비핵심어 모델의 가중치 기반 핵심어 검출 성능 향상에 관한 연구

김 학 진† · 김 순 협††

요 약

본 논문에서는 핵심어 검출기의 성능 향상을 위해 가베지 클라스 클러스터링과 함께 필러 모델에 가중치를 부여하는 방안 및 태스크 도메인 이용자들의 발화 음성의 성향 분석을 통해 핵심어 천이 확률을 계산하여 핵심어 검출기반 대화 음성처리 시스템의 처리 시간 단축 방안을 제안한다. 제안한 방법은 음성학적으로 유사한 음소끼리 묶어서 사용함으로써 하나의 음소는 잘 표현하지 못하지만 비슷한 음소 그룹의 표현에는 유용한 방법으로 본 논문에서는 한국어 형태론과 태스크 도메인으로 선정한 증권거래 대화음성처리 시스템에서 활용되는 발화 문장을 분석하여 5 음소군을 제시한다. 또한 이들 음소군에 태스크 종속적인 필러 모델 가중치를 부여하며, 두 번째로는 시스템의 처리시간 단축을 위해 연속 발화 문장 속에 포함되어 있는 핵심어 천이 확률을 계산하여 시스템에 적용 실험한다. 제안한 시스템의 성능 평가를 위해 태스크 도메인에 활용되는 4,970 문장의 코퍼스를 구축하고, 이용자 중 20대~30대 5명이 발성하게 하여 실험한 결과, 제안한 5 음소군에 가중치를 부여한 방법의 FOM은 87.5%로 Yapanel[1]의 7음소군 85.5%보다 우수한 성능을 보였으나, LVCSR의 89.8%보다는 약간 뒤지는 성능을 확인하였다. 계산시간에 있어서도 0.70초로 7음소군의 0.72초보다 우수한 성능을 보였다. 핵심어 천이 확률 분석을 통한 인식 시간 단축 실험에서는 천이 확률을 적용했을 때 약 0.04초~0.07초의 처리 시간을 단축하는 것을 확인하였다.

A Study of Keyword Spotting System Based on the Weight of Non-Keyword Model

Hack-Jin Kim† · Soon-Hyub Kim††

ABSTRACT

This paper presents a method of giving weights to garbage class clustering and Filler model to improve performance of keyword spotting system and a time-saving method of dialogue speech processing system for keyword spotting by calculating keyword transition probability through speech analysis of task domain users. The point of the method is grouping phonemes with phonetic similarities, which is effective in sensing similar phoneme groups rather than individual phonemes, and the paper aims to suggest five groups of phonemes obtained from the analysis of speech sentences in use in Korean morphology and in stock-trading speech processing system. Besides, task-subject Filler model weights are added to the phoneme groups, and keyword transition probability included in consecutive speech sentences is calculated and applied to the system in order to save time for system processing. To evaluate performance of the suggested system, corpus of 4,970 sentences was built to be used in task domains and a test was conducted with subjects of five people in their twenties and thirties. As a result, FOM with the weights on proposed five phoneme groups accounts for 85%, which has better performance than seven phoneme groups of Yapanel [1] with 85.5% and a little bit poorer performance than LVCSR with 89.8%. Even in calculation time, FOM reaches 0.70 seconds than 0.72 of seven phoneme groups. Lastly, it is also confirmed in a time-saving test that time is saved by 0.04 to 0.07 seconds when keyword transition probability is applied.

키워드 : 핵심어 검출(Keyword Spotting), 비핵심어 모델(Non-keyword Model), 필러 모델(Filler Model), 가중치(Weight)

1. 서 론

최근에 연구되고 있는 핵심어 검출기는 주로 HMM을 기반으로 개발되고 있다[2-5]. HMM 기반의 핵심어 검출기에서 핵심어의 수, 길이, 발음, 유사 핵심어의 존재 여부 등은 핵심어 검출 성능과 인식 속도 면에서 많은 영향을 미치고 있으며, 특히 정교한 비핵심어 모델링을 통해 검출기의 성능 향상을 가져올 수 있다[6]. 일반적으로 핵심어 모델과 필러 모델로 구성된 대어휘 연속음성인식(Large Vocabulary Continuous Speech Recognition : LVCSR) 시스템이 가장 좋은 성능을 보이고 있으나, 중요한 단점을 내포하고 있다[7]. 첫 번째로 계산시간이 과다하게 소요되고, 태스크 도메인에 대해 비독립적이며, 전체 어휘에 대한 지식과 훈련 데

※ 본 연구는 2002년도 광운대학교 교내학술연구비 지원에 의해 수행되었습니다.
† 정 회 원 : 명지전문대학 컴퓨터정보과 교수
†† 정 회 원 : 광운대학교 컴퓨터공학과 교수
논문접수 : 2003년 7월 10일, 심사완료 : 2003년 8월 8일

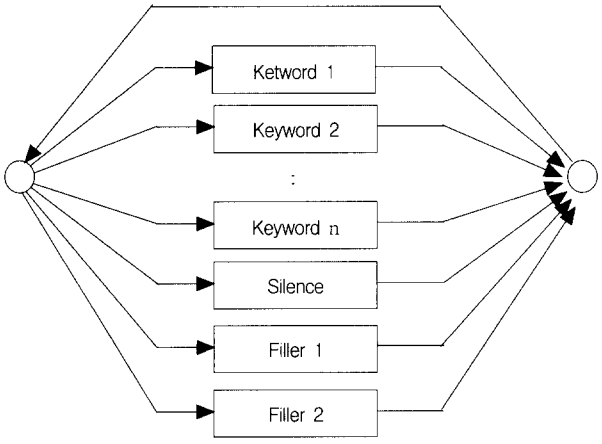
이터가 과다하게 필요한 문제점을 갖고 있다. 본 연구에서는 LVCSR과 같이 높은 성능을 갖으면서 이러한 단점을 극복하기 위한 방법론을 제시하였다.

먼저, 필러 모델의 구성을 위해 비핵심어 음소를 한국어 형태론에 의거한 7개 음소군으로 분류한 것과 이를 다시 태스크 도메인 분석을 통해 5개 음소군으로 분류하여 필러 모델로 채용하였다. 또한 핵심어와 비핵심어간의 로그 유사도 점수 차이를 제거하기 위하여 음소군 필러 모델에 가중치를 부여하는 방법을 여구하였다. 두 번째로는 태스크에 대한 이용자의 화행 분석을 통해 태스크 처리에 필요한 핵심어의 발화 순서를 미리 예측 가능하게 처리함으로써 대화 음성처리 시스템의 처리 시간을 단축하였다. 실험에서는 핵심어가 1개~4개까지 포함하고 있는 연속음성에 대한 실험으로 필러 모델이 1개, 5개, 7개를 포함하고 있는 인식 네트워크를 통해 시스템을 비교하고, 필러 모델이 5개, 7개인 네트워크에 각각 가중치를 부여한 실험 결과와 LVCSR를 비교하여 핵심어 검출 능력 및 실용 가능성을 제시하였다.

2. 관련 연구

2.1 HMM 기반 음성 인식

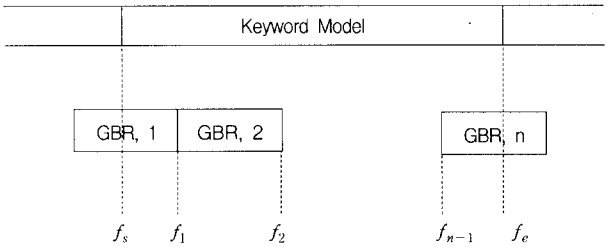
HMM 기반 음성인식기의 인식 수행은 일반적으로 Token Passing Paradigm[8]을 이용하며, 인식단계에서는 유한 상태 네트워크와 시간동기 비터비 빔 탐색 디코더(Viterbi beam search decorder)를 이용하고 있다. (그림 2.1)은 null grammar 핵심어 인식 네트워크를 나타내고 있으며, (그림 2.2)는 필러 모델만이 이용할 수 있는 2패스 과정을 보여주고 있다.



(그림 2.1) Null-grammar 핵심어 인식 네트워크의 예

(그림 2.2)에서 핵심어 인식 프레임 f_s 와 f_e 사이에 핵심어가 존재한다고 가정하고, 2 단계 인식과정을 통해 가베지 모델 점수를 구한다. 먼저, 각 프레임에 대한 핵심어 점수와가베지 점수를 계산하고, 이를 정규화를 통해 식 (2.1)와 같은 핵심어 스코어를 얻는다.

$$S_{KW} = \frac{M_{KW}}{f_e - f_s} \tag{2.1}$$



(그림 2.2) 2 단계 인식 과정

M_{KW}는 핵심어의 비터비 최대 로그 유사도 점수(Viterbi maximum log likelihood score)를 의미한다.

비슷한 방법으로 가베지 점수를 식 (2.2)와 같이 얻는다.

$$S_{GR} = \frac{M_{GR}}{f_e - f_s} \tag{2.2}$$

M_{GR}은 프레임 $f_s \sim f_e$간에 있는 모든 가베지 모델들의 최대 로그 유사도 점수의 합을 의미한다. 마지막 단계는 핵심어 결정을 위하여 이 두 점수를 결합시켜야 한다. 실제로, 그 두 점수를 결합시키는 몇몇 방법 중의 하나는 두 점수의 평균 차를 이용하는 것으로 이를 유사도 비 점수(Likelihood Ratio Scoring)라고 하며 식 (2.3)과 같이 표현한다.

$$S_{LR} = S_{KW} - S_{GR} \tag{2.3}$$

다른 방법은 Rose[9]에 의해서 제안된 것으로 프레임 $f_s < f < f_e$에서 로그 핵심어 확률과 로그 가베지 확률간의 최대 차로 표현된다. 즉, 핵심어 모델 점수 $L_{KW}(f)$와 가베지 모델 점수 $L_{GR}(f)$는 식 (2.4)와 같이 표현된다.

$$M_{LR} = \max_{f} \left(L_{KW}(f) - L_{GR}(f) \right) \tag{2.4}$$

$L_{KW}(f)$는 프레임 f_s로 시작하여 프레임 f까지의 비터비 최대 로그 유사도, $L_{GR}(f)$는 프레임 f_s로 시작하여 f_e로 끝나는 프레임의 비터비 최대 유사도의 합을 의미한다. 핵심어 점수와 가베지 점수간의 차이는 프레임과 프레임간의 차감을 통해 얻어질 수 있다. 다른 방법으로는 오류 검출과 핵심어 검출간의 차별성을 강화할 수 있는 방법으로 정규화를 통해 점수를 재계산하는 방법[10]으로 식 (2.5)에서 보여주고 있다.

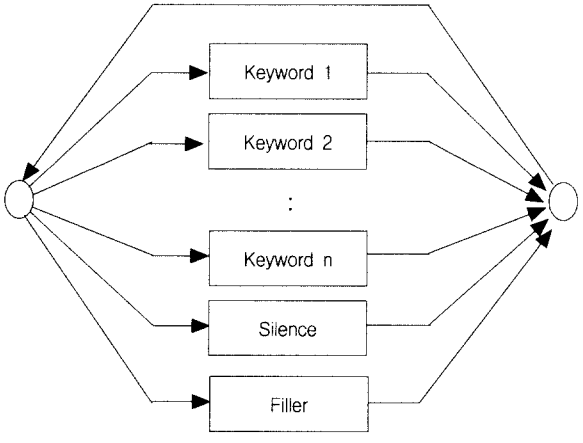
$$N_{LR} = \frac{S_{GR}}{S_{KW}} \times 100 \tag{2.5}$$

S_{GR}과 S_{KW}는 가베지 모델과 핵심어 모델의 비터비 로그 유사도 확률의 평균을 의미한다. HMM 기반 핵심어 인식

에서 2패스 방식을 개선한 방법으로 유사도비 점수(Likelihood Ratio Scoring(S_{LR}))를 이용하는 방법이 있다. 유사도 비 점수 방법은 2 패스 수행을 한번으로 줄일 수 있으나, 계산 시간은 증가한다. 단일 패스 인식 네트워크도 핵심어 모델, 가베지 모델, 묵음 모델 등을 병렬로 연결하여 구성한다. 핵심어들은 대부분 최대 점수를 갖으나, 가베지 패스에 약간의 가중치를 부여함으로써 이를 억제할 수 있다. 가중치는 핵심어 모델과 필러 모델 사이의 로그 유사도 점수 차이를 제거하기 위하여 활용한다[11]. 즉, S_{LR}에 대한 임계값을 설정하는 것보다는 필러 모델에 가중치를 더하여 제거하는 것이다.

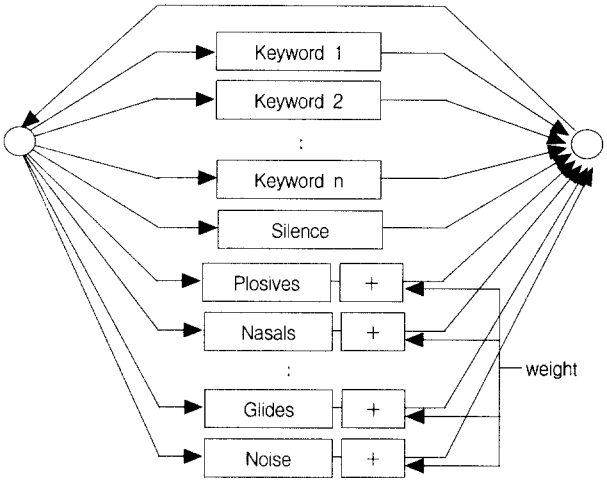
2.2 필러 모델

핵심어 검출은 일반적으로 핵심어 모델과 필러 모델을 연결하는 연결단어인식 알고리즘을 사용한다. 필러 모델의 구성은 비핵심어 각각을 모델링하는 방법[12]과 핵심어가 아닌 부분 전체를 모델링하는 방법[13, 14]으로 나누며, 이를 세부적으로는 단어 모델을 사용하는 방법과 부단어 단위를 사용하는 방법이 있다. 부단어 단위를 사용하는 방법으로는 트라이폰 모델 또는 모노폰 모델, 가베지 클라스 클러스터링, 온라인 가베지 모델을 사용하는 방법[13]이 있다. 본 연구에서는 가베지 클라스 클러스터링 방법을 이용하여 한국어 자음을 음성학적 특성이 유사한 음소끼리 클러스터링하여 필러 모델로 사용하였다. 이 방법은 각각의 비핵심어 모두를 훈련시키는 방법들과는 달리 비핵심어 각각에 대해 필러 모델을 만들지 않고 클러스터링된 음소군 모델을 필러 모델로 사용[13, 14]하며, 핵심어가 아닌 부분이 몇 개의 비핵심어 HMM으로 표현되는 관계로 방대한 훈련용 자료가 필요하지 않은 잇점을 갖고 있다. (그림 2.3)은 기본적인 인식 네트워크로 핵심어 모델, 필러 모델, 배경 묵음 모델 등으로 구성된다.



(그림 2.3) 1개의 필러 모델에 의한 Null-grammar 인식 네트워크

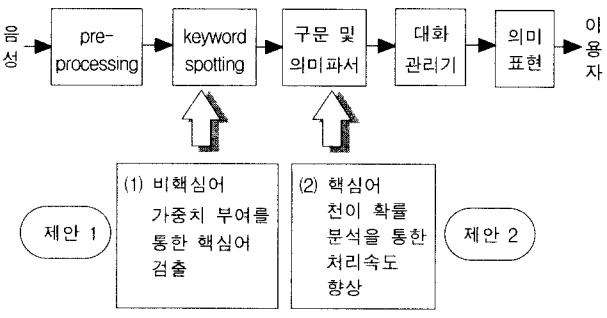
이 방법은 대부분의 필러 모델이 핵심어 모델보다 낮은 점수대를 보여, 인식 과정에서 핵심어 검출과 함께 많은 오류를 검출하는 문제점을 갖고 있어, 이를 방지하고 인식 성능을 향상시키기 위한 방안의 하나로 최근 필러 모델에 가중치를 주어 인식률을 높이는 방법이 활발한 연구가 진행되고 있다. 국내에서는 모노폰 클러스터링을 적용하여 모든 음소를 대상으로 '파열음', '마찰음', '파찰음', '유음 및 유성자음', '모음' 등 5개 그룹으로 클러스터링한 실험을 소개하고 있다[15]. Yapanel는 터키어를 위한 핵심어 검출 시스템의 연구에서 비핵심어 모델을 7 음소군으로 분류를 소개하였으며, 여기서 사용한 인식 네트워크는 (그림 2.4)과 같다.



(그림 2.4) 음소군 필러 모델에 의한 Null-grammar 인식 네트워크

3. HMM 기반 핵심어 검출시스템 성능 향상 제안

본 연구는 핵심어 검출기반 대화 음성처리 시스템의 성능을 개선하기 위한 연구로 2가지 관점에서 연구하였다. 먼저, 비핵심어 모델의 음소 클러스터링과 여기에 가중치를 부여하여 핵심어와 비핵심어간의 유사도 차를 제거하고 이를 통해 핵심어 인식 성능을 향상시키고자 하였으며, 두 번째로는 대화 음성처리 시스템의 실시간 처리를 위해 인식 성능과 계산 시간과의 trade-off를 보완하기 위하여 이용자 발화 성향 분석을 통해 처리 속도를 향상시키고자 한다. 이

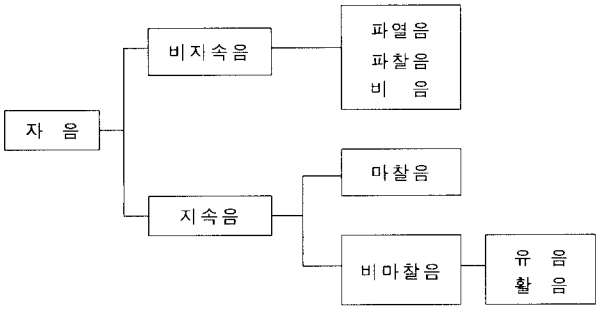


(그림 3.1) 대화음성처리 시스템을 위해 제안한 2가지 방법

용자 발화 성향 분석은 증권거래 이용자 200명을 대상으로 증권거래에 필요한 핵심어가 발화 문장 속에서 등장하는 순서를 분석하여 핵심어 천이 확률로 사용하였다. 즉, 핵심어 검출 우선 순위를 지정하기 위해 천이 확률을 적용함으로써 짧은 시간내에 화행에 적합한 핵심어의 검출로 인식 처리시간을 단축시키고자 한다. 이러한 방법을 대화 음성처리 시스템에 적용하면 (그림 3.1)과 같이 표현될 수 있다.

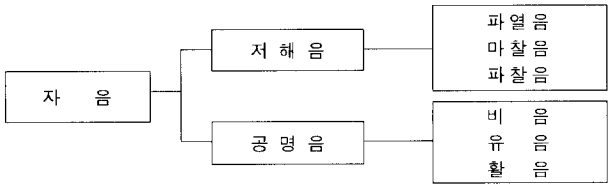
3.1 필러 모델 구성

본 연구에서 사용하는 가베지 클라스 클러스터링 방법(Garbage classes clustering method)은 트라이폰 또는 모노폰 모델을 음성학적으로 유사한 음소끼리 묶어서 사용함으로써 하나의 음소는 잘 나타내주지 못하지만 비슷한 음소 그룹의 표현에는 유용한 방법 중의 하나이다. 한국어의 자음은 조음방법 즉, 허파에서 올라오는 공기가 각각의 조음위치에서 어떠한 방식으로 방해를 받는가에 따라 파열음, 마찰음, 파찰음, 비음, 설측음, 접근음으로 분류한다[16]. 조음방법과 관련하여 또 다른 자음의 분류방법은 공기가 조음될 때 파열음이나 파찰음 등은 구강 내의 어느 곳이 완전히 막혔다가 터지는 비지속음(noncontinuant)이며, 마찰음, 유음, 활음 등은 그 음이 급격한 변화를 겪지 않고 똑 같은 상태로 발음되는 지속음(continuant)이다. 조음시 공기의 지속성에 따라 자음을 분류하면 (그림 3.2)와 같다.



(그림 3.2) 공기의 지속성에 의한 자음 분류

마지막 방법으로는 자음을 공명음(sonorant)과 저해음(obstruent)으로 분류할 수 있다. 공명음은 비음, 유음, 활음 등으로 악음(musical sound)의 특성을 갖으며, 저해음은 파열음, 마찰음, 파찰음 등으로 소음(noise)의 특성을 갖고 있다. 저해음은 유·무성의 대립이 있지만 공명음은 모두 유성음이다. (그림 3.3)은 소리의 공명성 여부로 자음을 분류하고 있다.



(그림 3.3) 소리의 공명성 여부에 의한 자음 분류

본 연구에서는 이들 자음의 조음 방법과 한국어 형태론[15]에 근거하여 자음을 7개 음소군 필러 모델로 분류하였다. 소음의 특성을 많이 갖고 있는 파열음, 마찰음, 파찰음 등과 혀끝을 잇몸에 가볍게 대었다가 떼거나, 혀끝을 잇몸에 댄 채 날숨을 양옆으로 흘려 보내면서 내는 소리인 유음, 콧소리인 비음, 일정한 음가가 없이 어떤 조음 위치에서 다른 조음위치로 빠르게 조음기관이 움직여가며 조음되는 활음과 숨소리, 입술소리, 문소리 등 잡소리를 각각 하나의 음소군으로 클러스터링하였다. <표 3.1>과 같다.

〈표 3.1〉 가베지 모델에서 이용한 음소군 분류

음 소 군	7 음소군	5 음소군
파 열 음	/ㅂ, ㅃ, ㅍ, ㄷ, ㄸ, ㅌ, ㄱ, ㄲ, ㅋ/	/ㅂ, ㅃ, ㅍ, ㄷ, ㄸ, ㅌ, ㄱ, ㄲ, ㅋ/
파 찰 음	/ㅈ, ㅊ, ㅉ/	/ㅈ, ㅊ, ㅉ/ + /ㅁ, ㄴ, ㅇ/
마 찰 음	/ㅅ, ㅆ, ㅎ/	/ㅅ, ㅆ, ㅎ/
비 음	/ㅁ,ㄴ,ㅇ/	—
유 음	/ㄹ/	/ㄹ/ + /y,w,r/
활 음	/y, w, r/	—
배경잡음	숨, 입술, 문소리 등	숨, 입술, 문소리 등

두 번째로는 필러 모델을 5개 음소군으로 분류하였다. 국내 증권시장에서 통용되는 71 문장을 35명이 각 2회씩 발음한 총 4,970 문장에 대해 필러 모델을 구성하였다. 이들 문장의 전체 비핵심어를 분석한 결과 파열음 77%, 파찰음 2%, 마찰음 9%, 비음 4%, 유음 2%, 활음 1%, 배경 잡음 5% 등으로 분석되어 음소군의 통합이 가능할 것으로 조사되었다.

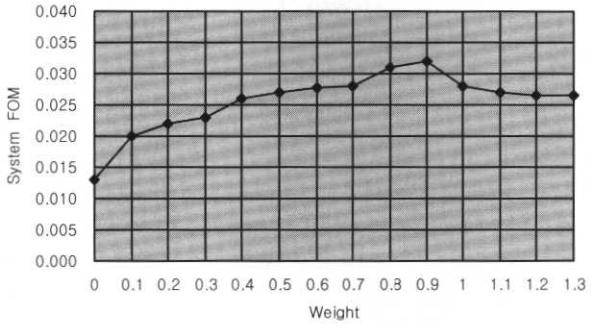
소군의 통합이 됨으로써 이들 음소군 중 파찰음과 비음을 통합하고, 유음과 활음을 통합하여 <표 3.1>와 같이 5 음소군을 제시하였다. 필러 모델에 대한 인식 네트워크는 (그림 2.4)를 사용하였다.

각 필러 모델별 HMM 초기화를 위한 로그 확률 수렴과정에서 여러 개 다른 음성신호 자질을 갖는 음소를 그룹화하기 때문에 수렴이 되지 않고 계속적인 초기화 과정이 반복(iteration) 발생한다. 이를 제한하기 위하여 최대 수렴 반복 횟수를 100회로 제한하여 초기화한 후 최종 필러 모델을 생성하였다.

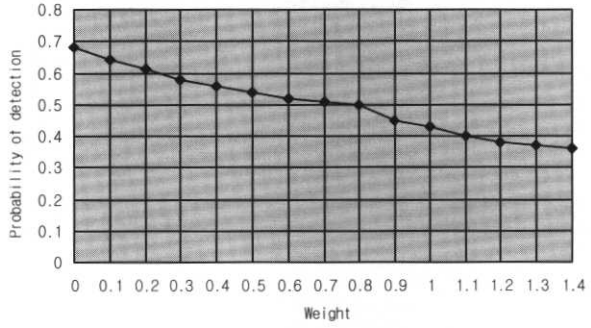
3.2 필러 모델 가중치 결정

필러 모델의 가중치는 핵심어 모델과 필러 모델사이의 로그 유사도 스코어 차이를 제거함으로써 임계 값을 설정하는 것보다 유용한 방법이나 태스크에 독립적인 가중치의 결정이 관건이다. Yapanel는 터키어를 위한 핵심어 인식 실험에서 가중치 변화에 따른 FOM, 최대 검출 확률 및 오류 검출수를 참조하여 인식기의 최적 운용상태에 필요한 가중치로 1을 적용하였다. 본 연구에서는 Yapanel이 제안한 방법을 적용하여 한국어 핵심어 인식기에서 가중치 변화에 따른 FOM, 최대 검출 확률, 오류 검출수를 확인하고 인식기의 최적 운용상태에 필요한 가중치로 0.9로 하였다. 이는 FOM이 최대 값을 나타낼 때 시스템의 최적 가중치로 결정

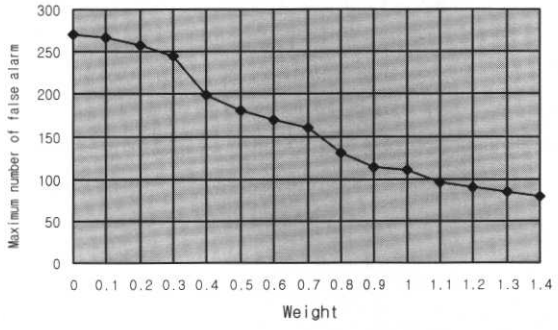
하였다. (그림 3.4)은 가중치에 대한 시스템 FOM의 변화를 보여주고 있다. (그림 3.5)에서는 가중치에 의한 최대 검출 확률을 나타내고 있으며, (그림 3.6)에서는 가중치에 의한 오류 검출수의 변화를 나타내고 있다.



(그림 3.4) 가중치 부여로 시스템 FOM의 변화



(그림 3.5) 가중치에 의한 최대 검출 확률의 변화



(그림 3.6) 가중치에 의한 오류 검출수의 변화

3.3 핵심어 천이 확률 분석을 통한 처리속도 향상

핵심어 인식기는 핵심어 검출 능력과 검출에 소요되는 계산 시간에 따라 음성 인식기로서의 실효성이 좌우된다. 음성 인식기의 실세계 적용을 위해서는 적정 핵심어 검출 능력과 적정 계산시간에 타협한 시스템이 필요하다. 본 연구에서는 인식기가 화자의 다음 발화 핵심어를 미리 예측할 수 있다면 처리 속도의 향상을 가져 올 수 있다. 본 장에서는 화자의 발화 행태를 분석하기 위해 증권거래 이용자를 대상으로 설문조사를 실시하였다. 설문 대상자는 현재 전화 음성으로 증권거래 이용 경험이 있는 자로서 주 이용대상 연령인 20~50대를 중심으로 200명을 참여시켰으며, 남녀의 구성은 각 연령별 동일하게 8 : 2의 비율로 하였다. <표 3.2>은 설문조사 대상자의 연령 분포를 보여주고 있다.

〈표 3.2〉 조사 연령별 현황

(단위 : 명, %)

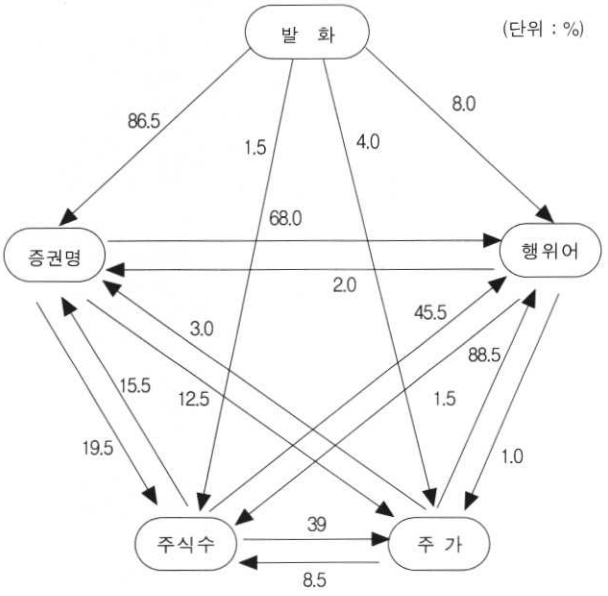
연령별	구성인원		구성비
	남	여	
20 대	20	4	10
30 대	80	16	40
40 대	70	14	35
50 대	30	6	15

설문조사 내용는 이용자가 음성을 통해 증권사에 증권거래를 문의 또는 매매를 요청할 경우 필요한 4개의 핵심어 그룹(증권명, 거래주식수, 주가, 행위어)의 발화 순서를 조사하였다. 설문조사 결과는 <표 3.3>과 (그림 3.7)과 같다.

〈표 3.3〉 발화 행태 분석 현황

(단위 : %)

구 분	증 권 명	주 식 수	주 가	행 위 어
발화시작	86.5	1.5	4.0	8.0
증 권 명	-	19.5	12.5	68.0
주 식 수	15.5	-	39.5	45.5
주 가	3.0	8.5	-	88.5
행 위 어	2.0	1.5	1.0	-



(그림 3.7) 발화 행태 분석에 의한 핵심어군 천이 확률

조사 결과 대상자의 86.5%가 발화의 첫 번째 핵심어로 증권명을 선택하였으며, 행위어를 첫 번째 핵심어로 발화한 대상자는 핵심어 1개만으로 처리할 수 있는 "종합주가지수", "종합지수", "코스닥" 등의 주가지수 조회에 한정되어 대상자 중 약 8% 정도로 나타났다. 또한 조사대상자의 68%는 "증

권명 + 행위어"를 이용하여 증권 매매보다는 거래가 또는 호가 등의 조회를 선호하고 있으며, 조사 대상자의 1.5%는 첫 발화 핵심어로 주식수를, 4%는 주가(매매가)를 발화하는 것으로 나타났다.

4. 인식 실험

시스템의 성능 분석을 위하여 다음과 같은 분석 관점을 선정하여 비교하였다. 먼저, 필러 모델 개수에 따라 FOM과 여기 각 필러 모델에 가중치를 부여했을 경우의 FOM을 상호 비교하고, 두 번째로는 각 필러 모델을 적용했을 경우에 음성 처리시간을 비교하였다. 세 번째는 제안한 핵심어 천이 확률을 음성 대화처리 시스템에서 적용하여 시스템 전체적인 처리시간을 확인하였다. 네 번째로는 각 필러 모델의 Mixture 개수의 변화를 통해 인식기의 인식률을 비교하였다.

4.1 음성 데이터 및 분석 조건

본 논문에서는 고립 단어 인식과 연속 음성 인식을 동시에 수행할 수 있는 증권거래 시스템을 구현하기 위해 증권명, 주가, 주식수, 행위어 등 핵심어 278단어와 이들로 구성된 71문장을 DB로 구축하였다. 실험에서 사용한 음성 데이타는 국내 증권시장에서 통용되는 문장 중 임의 71문장을 선정하고, 이를 남성화자 20명 여성화자 15명의 전화음성을 DB로 구축하였으며, 음성보드는 Dialogic의 D/120 JCT-LS Analog 12 Channel CSP(Continuous Speech Processing)음성보드를 사용하였다. 전화 음성은 mulaw로 하였으며, G.711의 규약에 의거하여 mulaw to linear PCM을 16bit로 변환한 음성 DB를 이용하였다. <표 4.1>는 데이터베이스의 일부를 보여주고 있으며, <표 4.2>은 음성의 분석조건을 나타내고 있다.

〈표 4.1〉 문장 데이터베이스의 일부 예

1) 하한가에 모아텍을 매도하겠습니다
2) 상한가에 백주
3) 강원산업우 매도호가
4) 경수종금을 칠백주 매도하겠습니다
5) 골드금고를 매수하겠습니다
:

〈표 4.2〉 입력 음성 데이터의 조건

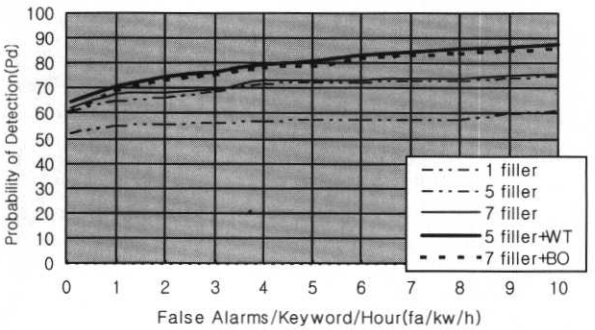
설 정 내 용	설 정 값
Sampling Rate	8000Hz
Channel	Mono
Quantization	16bit PCM(8bit u-law to linear)
음성 데이터 수	증권거래 71문장
화 자 수	남성화자 20명, 여성화자 15명
전화녹음	u-law, Analog Line
환 경	조용한 사무실
음성보드	D/120JCT-LS (Dialogic)

4.2 실험 결과 및 고찰

증권거래 주이용자 층인 20대~40대 남·녀 화자 5인을 대상으로 인식 실험하였으며, 적응화 데이터 및 인식 데이터를 SunMicrosystems Sparc station20에서 Off-line으로 실험하였고, PC상에서 실시간으로 인식 실험을 할 수 있도록 Pentium 2GHz 데스크 탑 개인용 컴퓨터에서 데모 시스템을 구성하였다.

4.2.1 필러 모델 가중치 부여를 통한 성능 비교

FOM은 핵심어 검출기의 보다 안정된 성능 측정 방법의 하나로 단위 시간에 핵심어 당 발생한 오류 검출수를 나타내는 오경보율이 0fa/kw/hr에서 10fa/kw/hr로 변하는 과정에서의 평균 핵심어 검출률을 의미하며, ROC의 전체적인 흐름을 단일 수치로 표현한 것이다. 각 필러 모델에 대한 ROC 커브는 (그림 4.1)과 같이 나타났다.



(그림 4.1) 필러 모델에 의한 ROC 도표의 비교

(그림 4.1)에서 제안한 필러 모델 및 가중치의 실험 결과를 위한 비교 실험은 핵심어 검출기의 가장 간단한 모델인 필러 1개 보유한 네트워크와 본 연구에서 제안한 필러모델 5개와 가중치를 부여한 방법 및 필러 모델 7개와 가중치를 부여한 방법들의 인식 성능을 비교 실험하였다. <표 4.3>에서 제안한 5 필러 모델에 가중치 0.9를 반영한 경우 FOM이 87.5%로 가장 높게 나타났으나, LVCSR의 89.8% 보다는 약간 못 미치는 결과를 보이고 있다. 또한 태스크 도메인으로 선정한 증권거래 시스템의 경우에는 7 필러 모델보다 5 필러 모델을 적용하고 가중치를 반영하는 것이 더 양호한 것으로 나타났다.

〈표 4.3〉 FOM 결과 요약표

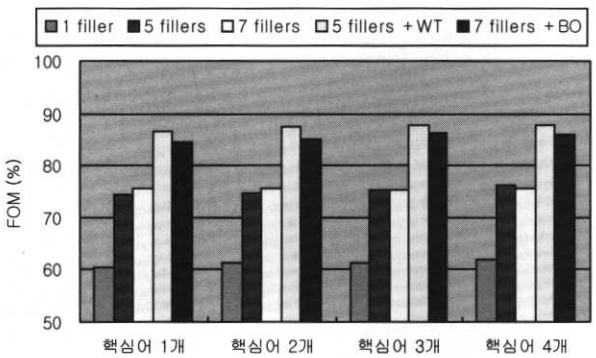
구 분	FOM
1 filler	61.3%
5 filler	75.1%
7 filler	75.5%
5 filler + weight	87.5%
7 filler + bonus	85.5%

<표 4.4>와 (그림 4.2)는 화자의 발화 음성 중에 포함된 핵심어 수에 따라 문장을 구분하여 각 20문장에 대해 화자 5명이 각 2회씩 발화하여 실험한 FOM 결과를 보여주고 있

다. 실험 결과 <표 4.3>과 비슷한 현상으로 5 음소군의 필러 모델에 가중치를 부여한 경우가 7 음소군의 필러 모델에 가중치를 부여한 경우보다 FOM이 평균 1.4% 양호한 것을 보여주고 있다.

〈표 4.4〉 핵심어 수에 따른 FOM 결과

필러 모델	발화 문장 중 핵심어 보유 수				평 균
	1개	2개	3개	4개	
1 filler	60.4	62.3	62.4	62.1	61.8
5 fillers	74.1	74.6	75.2	76.5	75.1
7 fillers	75.5	76.3	76.4	75.6	76.0
5 fillers + WT	86.6	87.6	87.1	88.1	87.4
7 fillers + BO	84.5	86.2	87.1	86	86.0



(그림 4.2) 필러 모델 및 핵심어 수에 의한 FOM 변화
(〈참조〉 BO (bonus) = 1, WT(weight) = 0.9)

핵심어 인식 시스템의 성능을 평가를 위한 또 다른 하나는 매 음성마다 평균 계산 시간을 계산하는 것이다. 각 화자로부터 하나의 음성을 선택하여, 각 음성마다 결과가 나오기까지의 평균 계산 시간으로 성능을 측정하였으며, 모든 실험 결과들은 시스템 종속성을 제거하기 위하여 정규화하였다. <표 4.5>는 실험 결과를 보여주고 있다.

〈표 4.5〉 가중치에 의한 계산시간의 변화
(〈참고〉 BO(bonus) = 1, WT(weight) = 0.9)

(단위 : 초)

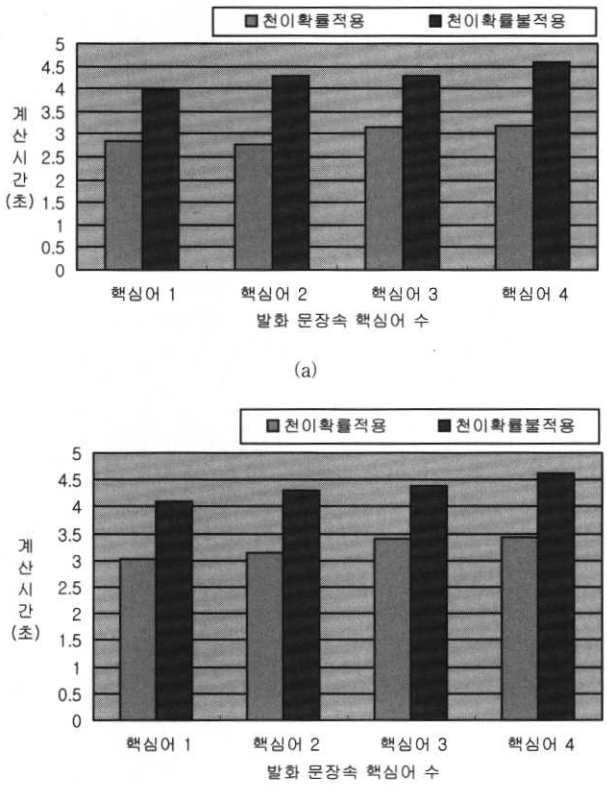
1 filler	5 filler	7 filler	5filler+WT	7filler+BO
0.42	0.67	0.81	0.70	0.72

4.2.2 핵심어 천이 확률 분석을 통한 인식시간 단축

핵심어 천이 확률에 의한 음성 인식 시간의 단축 실험은 먼저, 핵심어 그룹간의 천이 확률을 반영하지 않은 상태에서의 처리시간과 본 논문에서 제안한 핵심어 그룹의 천이 확률을 적용한 인식 실험을 각각 비교하였으며, 필러 모델은 5 음소군 모델과 가중치 0.9를 적용하였다. <표 4.6>과 (그림 4.3)(a)는 핵심어 천이 확률을 적용하지 상태에서의 인식시간이며, (그림 4.3)(b)는 핵심어 천이 확률을 적용하였을 때의 인식 결과이다. 인식 결과는 화자를 남·녀로 구분하였으며, 핵심어가 1~4개가 포함되어 있는 발화 문장을 각 5회씩 발음하여 그 평균 인식시간을 보여주고 있다.

〈표 4.6〉 계산시간 결과 요약표

핵심어 수	천이확률 적용		천이확률 불적용	
	남 성	여 성	남 성	여 성
1 개	0.65	0.63	0.71	0.71
2 개	0.65	0.64	0.73	0.72
3 개	0.68	0.66	0.69	0.69
4 개	0.7	0.7	0.72	0.74
평 균	**0.67**	**0.65**	**0.71**	**0.72**



(그림 4.3) 핵심어 수에 의한 계산 시간 변화

실험 결과 천이 확률을 적용한 경우 남성 화자 평균 0.04초, 여성 화자 0.07초의 인식처리 시간 단축 효과를 확인하여 증권거래 시스템과 같이 한정된 핵심어만으로 구현이 가능한 태스크에 대해서는 발화 문장속의 핵심어 천이 확률의 적용이 시스템의 최적화에 유용함을 확인하였다.

5. 결 론

본 논문에서는 먼저, 태스크 도메인 분석을 통해 가베지 클라스 클러스터링으로 필러 모델을 설정하고, 여기에 가중치를 부여하여 핵심어 검출기의 성능 향상과 처리시간을 단축하고, 두 번째로는 태스크 도메인의 이용자들에 대한

발화 성향 분석을 통해 얻은 핵심어 천이 확률을 대화 음성처리 시스템에 적용하여 시스템의 처리 시간 단축을 통해 실효성을 향상시키는 방안을 제안하였다. 제안한 방법인 가베지 클라스 클러스터링은 음성학적으로 유사한 음소끼리 묶어서 사용함으로써 하나의 음소는 잘 표현하지 못하지만 비슷한 음소 그룹의 표현에는 유용한 방법으로 본 논문에서는 한국어 형태론과 태스크 도메인으로 선정한 증권거래 대화음성처리 시스템에서 활용되는 발화 문장을 분석하여 5 음소군을 제시하였다. 그러나 이들 음소군만으로는 LVCSR 인식 능력에 걸맞는 성능을 구현하기 어려워, 한국어 음성 인식에 적용할 수 있는 필러 모델 가중치를 결정하고 시스템에 적용하여 비교 실험하였다. 두 번째로는 대화 음성처리시스템의 실용화에 필수적인 처리시간 단축을 위해 연속 발화 문장 속에 포함되어 있는 핵심어 천이 확률을 계산하여 시스템에 적용 실험하였다.

실험 결과, 제안한 5 음소군에 가중치를 부여한 방법의 FOM은 87.5%, 7 음소군에 가중치를 부여한 방법 85.5% 보다 우수한 성능을 보였으나, LVCSR의 89.8%보다는 약간 뒤지는 성능을 확인하였다. 그러나 계산시간에 있어서도 0.70초로 7 음소군의 0.72초보다 우수한 성능을 보였다. 핵심어 천이 확률 분석을 통한 인식 시간 단축 실험에서는 천이 확률을 적용했을 때 약 0.04~0.07초의 처리 시간을 단축하는 것을 확인하였다.

향후에는 음성 대화처리 시스템의 성능 및 실효성 향상을 위하여 한국어 음소의 필러 모델 구성에 있어 태스크 독립적인 음소 클러스터링의 연구를 계획중이다.

참 고 문 헌

[1] Ümit Yapanel, "Garbage modeling techniques for a turkish keyword spotting," Bogazici univ., 2000.

[2] R. Rose, "Definition of subword acoustic units for word spotting," Proc. EURO SPEECH 93, pp.1049-1052, 1993.

[3] P. Jeanrenaude, K. Ng, M. Siu, J. R. Rohlicek and H. Gish, "Phonetic-based word spotter : various configurations and application to event spotting," Proc. EURO SPEECH 93, pp.1057-1060, 1993.

[4] E. Lleida, J. B. Marino, J. Salavedra, A. Bonafonte, E. Monte and A. Martinez, "Out-of-vocabulary word modelling and rejection for keyword spotting," Proc. EURO SPEECH 93, pp.1265-1268, 1993.

[5] M. Weintraub, "Keyword-spotting using SRI's DECIPHER large-vocabulary speech recognition system," Proc. ICASSP 93, pp.463-466, 1993.

[6] 오영환, "음성 언어 정보처리", 홍릉과학출판사, 1998.

[7] Alexandros S. Manos and Victor W. Zue, "A segment based word spotter using phonetic filler models," Spoken Language Systems Group Laboratory for Computer Science Massachusetts Institute of Technology Cam bridge, 1996.

[8] Young, S. J., Russell, N. H., Thornton J. H. S., Token Passing : a Simple Conceptual Model for Connected Speech Recognition Systems, Technical Report, Cambridge University Engineering Department, July, 1989.

[9] Rose, R. C., Discriminant Word spotting Techniques for Rejecting Non-vocabulary Utterances in Unconstrained Speech, Proceedings of the1992 International Conference on Acoustics, Speech and Signal Processing, March, 1992.

[10] Knill, K. M. and Young S. J., "Speaker Dependent Keyword Spotting for Accessing Stored Speech," CUED/F-INFENG /TR 193, October, 1994.

[11] "Garbage modeling techniques for a Turkish keyword spotting system," 2001.

[12] Rose R. C. and Paul D. B., "A Hidden Markov Model Based Keyword Recognition System," in Proc. IEEE ICASSP, pp. 129-132, April, 1990.

[13] Herve Bourlard, Bart D'hoore and Jean-Marc Boite "Optimizing recognition and rejection performance in word spotting systems," in Proc. IEEE ICASSP, pp.1-373-376, 1994.

[14] J. G. Wilpon, L. R. Rabiner, C. H. Lee and E. R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden markov models," IEEE Trans. Acoust., Speech, Signal Processing, Vol.38, No.11, pp1870-1878, 1990.

[15] 김형순, "연속음성 인식에서의 Keyword spotting 적용방식 연구", 한국전자통신연구소, 1995.

[16] 허 웅, "국어음운론", 정음사, 1987.

김 학 진

e-mail : kimhj@shinbo.co.kr

1986년 건국대학교 전자계산학과(학사)

1997년 연세대학교 전자공학과(석사)

2002년 광운대학교 컴퓨터공학과 박사과정 수료

1986년~현재 신용보증기금 근무

2000년~현재 명지전문대학 컴퓨터정보과 겸임교수

관심분야 : 음성신호처리, 음성인식, 대화처리시스템, 자연어처리

김 순 협

e-mail : kimsh@daisy.gwu.ac.kr

1974년 울산대학교 전자공학과

1976년 연세대학교 석사과정

1983년 연세대학교 공학박사

1998년~1999년 한국음향학회 회장

1999년~2003년 광운대학교 사회교육원장

1979년~현재 광운대학교 컴퓨터공학과 교수

2000년~현재 한국음향학회 명예회장

2003년~현재 광운대학교 정보과학기술대학 원장

관심분야 : 음성신호처리, 음성인식, 자연어처리, 3-D 멀티미디어