults with cerebral palsy. *Clinical Linguistics & Phonetics, 14*(2), 111-130.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., & Woodland, P. (2006). The HTK book (for HTK version 3.4).

Yu, D., & Deng, L. (2014). *Automatic speech recognition: A deep learning approach (Signals and Communication Technology)*. Springer.

신지영. (2011). 한국어의 말소리. *지식과 교양*.

부록

1. 한국어 의사음소단위(Phone-Like Unit, PLU) 정의

번호	음소	PLU	번호	음소	PLU
1	ㅂ	P	26	ㅏ	AA
2		PQ	27	ㅓ	AX
3	ㅃ	PP	28	ㅗ	OW
4	ㅍ	PH	29	ㅜ	UW
5	ㄷ	T	30	ㅡ	WW
6		TQ	31	ㅣ	IY
7	ㄸ	TT	32	ㅐ	EH
8	ㅌ	TH	33	ㅔ	EY
9	ㄱ	K	34	ㅚ	OI
10		KQ	35	ㅟ	UI
11	ㄲ	KK	36	ㅑ	JA
12	ㅋ	KH	37	ㅕ	JX
13	ㅈ	Z	38	ㅛ	JO
14	ㅉ	ZZ	39	ㅠ	JU
15	ㅊ	CH	40	ㅒ	JH
16	ㅅ	S	41	ㅖ	JE
17	ㅆ	SS	42	ㅘ	WA
18	ㅎ	HH	43	ㅝ	WX
19	ㅁ	M	44	ㅙ	WH
20		MM	45	ㅞ	WE
21	ㄴ	N	46	ㅢ	WI
22		NN	47	묵음 1	sp
23	ㅇ	NX	48	묵음 2	sil
24	ㄹ	L			
25		R			

2. 173 단어 리스트

구분	단어	구분	단어	구분	단어
APAC 37	포도	APAC 37	이빨	KP 36	한강
	딸기		거북이		아버지
	사탕		뱀		야자수
	햄버거		호랑이		어머니
	옥수수		고래		연못
	컵		찢어요		오징어
	빨대		싸워요		요지경
	책		아파요		우편
	색종이		병원		유달산
	머리		안경		은방울
	양말		없어요		이순신
	단추		올라가요		앵무새
	모자	KP 36	기러기		엑스레이
	장갑		나폴리		하나
	빗		도라지		둘
	우산		로마		삼
	침대		미나리		넷
	화장실		바가지		오
	나무		서울		여섯
	꽃		잉어		칠
	바퀴		지게		팔
	그네		치마		아홉
	시소		키다리		공
	눈사람		통신	CW 100	거절
	토끼		파고다		검색

2. 173 단어 리스트

구분	단어	구분	단어	구분	단어
CW 100	곡선택	CW 100	문장선택	CW 100	열번째
	구번		물음표		영상통화
	기호		복사		영어
	김치		붙이기		오른쪽
	끄기		빠르게		오번
	네번째		사번		왼쪽
	느낌표		사진찍기		위로
	느리게		사진촬영		육번
	다섯번째		삼번		음악재생
	다음		선택		음악제목
	다음칸		설정		응급상황
	단어선택		세번째		이동
	동영상		숫자		이메일
	두번째		쉼표		이메일보기
	뒤로		스페이스		이메일보내기
	디엠비		시작		이메일쓰기
	마침표		실행		이번
	맨뒤로		십번		이전
	맨앞으로		십일번		인식결과선택
	메모장		아래로		인식시작
	명령어리스트		아롱아		인식종료
	문자		아홉번째		인터넷
	문자보기		앞으로		일곱번째
	문자보내기		여덟번째		일번
	문자쓰기		여섯번째		임의재생

2. 173 단어 리스트

구분	단어	구분	단어	구분	단어
CW 100	작게	CW 100	정지	CW 100	카메라
	재생		종료		켜기
	저장		채널		크게
	전송		첫번째		팔번
	전체		축소		한글
	전화걸기		취소		한칸띄우기
	전화받기		치즈		확대
	정열		칠번		

3. 500 단어 리스트

구분	기능	단어	구분	기능	단어
기기 제어 명령어 105	전화	전화	기기 제어 명령어 105	위치알림	위치알림
		연락			위치
		걸기			자리
		다이얼			어디
		자라			비둘기
	메시지	메시지		네이버	네이버
		문어			누나
		마당			냉이
		미니			레몬
		만일			머루
	일정	일정		다음	한메일
		하루			다리미
		할일			대나무
		모임			단골
		참새			다래
	날씨	날씨		구글	구글
		기상			거미
		일기			기차
		예보			보리수
		개나리			구구구
	뉴스	뉴스		탐색기	탐색기
		신문			찾기
		소식			터미널
		기사			태아
		제비			토란

3. 500 단어 리스트

구분	기능	단어	구분	기능	단어
기기 제어 명령어 105	계산기	계산기	기기 제어 명령어 105	확인	확인
		더하기			오케이
		연산			승인
		셈			비버
		가마니			자두
	위로	아이		취소	다시
		나비			정정
		앵두			친구
		비행기			치타
		도토리			오미자
	아래로	두더지		이전	요전
		밑으로			오리
		아리아			이리
		아이스			이구아나
		지렁이			노루
	왼쪽	좌측		다음	이후
		왼편			악어
		파파야			새우
		고니			두루미
		파리			기린
	오른쪽	우측		보내기	보내기
		오른편			전달
		망고			발송
		오소리			발신
		모기			개구리

3. 500 단어 리스트

구분	기능	단어	구분	기능	단어
기기 제어 명령어 105	종료	끝내기	통화표 108	ㅇ	아줌마
		마무리			엄마
		그만			아내
		완료		ㅈ	자유
		부엉이			지우개
통화표 108	ㄱ	가구			주머니
		고구마		ㅊ	치약
		가게			차표
	ㄴ	너구리			차이
		나라		ㅋ	카레
		노인			카라멜
	ㄷ	도마			카드
		다리		ㅌ	타조
		단어			타잔
	ㄹ	라면			토마토
		라디오		ㅍ	파자마
		라이터			피자
	ㅁ	마루			파도
		매미		ㅎ	하와이
		모내기			하마
	ㅂ	바나나			호미
		바다		ㅏ	아몬드
		비타민			아기
	ㅅ	사다리			아이고
		소리			
		소나기			

3. 500 단어 리스트

구분	기능	단어	구분	기능	단어
통화표 108	ㅑ	야구	통화표 108	ㅣ	이미지
		야후			이모
		야간			인간
	ㅕ	엉덩이		ㅐ	애인
		언니			애니콜
		언어			애벌레
	ㅖ	여자		ㅔ	에어컨
		여보			에누리
		여우			에너지
	ㅗ	오디오		1	일지매
		오뚜기			일본
		오이			일반인
	ㅛ	요가		2	둘째
		요리			이사
		요일			이마
	ㅜ	우리		3	석삼
		우유			삼다수
		운반			삼거리
	ㅠ	유도		4	사오정
		유리			사투리
		유머			사나이
	ㅡ	으악새		5	오늘
		으차차			오렌지
		음악			오토바이

3. 500 단어 리스트

구분	기능	단어	구분	기능	단어
통화표 108	6	육아	통화표 108	9	구렁이
		육군			구두
		육개장			구구단
	7	칠뜨기		0	십자가
		칠판			십자수
		칠면조			공기
	8	팔봉산			
		팔공산			
		팔다리			

3. 500 단어 리스트

구분	단어	구분	단어	구분	단어
	햇볕		쾌감		빠른우편
	국회의원		낚싯대		재활용품
	한의원		옆집		바퀴벌레
	뜻밖		뼈다귀		스웨터
	외갓집		생활용품		부잣집
	밀크쉐이크		사회생활		치과의사
	구슬꿰기		전기밥솥		집단적
	불쾌지수		계좌이체		텍스트
	앞뒤		웨이브파마		빨래건조대
	돼지갈비		위원장		중계방송
	계약서		영향력		해수욕장
	얘기하기		교육비		취향
PBW	떡볶이	PBW	적극적	PBW	퇴직금
287	얘기	287	일회용품	287	식생활
	의원		인터넷예매		망원경
	직접적		체육관		바깥쪽
	귓속		편의점		휴대폰로밍
	생활환경		돼지고기		베스트셀러
	사회주의		초등학교		반딧불이
	소프트웨어		슈퍼마켓		호우주의보
	햇빛		된장찌개		충격적
	학교생활		초등학생		휴대전화
	월드컵		멀리뛰기		오랫동안
	계획		밥그릇		돼지
	뒷골목		퇴원		하드웨어
	콤플렉스		옛날이야기		고등학교

3. 500 단어 리스트

구분	단어	구분	단어	구분	단어
	가위바위보		인쇄		전통문화
	체크카드		예식장		횡단보도
	횟수		대중교통		탈의실
	답변		컴퓨터		꼭대기
	무선인터넷		빨래바구니		대합실
	세계관		샤워		환경오염
	유치원		뒷모습		느타리버섯
	집중적		편의		출퇴근
	분실물센터		문화센터		효율적
	고등학생		저녁때		어젯밤
	약혼녀		체계적		이해관계
	돛단배		웨이터		형식적
PBW	자연환경	PBW	외교관	PBW	볶음밥
287	대학교수	287	이것저것	287	애견카페
	최고급		롤러코스터		쓰레기통
	규칙적		출입국		월간도서
	예의		약속		후춧가루
	윗사람		해외여행		웬일
	데칼코마니		계란프라이		민주주의
	특별		중국집		평양
	회원		객관적		치료법
	관광버스		뒷산		생활수준
	귀뚜라미		정치권		등록증
	졸업생		베네수엘라		오케스트라
	유채꽃		교육자		떡국
	우편배달부		빗방울		인간관계

3. 500 단어 리스트

구분	단어	구분	단어	구분	단어
	팩시밀리		동창회		밥솥
	권위		경복궁		세종대왕
	고속버스		딸기우유		예약
	노랫소리		하룻밤		휠체어
	중학교		공중전화		껍질
	에스프레소		동물원		물뿌리개
	간접적		특급		식료품
	회의		휘발유		안전벨트
	루미큐브		철학적		등록금
	계약		네비게이션		대중문화
	미술치료사		양배추		증권사
	태권도		플라스틱		고속도로
PBW	본격적	PBW	의도적	PBW	칼국수
287	촛불	287	습관	287	외야수
	유학생		티머니카드		육체적
	텔레비전		학생증		우편번호
	대피훈련		백두산		의욕
	샌드위치		재활용		동대문시장
	수도꼭지		경쟁력		국제화
	핫도그		세계적		초콜릿
	양옆		월급		옆방
	물리치료사		밑바닥		삼계탕
	일상생활		일회용		달맞이꽃
	신입생		전화번호		뒤편
	이곳저곳		큰딸		인천일호선
	메니큐어		고춧가루		특수성

3. 500 단어 리스트

구분	단어	구분	단어	구분	단어
PBW 287	대표적	PBW 287	양쪽	PBW 287	엘리베이터
	딸기주스		크리스마스		하숙집
	약점		시청률		아랫배
	원피스		딴지일보		관광객
	비스켓		위협		뮤지컬
	의류		레몬레이드		토론회
	남대문시장		관계자		홈쇼핑
	협상		팝송		낚시꾼
	최대한		언어치료사		뜨게질
	대학교		과학적		오천원
	경찰관		꼬꼬면		갈비뼈
	불꽃놀이		종업원		양상추
	장례식		트위터		아랫사람
	백화점		한의사		배드민턴
	빨간색		점심때		신혼여행
	뱃사람		약혼자		뒤쪽
	월요일		한평생		비뇨기과
	외할아버지		초록색		

Abstract

Optimizing Vocabulary Modeling for Dysarthric Voice User Interface

Minsoo Na

Interdisciplinary Program in Cognitive Science

The Graduate School

Seoul National University

Articulation errors, disfluency, impulsive pause, low speaking rate have been suggested as factors of recognition error for dysarthric speakers using voice user interface. In related works, methods for correcting dysarthric speech, AM adaptation, pronunciation variation modeling, grammar modeling and vocabulary modeling based on acoustic and phonetic analyses on dysarthric speech were proposed to compensate those factors. In this paper, acoustic model was optimized. And words in the vocabulary were selected by the GLMM which had modeled relationship between recognition errors and articulatory features for phonetic class and optimized by lowering similarity between words.

Problems in training AM for dysarthric speech recognition were addressed: firstly low speaking rate was compensated by varying the window length of FFT and the number of states of HMM. Secondly the efficiency of models for emission probability of HMM was compared. Thirdly AM trained using large amount of non-dysarthric speech was experimented. Fricative and nasal were statistically significant in the analysis of relation between recognition error and consonant classes. And all vowel classes were significant. AIC was lower by classifying consonants based on manner of articulation than based on place and by classifying vowels based on position of tongue than based on height. Fricatives increased WER and nasals increased accuracy of recognition. Estimates of plosive, affricate, liquid were close to zero. All vowel classes increased accuracy. Estimate of central vowel was the largest followed by back vowel, front vowel and diphthong. Triggering recognition error by competitive words was modeled by similarity between words based on Levenshtein and cosine distance respectively. Effect of similarity between words on recognition result was confirmed by the minimum-maximum similarity contrast and the N-best prediction. Prior to model vocabulary, articulation score for each word was calculated. Firstly the vocabulary was composed of the words with the

maximum articulation scores. Secondly the words in the vocabulary with high similarity were replaced by the word with less similarity and large articulation score. In dysarthric speech recognitions, the optimized vocabulary lowered WER by 5.72% (34.60% ERR).

Keywords: Automatic speech recognition, Dysarthria, Acoustic modeling, Vocabulary modeling, Articulation error, Similarity between words

Student Number: 2006-30742