한국음향학회지 제21권 제3호 pp. 330~338 (2002)

# 대화체 연속음성 인식을 위한 한국어 대화음성 특성 분석

## Analysis of Korean Spontaneous Speech Characteristics for Spoken Dialogue Recognition

박 영 희*, 정 민 화*
(Young-Hee Park*, Minhwa Chung*)

*서강대학교 컴퓨터학과 음성언어처리연구실
(접수일자: 2002년 2월 5일; 채택일자: 2002년 2월 14일)

대화체 연속음성은 자연스러운 발화로 낭독체 문장에 비해 잡음, 간투어와 같은 비문법적인 요소가 많고, 발음의 변이가 심하다. 이런 이유로 대화체 연속음성을 인식하기 위해서는 대화 현상을 분석하고 그 특징을 반영하여야 한다. 본 논문에서는 실제 대화음성에 빈번히 나타나는 대화 현상들을 분류하고 각 현상들을 모델링하여 대화체 연속음성 인식을 위한 기본 베이스라인을 구축하였다. 대화 현상을 묵음 구간과 잡음, 간투어, 반복/수정 발화의 디스플루언시 (disfluencies), 표준전사와 다른 발음을 갖는 발음변이 현상으로 나누었다. 발음변이 현상은 다시 양성음의 음성음화, 음운축약/탈락 현상, 패턴화된 발음변이, 발화 오류로 세분화하였다. 대화체 음성인식을 위해서 빈번히 나타나는 묵음 구간을 고려한 학습과 잡음, 간투어 처리를 위한 음향모델을 각각 추가하였다. 발음변이 현상에 대해서는 출현빈도수가 높은 것들만을 대상으로 발음사전에 다중발음열을 추가하였다. 대화 현상을 고려하지 않고 낭독체 스타일로 음성인식을 수행하였을 때 형태소 에러율 (MER: Morpheme Error Rate)은 31.65%였다. 이에 대한 형태소 에러율의 절대값 감소는 묵음 모델과 잡음 모델을 적용했을 때 2.08%, 간투어 모델을 적용했을 때 0.73%, 발음변이 현상을 반영했을 때 0.92%였으며, 최종적으로 27.92%의 형태소 에러율을 얻었다. 본 연구는 대화체 연속음성 인식을 위한 기초 연구로 음향모델과 어휘모델, 언어모델 각각에 대한 베이스라인으로 삼고자 한다.
**핵심용어:** 대화체 연속음성 인식, 대화 현상, 대화체 발음변이, 간투어, 잡음 모델
**투고분야:** 음성처리 분야 (2.5, 2.7)

Spontaneous speech is ungrammatical as well as serious phonological variations, which make recognition extremely difficult, compared with read speech. In this paper, for conversational speech recognition, we analyze the transcriptions of the real conversational speech, and then classify the characteristics of conversational speech in the speech recognition aspect. Reflecting these features, we obtain the baseline system for conversational speech recognition. The classification consists of long duration of silence, disfluencies and phonological variations; each of them is classified with similar features. To deal with these characteristics, first, we update silence model and append a filled pause model, a garbage model; second, we append multiple phonetic transcriptions to lexicon for most frequent phonological variations. In our experiments, our baseline morpheme error rate (MER) is 31.65%; we obtain MER reductions such as 2.08% for silence and garbage model, 0.73% for filled pause model, and 0.92% for phonological variations. Finally, we obtain 27.92% MER for conversational speech recognition, which will be used as a baseline for further study.
***Keywords:*** *Conversational speech recognition, Spontaneous speech recognition, Disfluencies, Noise, Pronunciation variations, Filled pauses, Garbage model*
***ASK subject classification:*** *Speech signal processing (2.5, 2.7)*

책임저자: 박영희 (yhpark@nlpeng.sogang.ac.kr)
127-742 서울시 마포구 신수동 1번지
서강대학교 컴퓨터학과 음성언어처리연구실
(전화: 02-712-8023; 팩스: 02-704-8273)

# I. 서 론

음성은 컴퓨터와 정보를 전달하는 매우 편리한 수단으로 음성 인식을 위한 연구가 활발히 진행 중이다. 고립단어 인식의 단계를 넘어 대어휘의 연속음성 인식 단계로 연구가 진행되고 있다. 그러나 이들 연구는 대부분이 낭독체 연속 음성을 대상으로 하고 있다. 한국어 대화체 인식 시스템에 대한 연구[3]에서는 대화체에서 많이 발생하는 잡음, 간투어 (filled pause)를 하나의 음향모델로 모델링하여 적용하였다. 그러나 아직까지 자연스러운 발화에 대한 체계적인 연구가 미비한 상태이다. 자연스러운 연속음성을 인식하기 위해서는 한국어 대화체 음성의 대화 현상을 분석하고, 분석을 통해 얻어진 특징들을 반영하여 대화체 연속음성 인식의 성능을 개선하는 노력을 하여야 하겠다.

한국어 대화 특성에 대한 기존의 연구는 대화체에 대해 자연어처리 측면에서 대화체 문장들을 분석하는 것이 일반적이며, 언어학적인 지식을 이용하여 대화체 문장의 특징을 분류하고 있다[1]. 그러나 대화체 음성의 인식을 위한 분석은 그리 많지 않은 편이다. 영어나 타 언어에 대해서는 대화체 연속음성 인식을 위한 대화 특성의 분석이 매우 체계적으로 이루어지고 있으며, 대화체 연속음성의 인식뿐 아니라 대화 시스템과의 연결을 위한 자연어처리 단계와 연속성을 갖도록 하는 연구가 진행되고 있다[4-9,12-15].

낭독체 음성은 발화시 발음이나 문법적으로 오류가 없도록 녹취가 이루어지지만, 자연스러운 발화에서는 "하구요", "어트케"와 같이 음향학적으로 표준 발음이 아닌 발음들과 문법적으로도 완전하지 않은 문장의 형태, 간투어와 같은 비문법적인 발화, 입술소리 같은 발화 도중의 잡음 등의 많은 대화 현상을 포함한다.

영어 대화체에 대한 연구는 Switchboard[10] 코퍼스를 주 대상으로 하여 연구가 진행되고 있으며, 대화현상에 대해서 디스플루언시의 분석[4,7-9,12]과 발화 스타일에 따른 발음변이의 연구[5,6,13,14]가 주류를 이루고 있다.

[7,8]의 연구는 디스플루언시를 체계적으로 분류하고, 다양한 분석을 통해 예전의 연구에서는 단순한 잡음으로 분류되던 디스플루언시가 체계적인 분포를 보일 뿐 아니라 예측 기능이 있음을 보였다. [4,12]에서는 디스플루언시 정보를 통계적 언어모델 생성에 이용하였다. 큰 성능개선을 얻지는 못했지만 반복, 삭제 현상이 혼잡도의 감소시키므로 정보를 포함하고 있음을 확인하였다. 또한 간투어는 위치에 따른 혼잡도가 다르고 대부분이 문장의 시작부분에 나타나므로 음성 분할시에 이용할 수 있음을 말하였다.

발음변이에 대한 연구는 단어내부의 발음변이와 단어간의 발음변이로 나뉘고, 단어내부의 발음변이는 다중 발음 사전을 이용하고, 단어간의 발음변이는 *multiwords*[1]를 발음사전에 추가하는 것이 일반적이다[5,6,13]. 또한 [13]의 연구에서는 같은 대화를 각기 다른 스타일로 녹음하여 인식 실험을 수행하였다. WER가 낭독체일 때 28.8%에서 자연스러운 발화시 52.6%로 떨어져 자연스러운 발화의 인식이 어려운 과제임을 보여주고 있다.

본 논문에서는 자연스러운 발화시 나타나는 대화 현상을 음성인식 측면에서 분류하고, 기본적인 낭독체 인식 기술을 적용할 때의 문제점 및 개선 사항들에 대해 논의하고자 한다.

# II. 대화체 음성 데이터베이스의 특징

본 연구에 사용된 대화체 음성 데이터베이스는 서강대학교 음성언어처리연구실에서 한국전자통신연구원의 용역으로 98, 99년도에 C-STAR 과제를 위해서 구축한 대화체 음성 데이터 베이스이다. 여행 계획을 위한 가상의 대화이며 시나리오 설정과 대화시에 발화의 자유도에 차이를 두었다. 앞으로 TP#1, TP#2의 명칭을 사용한다.

두 데이터 베이스 모두 여행사 직원과 고객의 두 사람이 한 조가 되어 대화를 이끌어 나간다. 각 시나리오는 호텔 예약과 교통편 문의 등을 포함하는 복합적인 내용으로 구성된다.

TP#1은 총 25개의 시나리오로 구성되어 있고, 25조가 한 조당 4개의 대화를 발화하여 총 100개의 대화로 구성된다. 자연스러운 발화를 위해 여행사 직원에게만 자세한 정보를 주고 고객은 문의하면서 예약/변경/취소 등의 목적을 달성하도록 하였다. 즉 완성된 형태의 문장이 주어지지 않으므로 발화 중간에 머뭇거림이나 반복, 수정, 재질문 등이 빈번히 일어나므로 실제 대화와 매우 유사하다.

TP#2는 15개의 시나리오를 25조가 각 조당 5개의 대화를 발화하여 총 125개의 대화로 이루어져 있다. 난이도에 있어서 TP#2는 시나리오도 단순하고 대화시 선택 사항도 적게 하는 등의 제약 사항을 많이 주었다. 발화시에도 표준발음에 가깝도록 유도하여 대화체의 특성을 많이 포함하지 않는다.

---

1) "kind_a", "going_to"와 같이 발음변이가 심한 단어 열을 연결하여 사전의 표제어로 사용

음성데이터는 한 사람씩 번갈아가며 말한 것을 하나의 발화 (utterance)로 하여 하나의 파일로 분할하였다. 한 발화는 "예", "네"와 같이 짧은 문장도 존재하고, 여러 문장이 하나의 발화를 구성하기도 한다.

사람이 직접 음성을 듣고 입술소리, 찹소리 등의 잡음, "아", "어" 등의 간투어, 발음변이 (표준전사와 틀린 발음), 수정 또는 잘못된 발화 등을 전사하였다 (전사 예는 부록 A 참조). "/"를 기준으로 왼쪽의 텍스트는 실제 음성을 그대로 받아쓴 것이고, 오른쪽의 텍스트는 대화분석이나 언어모델 생성을 위해 문어체의 올바른 형태로 고쳐 썼다. 예를 들어 "어트케/어떻게"에서 실제 음성은 "어트케"라고 발화했지만 언어모델 생성을 위해서는 "어떻게"를 이용하기 위함이다. "어/"는 간투어의 전사 예로, 실제 음성은 존재하지만 대화분석에는 사용되지 않음을 나타낸다. 일반적으로 나타나는 모든 대화현상에 대해서 이와 같이 전사하였다. 단 문법에 틀리게 말한 것은 발화한 그대로 기록하였다.

표 1은 두 음성 데이터 베이스의 전사된 텍스트를 대상으로 형태소 분석을 수행한 후의 전체 크기이다. 잡음을 제외한 간투어 등을 모두 포함하였다. 유일 (unique) 형태소 수는 TP#1이 1,994이고 TP#2는 1,174로, 두 데이터 베이스 모두 사전 크기가 크지는 않지만 TP#1이 TP#2보다는 많은 어휘의 사용을 보여준다. 발화의 길이에 있어서도 TP#2는 평균 6.6어절, 12.6형태소이고 가장 긴 발화가 59어절, 108형태소인 반면, TP#1의 발화는 평균 10.5어절, 17형태소로 이루어져 있고 가장 긴 발화는 170어절, 320형태소로 매우 길고 한 발화가 여러 문장으로 이루어진 발화가 많아서 인식을 더 어렵게 한다.

생성된 음성 데이터 베이스의 바르게 말한 정도를 평가하기 위하여 코퍼스의 특성을 잘 설명할 수 있는 지수함수, $C^*b^L$를 사용하였다[7,8]. 여기서 $L$은 "efficient" (excluding edited) 단어들로 이루어진 문장 길이, $b$는 바르게 말한 정도 ($1-b$는 디스플루언시의 전체 비율)를 의미한다. $C$는 상수로 y축과의 교점을 나타내고, 문장 길이 $L$을 이용하여 $b$를 추정한다. b값은 낭독체 데이터 베이스인 ATIS (Air Travel Information System)가 0.9922이고 대화체인 Switchboard는 0.9447로 낭독체일수록 $b$가 크고, 디스플루언시를 많이 포함한 대화체일수록 $b$값이 작은 특징이 있다.

본 실험에서는 데이터 베이스의 전사가 정확하지 않음을 감안한 대략적인 수치로, TP#1의 $b$는 약 0.95이고 TP#2는 0.985이다. $b$ 값을 비교해 보면, TP#1이 TP#2보다 디스플루언시를 더 많이 포함하고 있고, 디스플루언시의 정도도 낭독체보다는 자연스러운 발화에 더 가까운 것을 알 수 있다.

디스플루언시의 위치 정보에 대해서도 위치정보를 문장의 앞과 중간으로 분류했을 때 TP#1은 73%, TP#2는 87%가 문장의 앞부분에 나타나서 [7,8]의 연구 결과와 일치함을 보여준다.

표 1. 데이터베이스의 크기
Table 1. Total size of databases.

| | 발화 수 | 어절 수 | 형태소 수 |
|---|---|---|---|
| TP#1 | 6,006 | 62,946 | 103,406 |
| TP#2 | 5,491 | 36,084 | 69,421 |

## III. 대화 특성의 분류

자연어처리 측면에서 대화체 문장의 주요 특징들로 음운축약 또는 탈락현상, 양성음의 음성음 발화, 어미 '요'의 출현, 띄어쓰기 오류, 격조사와 보조사의 결합, 조사 생략, 존칭어 사용, 문장성분의 생략, 간투어 등을 들 수 있다[1]. 음성 인식을 위해서는 이러한 특징들을 모두 고려할 필요는 없으며, 단지 기존의 낭독체 인식과 비교되는 특징들 중에서 인식 성능을 저하시키는 요소를 찾아내고, 대화체 인식을 위해서 낭독체와 다른 대화 현상들을 어떻게 모델링할 것인가에 대한 고려가 필요하다.

낭독체와 다른 대화체 문장의 특징들 중 텍스트 상에 나타나지 않는 대화 현상으로 빈번한 묵음 구간을 들 수 있고, 대화체 문장에 나타나는 대화 현상들을 음성인식

표 2. 분류 I: 대화 특성의 분류
Table 2. Classification I: Korean spontaneous speech characteristics.

| 분류 | | 예 |
|---|---|---|
| DFs | 잡음 (Noise) | N/ is/ 예 (예정입 is/ 니까)/예정입니까 |
| | 간투어 (Filled pause) | 예/ 저/ 어/ 어/ 기차가<br>예/ 혹시 뭐/ 예약 |
| | 반복/수정발화 (Repeat/repair) | 예약/ 예약하신다구요<br>연희장이/ 대연희장이<br>맞/ 맞습니까<br>예약하시/ 예약하셨습니다<br>(호텔 이/ 에는)/호텔에는<br>(예약하 하시면)/예약하시면 |
| 발음변이* (표준전사와 틀린 발음) | | 했구여/했고요<br>알겠슴다/알겠습니다<br>그르구/그리고<br>어트케/어떻게<br>그러믄/그러면 |

* 발음변이는 표 6과 같이 다시 세분화 함.

표 3. 분류 I: 대화 특성의 분포
Table 3. Distribution of classification I.

| | DFs | 발음변이 | 합 계 |
|---|---|---|---|
| TP#1 | 7,022 (11.2%) | 4,400 (7%) | 11,422 (18.2%) |
| TP#2 | 1,712 (4.7%) | 1,015 (3%) | 2,727 (7.7%) |

측면에서 표 2, 표 5와 같이 분류하였다.

낭독체와 비교되는 대화체의 특징을 분류 I과 같이 디스플루언시 (DFs)와 발음변이의 두 가지로 분류하였다. 디스플루언시는 낭독체에서는 전혀 나타나지 않는 현상으로 추가 고려가 필요한 부분이다. [7,8]에서는 디스플루언시를 간투어, 반복, 대치, 삽입, 삭제, 음성 오류로 세분화하였지만, 본 논문에서는 잡음, 간투어, 반복/수정 발화의 세 부분으로 간략하게 분류하였다. 발음변이는 표준전사와 다른 발음이라고 명시한 것처럼 텍스트 상으로는 낭독체와 같지만 실제 음성은 매우 다른 발음 현상을 보인다.

표 3은 데이터베이스 TP#1, TP#2에서의 대화 현상 각각의 분포를 보여준다. DFs는 잡음을 제외한 수치임에도 전체 어절의 11.2%로 매우 많은 부분을 차지한다. TP#1의 대화 현상은 전체 어절의 18.2%로 TP#2보다 많은 대화현상을 포함하는 것을 볼 수 있다.

## 3.1. 묵음 구간

전사된 텍스트만을 대상으로 대화 특성을 분류하였으므로 묵음 구간은 분류 I에서 제외하였다.

미리 준비된 문장을 발화하는 낭독체와 달리 한 발화 안에서 묵음 구간이 빈번하게 나타나고, 묵음 구간의 길이 또한 길어지는 현상을 보였다. 낭독체의 음향모델 학습시에는 묵음 구간을 크게 고려하지 않아도 상관없었지만, 대화체 음성 인식을 위한 음향모델 학습과 언어모델에서는 묵음을 고려하지 않았을 때 많은 삽입오류를 유발하였다.

대용량의 음향모델 학습은 비지도 학습 (unsupervised training)을 수행하므로 묵음 구간을 학습하도록 정보를 주는 것이 필요하다. 그러나 묵음 구간의 위치를 파악하기 위해서는 사람이 직접 음성을 듣고 레이블링을 수행해야 하는 어려움이 있다. 이를 해결하기 위하여 학습의 폴스 얼라인먼트 (force alignment) 단계에서 묵음 구간을 찾을 수 있도록 모든 발음열의 끝에 짧은 묵음과 묵음 모두를 첨부하도록 발음 사전을 구성하였다. 이렇게 학습한 음향 모델을 사용하여 인식 실험을 수행하였을 때 형태소 에러율이 1% (absolute) 감소하였다.

표 4. 잡음의 분포
Table 4. Distribution of noises.

| | TP#1 | TP#2 |
|---|---|---|
| 입술소리(ls) | 2,210 (19.9%) | 3,145 (40.5%) |
| 숨소리(h) | 6,964 (62.8%) | 1,815 (23.4%) |
| 기타 사람잡음 (other) | 767 (6.9%) | 448 (5.8%) |
| 주변잡음(N) | 1,142 (10.3%) | 2,356 (30.3%) |
| 합 계 | 11,083 | 7,764 |

* ( )안의 영문은 전사 기호를 명시하였음.

## 3.2. 잡음

입술소리 (입술이 붙었다 떨어지거나, 혀가 입천장에서 떨어지는 소리), 숨소리, 웃음소리, 기침소리 등의 잡음은 자연스러운 발화에서는 위치에 상관없이 빈번히 나타나는 현상으로 음성인식 단계에서 많은 삽입 오류를 유발시키는 요소이다.

표 4는 잡음을 각각의 특성에 따라 분류하고, 음성 데이터 베이스를 전사한 후의 분포를 나타낸다. 음성과 함께 나타나는 잡음은 제외시켰으며 웃음소리, 기침소리 등은 매우 적게 나타나므로 "기타 사람잡음"에 포함시켰다. 한 발화당 평균 잡음 수는 TP#1이 1.8, TP#2가 1.4이고 "예", "네"와 같이 한 어절로 구성된 발화와 전사 오류를 고려하면 한 발화당 평균 2개 이상의 잡음이 있다고 볼 수 있다.

잡음이 음성과 함께 나타나는 경우에도 잡음기호는 함께 사용하여 "여정입 ls/ 니까", "매 N/ 표 N/ 소에서"와 같이 표기하였으나, 어절과 어절 사이에 나타나는 잡음과 함께 처리할 수 없으므로 구별이 필요하다. 이러한 잡음은 음향 모델의 질과 인식 성능을 떨어뜨리는 요소로 작용하므로 견고한 음향 모델을 사용하는 것이 필요하다.

본 논문에서는 어절과 어절 사이에서 나타나는 잡음만을 대상으로 HMM 모델링하였다[2]. 각 잡음들의 특성에 따른 세분화된 HMM 모델링이 필요하지만, 본 논문에서는 한 개의 HMM 모델을 사용하였다. 이 모델을 사용하여 형태소 에러율을 1.08% (absolute) 감소시켰다.

## 3.3. 간투어

낭독체 음성과 비교할 때 대화체에서 가장 빈번히 나타나는 현상으로 음성 인식 성능을 떨어뜨리는 요소 중의 하나이다[8]. 이전의 여러 연구에서 간투어는 입술소리, 숨소리 등과 같은 비언어적인 요소로 분류되기도 했으나, 언어적인 경계 정보를 가지고 있어서 발화 위치에 따라 다음 단어에 대한 예측 기능을 가지고 있다. 간투어들은 문장의 시작 부분에서 가장 많이 나타나기 때문에 음성을

표 5. 간투어의 분포
Table 5. Distribution of filled pauses.

| | TP#1 | TP#2 |
|---|---|---|
| DFs | 7,022어절 (11.2%) | 1,686 어절 (4.7%) |
| 간투어 | Top10: DFs의 80% | Top10: DFs의 90% |
| 분포 | 예(29.4%), 어(26.4%), 아(9.2%), 음(3.8%), 그(3.1%), 좀(2.9%), 네(2.1%) | 예(61%), 네(11%), 어(7%), 아(6%) |

* 잡음은 포함시키지 않았음.

문장 단위로 분할할 때 간투어 정보를 효과적으로 사용할 수 있다[4,12].

표 5는 여행계획 데이터 베이스에 나타난 간투어의 분포이다. 약 10여 개의 간투어가 여행계획 음성 데이터에 빈번히 나타났으며, 두 사람이 문의하고 대답하는 형식의 대화이고 발화 자체가 길기 때문에 응답성의 간투어 "예"가 가장 많이 나타났다.

두 데이터 베이스의 생성시 자유도에 따라 간투어의 분포와 현상이 달라지는 것을 확인할 수 있다. TP#2는 응답성의 "예", "네"가 70% 이상을 차지하는 반면, TP#1에서는 "어", "음"과 같이 발화 도중에 생각하거나 "아"와 같이 발화의 수정을 위한 간투어들이 많이 나타나는 현상을 보였다. 이는 TP#1의 발화가 더 자연스러운 발화임을 보여준다.

음향모델의 경우, 간투어를 잡음과 함께 하나의 잡음 모델로 생성하는 예도 있으나[3], 간투어는 잡음과는 다른 특성을 가지므로 본 논문에서는 잡음과 분리하여 모델링하였다. 또한 머뭇거리거나 생각할 때 나타나는 "어", "음" 등과 같은 간투어의 발화구간이 어절 안에서 나타나는 같은 음절보다 길게 나타나는 특성을 보이므로, 간투어에 따라 다른 PLU (Phone-Like Unit)로 모델링하는 것이 효과적이라 여겨진다. 본 논문에서는 빈도수가 높은 "예", "어", "아"에 대해서 다른 PLU를 사용하여 모델링하였다. 실험 결과로부터 "어"는 인식 성능은 개선하였다. 그러나 "예"와 "네" 같은 간투어는 대답과 간투어를 구분하기 어렵기 때문에 언어모델에서는 간투어를 특별히 구별하지 않고 모델링하였다.

이외에도 "예_어", "예_예" 등을 하나의 어휘모델로 할 것인지 분리할 것인지의 여부와 평가시에 간투어를 인식률에 포함시킬지 등이 여전히 문제로 남는다. 본 논문에서는 간투어도 사전의 한 단어로 간주하여 실험하였다.

### 3.4. 반복/수정 발화

발화 도중 같은 단어 또는 어절을 반복해서 말하거나 다른 단어나 어절로 수정해서 말하는 현상을 말한다. [4]의 연구에서 반복된 단어들의 정보를 언어모델 생성에 이용하였으나 많이 나타나는 현상이 아니므로 성능의 향상에 큰 공헌을 하지는 않았다.

한국어에서는 온전한 한 어절이 반복되는 현상은 매우 드물고, 주로 어절의 일부분만 발화하거나 어미나 조사가 변형된 형태로 수정되는 것이 일반적이다. 이때 단어나 어절을 온전하게 다 발화하지 않고 발화 도중에 중단하여 어절의 일부만 발화하는 현상을 단어의 조각화(word fragmentation)라고 한다. "맞/ 맞습니까", "시/ 신혼여행"과 같은 예에서 "맞", "시"가 조각난 단어들인데 이들을 어떻게 분류하여 처리할 것인가가 논의의 대상이 된다. 대체로 이들 조각난 단어들은 사전에 포함시키지 않는 것이 일반적이다.

한국어에 있어서 인식 단위를 형태소로 하는 경우 이런 조각난 단어의 처리에 있어서, "예약하셨/ 예약하셨는데요"라고 발화했을 경우 "예약하셨"을 형태소 분석하면 "예약+하+셨"으로 분할 가능하다. 이런 경우 형태소 분석에 상관없이 "예약하셨"을 무시할 것인지, 아니면 형태소 정보를 사용할 것인지는 더 많은 연구를 필요로 한다. 본 논문에서는 조각난 단어도 형태소 분석을 수행하는 것을 원칙으로 하였다.

이러한 반복/수정 발화 현상이 TP#1에서는 총 690회(전체 어절의 1%)가 나타났으며 한 어절 이하의 반복이 565회로 대부분을 차지하고 있으며, TP#2에서는 170회로 적은 부분을 차지하였다. 반복/수정 발화현상은 많은 데이터베이스의 분석을 통해 이루어져야 하지만, 많이 나타나는 현상이 아니므로 본 연구에서는 논외로 하였다.

### 3.5. 발음변이

표준전사 또는 문어체와 다르게 발화하는 현상을 모두 발음변이로 포함시켰다. 전체 어절의 7%에 해당하며 대화체의 인식 성능을 떨어뜨리는 요소들 중의 하나이다.

여행계획 데이터 베이스, TP#2는 발화시 제약을 많이 주었기 때문에 올바르게 말하는 경향이 있어서 대화체의 특징이 잘 나타나지 않으므로 발음변이는 TP#1만을 대상으로 하였다. 표 6의 분류 II는 TP#1에 나타나는 발음변이 현상을 세분화한 것으로, 음성데이타를 전사한 문장에 나타난 현상들이 공통된 특징을 갖도록 분류하였다. TP#1에 대한 분류 II의 분포를 표 7에 나타내었다.

가장 많은 부분을 차지하는 양성음의 음성음 발화는 사람들이 자연스럽게 말할 때 흔하게 나타나는 현상으로 71.8%에 달하지만, 제약을 많이 준 TP#2의 경우는 38%로 TP#1과 비교할 때 상대적으로 적게 나타났다. 형태소의

표 6. 분류 II: 발음변이의 분류
Table 6. Classification II: Phonological variations.

| 분 류 | 예 |
|---|---|
| 양성음의 음성음 발화 | ~구여, ~구요, ~고여<br>데여, 알구, 하구 |
| 음운축약 / 탈락 | 했습다, 주십쇼, 일임다<br>예맬/예매를<br>오심/오시면<br>김철숩니다/김철수입니다<br>까집니다/까지입니다 |
| 패턴화된 발음변이 | 그르구/그리고<br>어트케/어떻게<br>그러믄/그러면 |
| 발화 오류 | 오우/오후<br>예앙하고/예약하고 |

표 7. 분류 II: 발음변이의 분포
Table 7. Distribution of classification II.

| 분류 | 빈도수 | 백분율 |
|---|---|---|
| 양성음의 음성음 발화 | 3,164 | 71.8% |
| 음운축약 / 탈락 | 317 | 7.2% |
| 패턴화된 발음변이 | 518 | 11.8% |
| 발화 오류 | 405 | 9.2% |
| 합 계 | 4,404 | 100% |

어미 부분에서 이 현상이 나타나므로 인식 단위를 형태소로 하면, 대화체 인식을 위해서 단순히 발음사전에 변형된 어미를 표준 어미의 다중 발음의 하나로 추가하여 인식 성능을 개선할 수 있다. 예를 들면, "구여", "구요", "고요"의 표준 전사는 "고요"이므로 발음사전의 "고요"의 다른 다중 발음으로 "구여", "구요", "고요"의 발음을 추가하였다. 본 실험에서는 단지 37개의 발음열을 추가하였을 뿐이다.

음운축약은 어미나 조사의 변형을 가져오는데, 이 중 "~습다/~습니다", "~십쇼/~십시오", "~ㅁ다/~ㅂ니다"와 같이 종결어미가 가장 많은 부분을 차지한다. 탈락현상은 서술격 조사 "이"의 탈락이 대부분이다. 대화체 인식 실험을 위해서 변형된 종결어미를 표준 전사의 다중 발음으로 발음사전에 추가하였다. 탈락현상은 인식해야 할 형태소가 없어지기 때문에 언어모델에 반영해야 하므로 전사된 텍스트를 탈락된 형태로 수정하였다. 수정된 전사 텍스트로부터 얻은 언어모델을 기본 인식 실험에 이용하여 탈락현상은 베이스 인식률에 포함시켰다.

패턴화된 발음변이는 위 두 현상과 발화 오류를 제외한 것을 모두 포함하는데, 대체로 발화 경향이 일정하여 패턴화 가능한 부분이므로 발생빈도가 높은 것만을 대상으로 발음 사전에 추가하였다. "어트케"의 발음열을 "어떻게"의 다중 발음열로 추가하는 등, 총 13개 형태소에 대해 24개의 다중발음열을 추가하였다.

발화 오류도 9.2%로 대화체 음성 인식의 성능을 저하시키는 요인이 되고 있지만, 의미정보와 구문정보 등의 자연어처리 레벨의 정보를 이용하지 않고 음성 인식 단계에서 해결할 수 없는 문제이므로 논외로 하였다.

### 3.6. 그 밖의 특징

위의 분류 외에 TP#1의 경우, "하구요"에서처럼 어미 "요"가 매우 빈번히 나타나고, "과", "와"보다 "하고"의 사용이 더 빈번하였다. 이런 현상은 낭독체에서는 나타나지 않는 현상들이다.

## IV. 대화체 연속음성 인식

대화체 연속음성 인식을 위해 HTK (Hidden Markov Model Toolkit)[11]를 이용하여 인식 실험을 수행하였다. 본 실험에서 사용한 음향모델은 CHMM (Continuous Hidden Markov Model)을 기반으로 하였으며 6개의 가우시안 믹스츄어를 사용하였다. 잡음 처리를 위하여 하나의 잡음 모델을 사용하였다.

학습 및 테스트에 사용된 음성 데이터 베이스는 표 8과 같다. 두 데이터 베이스 모두 화자 독립이 되도록 4조, 8화자의 모든 음성을 테스트에 이용하였고, 테스트에 이용되지 않는 나머지를 학습에 이용하였다.

발음사전은 각 여행계획 데이터 베이스로부터 미등록어가 없도록 다중발음 형태소 사전을 생성하였으며, 간투어와 조각난 단어를 모두 포함하였다.

좀더 안정적인 음향모델 생성을 위하여 대화체 음성 데이터 베이스만으로 학습하지 않고 낭독체 데이터 베이스를 함께 사용하였다. 낭독체 음성 데이터 베이스 18000 문장으로 학습하여 생성된 음향 모델을 대화체 학습 데이터 베이스로 적응 훈련하였다. 그러나 적응 훈련만으로는 발화 길이와 잡음, 간투어를 반영하지 못하므로 실험에서는 묵음모델, 잡음모델, 간투어 모델은 대화체 음성 데이터 베이스로 학습한 음향 모델을 사용하였다. 또한

표 8. 학습 및 테스트용 음성 데이터 베이스 크기
Table 8. DB sizes for train and test.

| | 학습 | 테스트 |
|---|---|---|
| TP#1<br>(약 10시간) | 21조, 84 대화<br>5,021 발화 | 4조, 16 대화<br>834 발화 |
| TP#2<br>(약 7.5시간) | 21조, 105 대화<br>4,621 발화 | 4조, 20 대화<br>870 발화 |

TP#1과 TP#2는 각기 다른 학습 모델을 생성하였다.

언어모델은 학습데이터로부터 백오프 바이그램 (backoff bigram)을 생성하였으며 TP#1은 262, 87.13%, TP#2는 98.86, 91.46%의 언어모델 혼잡도와 바이그램 히트율 (bigram hit ratio)을 얻어 두 데이터 베이스 모두 언어모델 생성을 위한 텍스트 데이터가 매우 부족함을 알 수 있다. 데이터 부족을 줄이기 위하여 두 데이터 베이스의 학습 데이터를 합쳐서 언어모델을 생성했을 때 언어모델 혼잡도가 174, 49.6으로 감소하지만, 두 데이터 베이스의 이질성이 크기 때문에 인식 성능에는 영향을 미치지 못하였으므로 본 실험에서는 각각의 학습데이터로 생성한 언어모델을 사용하여 인식 실험을 수행하였다.

대화체 음성 데이터 베이스의 자유도에 따른 인식 성능을 비교하기 위하여 TP#1과 TP#2의 인식 실험을 수행하였다. 인식 실험은 대화 현상을 반영하지 않은 모델을 사용하였다. 인식 결과는 표 9과 같다. 두 데이터 베이스 모두 대화체의 특성을 가지므로 낭독체 데이터 베이스보다 인식 성능이 떨어지지만, 대화체의 특성을 많이 포함하는 TP#1의 형태소 에러율이 TP#2와 비교할 때 큰 것을 볼 때 대화체 음성의 인식이 쉽지 않은 것을 알 수 있다.

표 10은 대화체의 특성을 반영했을 때의 인식 성능을 TP#1을 대상으로 실험한 결과이다. Sil은 긴 묵음 구간을 학습하도록 음향모델을 생성한 것이고, GBM은 잡음 모델을 사용했을 때이다. 새로운 묵음 구간 학습과 잡음 모델을 적용하여 2.08%의 형태소 에러율을 감소시켰다.

FP는 간투어 모델을 사용했을 때이고, 0.73%의 형태소 에러율을 감소시켰다. 출현 빈도가 높은 간투어 "예", "어", "아"를 다른 PLU로 모델링하여 인식 실험을 수행하였다. 인식 실험을 수행한 결과 간투어 "어"만 형태소 에러율을 감소시켰는데, 이는 어절 중간에 나타나는 "어"는 짧게 발화되지만 간투어 "어"는 주로 머뭇거릴 때 사용되므로 길게 발화되는 특성을 갖기 때문이다. "예"는 어절 안에서 잘 나타나지 않고, "아"는 어절 안에서와 간투어 모두 발화 길이가 비슷하여 간투어 모델이 성능 개선에 도움을 주지 못하였다.

발음변이 현상들을 양성음의 음성음화 (P1), 축약과 패턴화된 발음변이 현상 (P2)으로 나누어 다중발음사전에 반영한 후에 인식 실험을 수행한 결과에서는 0.92%의 형태소 에러율 감소를 얻었다. P1의 형태소 에러율 감소가 P2보다 더 큰 이유는 양성음의 음성음 발화 현상의 출현 빈도가 매우 높기 때문으로 분석된다.

대화 현상의 기초적인 분석을 통해서 절대치로 3.73%, 상대치로 12%의 형태소 에러율을 감소시켰다. 묵음과 잡음 모델의 추가로 가장 많은 형태소 에러율 감소를 얻었는데 이는 대화체 음성이 기본적으로 낭독체 음성과 매우 다른 특성을 가지고 있음을 단적으로 보여주는 예라고 하겠다.

표 9. 자유도에 따른 형태소 에러율
Table 9. MER (Morpheme Error Rate) vs. speaking styles.

| | TP#1 | TP#2 |
|---|---|---|
| MER(%) | 31.65 | 17.34 |

표 10. 대화특성 반영에 따른 형태소 에러율 (%)
Table 10. MER (%) vs. Korean spontaneous speech characteristics.

| 분류 | MER | 감소율 |
|---|---|---|
| Baseline | 31.65 | |
| Base+Sil | 30.65 | 2.08 |
| Base+Sil+GBM | 29.57 | |
| Base+Sil+GBM+FP | 28.84 | 0.73 |
| Base+Sil+GBM+P1 | 28.99 | 0.92 |
| Base+Sil+GBM+P2 | 29.33 | |
| Base+Sil+GBM+P1+P2 | 28.73 | |
| Base+Sil+GBM+FP+P1 P2 | 27.92 | |
| 전체 감소율 (절대치) | | 3.73 % |

Sil: 묵음 구간을 반영한 모델
GBM: 잡음 모델
FP: 간투어 모델
P1: 양성음의 음성음화 반영
P2: 축약, 패턴화된 발음변이 현상

## V. 결론 및 향후과제

본 논문에서는 대화체 연속음성의 특성을 분석하고 낭독체 인식 기술을 기반으로 대화 특성을 반영하여 대화체 연속음성 인식을 위한 기본 인식 실험을 수행하였다.

대화체 연속 음성 데이터베이스의 분석을 통해 대화체 음성은 낭독체 음성과 달리 묵음 구간이 길고 많은 잡음을 포함할 뿐 아니라, 발화 사이사이에 존재하는 간투어, 발음변이 현상들이 빈번히 나타나고 있음을 확인하였다.

인식 실험을 통해 발화 스타일이 대화체에 가까울수록 디스플루언시와 발음변이가 많아지고 인식 성능도 저하되었으며, 각 대화현상을 반영하여 12% (relative)의 형태소 에러율을 감소시켰다.

본 논문에서는 대화체 연속음성에 대한 기초적인 연구이므로 대화 현상에 대한 깊이 있는 연구를 필요로 한다.

음향모델과 어휘모델, 언어모델에서 데이터 부족 문제, 인식 실험에 간투어를 포함시킬지의 여부 등에 대한 연구가 필요하지만, 대화체 연속음성 인식을 위해 대화 음성을 어떤 정보를 갖도록 어떻게 전사할 것인가도 무엇보다 중요한 과제라 하겠다.

## 감사의 글

본 연구는 과학기술부의 특정연구개발사업 (M1010700015) 지원으로 수행되었으며, 실험에 사용된 한국전자통신연구원의 대화체 연속음성 데이터 베이스와 삼성종합기술원의 낭독체 연속음성 데이터 베이스의 사용허가에 감사드립니다.

## 참 고 문 헌

1. 왕지현, 서영훈, "개념 및 구문정보를 이용한 한국어 대화체 분석 시스템," 제9회 한글 및 한국어 정보처리 학술발표 논문집, 341-346, 1997.
2. 이경님, 정민화, "한국어 낭독체 인식의 발성 잡음처리를 위한 Human Garbage 모델링," 한국음향학회 하계학술대회논문집, 323-326, 2001
3. 이황섭, 박준, 권오욱, "한국어 대화체 인식 시스템의 구현," 제13회 음성통신 및 신호처리 워크샵, 13 (1), 145-148, 1996.
4. A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," *Proc. of International Conference on Acoustics, Speech, and Signal*, vol. 1, 405-408, 1996.
5. A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Graoble, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng and J. Zheng, "The SRI march 2000 HUB-5 conversational speech transcription system," *Proc. of NIST Speech Transcription Workshop*, 2000.
6. B. Byrne, M. Finke, S. Khudanpur, J. McDounugh, H. Nock, M. Riley, M. Saraclar, C. Wooters and G. Zavaliagkos, "Pronunciation modeling using a Hand-labelled corpus for conversational speech recognition," *Proc. of International Conference on Acoustics, Speech, and Signal*, vol. 1, 313-316, 1998.
7. E. Shriberg, "Preliminaries to a Theory of Speech Disfluencies," Ph. D. thesis, University of California at Berkeley, 1994.
8. E. Shriberg, "Disfluencies in switchboard," *Proc. of International Conference on Spoken Language Processing*, vol. 3, 1301-1305, 1996.
9. E. Shriberg and A. Stolcke, "Word predictability after hesitations: A corpus-based study," *Proc. of International Conference on Spoken Language Processing*, vol. 3, 691-695, 1996.
10. J. J. Godfrey, E. C. Holliman and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," *Proc. of International Conference on Acoustics, Speech, and Signal*, 1992.
11. HTK Hidden Markov Model Toolkit, Version 2.2, http://htk.eng.cam.ac.uk/index.shtml
12. M. H. Siu and M. Ostendorf, "Modeling disfluencies in conversational speech," *Proc. of International Conference on Spoken Language Processing*, vol. 1, 621-625, 1996.
13. M. Finke and A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," *Proc. of EUROSPEECH*, vol. 5, 2359-2382, 1997.
14. M. Weintraub, K. Taussig, K. H. and A. Snodgrass, "Effect of speaking style on LVCSR performance," *Proc. of International Conference on Spoken Language Processing*, vol. 3, 1036-1039, 1996.
15. R. Rosenfeld, R. Agarwal, R. Iyer, L. Shriberg and D. Vergyri, "Error analysis and disfluencies modeling in the Switchboard domain," *JHU Summer Workshop*, 1995.

## A. 대화체 전사 예

갑: 예/ 안녕하십니까? h/ 월드 와이드 여행사 김철습니다/ 김철수입니다.
을: ls/ 아/ 예/ 제가 시월/10월 이/2 일부터 육/6 일까지 신혼여행을 가려/ 가려고 하는데요. 어/ h/ 항공편이나 뭐/ 호텔 예약 같은 걸 하고 싶습니다.
갑: h/ 예/ 그럼 우선 항공편부터 예약하시겠습니까?
을: 예. 그렇게 하죠. 저기/ h/ 제가 시월/10월 이/2 일 날 오후 다섯/5 시 경에 출발하려고 하는데요. h/
갑: 예/ 어/ 오후 다섯/5 시 경에 지금 어/ h/ 예약 가능한 항/ 항공편이요. 어/ h/ 아시아나 항공편어 지금 두/2 편 남아 있습니다. h/ 어/ 하나/1 가 오후 네/4 시 이십/20 분에 출발하고 h/ 하나/1 가 오후 다섯/5 시 이십/20 분에 출발합니다. 어/ 오후 네/4 시 오십/50 분 편도 있었는데 방금 매진 됐습니다. h/ 어/ 일/1 인당 요금이 오만/50000 구천/9000 원인데요 h/ 어떤 것으로 예약하시겠습니까? ls/
을: 예/ 다섯/5 시 이십/20 분에 비행기가 있다고 하셨죠?
갑: 예. 있습니다.
을: 예/ 그럼 그 걸로 해주십시오.
갑: 예/ 그/ 그렇게 하겠습니다. 어/ 매수는 어/ 두/2 매 되겠습니까?
을: 예/ 예

* TP#1의 전사 예
* 띄어쓰기 단위를 어절 단위로 제산하였다.
* "예"와 같이 짧은 발화뿐 아니라 5번째 발화처럼 매우 긴 발화도 존재하고, 고두 쉬지 않고 한번에 발화하였다.
* 짧은 발화에서도 대화 현상이 빈번하게 나타난다.

## 저자 약력


● 박 영 희 (Young-Hee Park)

1994년 2월: 동국대학교 전자계산학과 (공학사)
1999년 2월: 서강대학교 컴퓨터학과 대학원 (공학석사)
1999년 3월~현재: 서강대학교 컴퓨터학과 대학원 박사과정
※ 주관심분야: 음성인식 및 언어처리, 언어모델

● 정 민 화 (Minhwa Chung)

1984년 2월: 서울대학교 제어계측학과 (공학사)
1988년 5월: Univ. of Southern California 전기공학과 (M.S.)
1993년 8월: Univ. of Southern California 전기공학과 (Ph.D.)
1995년 9월~현재: 서강대학교 컴퓨터학과 부교수
※ 주관심분야: 음성언어처리, 자연어처리