한국어 핵심어 추출 및 연속 음성 인식을 위한 다목적 전처리 프로세서 설계

김동헌*, 이상준**

Design of Multi-Purpose Preprocessor for Keyword Spotting and Continuous Language Support in Korean

Dong-Heon Kim*, Sang-Joon Lee**

요 약 음성인식 기술은 단순한 단어 인식을 넘어 자연스럽게 발성한 연속 음성도 인식할 수 있는 수준으로 발전해 왔다. 아이폰에 탑재된 자연어 음성인식 처리 소프트웨어인 시리(Siri)가 2010년에 발표되면서, 음성인식에 대한 연구가 관심을 받고 있다. 한국어 음성 인식 소프트웨어들은 대부분 단어 위주의 인식 서비스로 구성 되어 있으며, 잡음 처리 및 음성 에너지 조절 기능들이 부족해 만족할 만한 인식률을 보이지 못하고 있다. 또한 요구된 발성 규칙을 따르지 못한 음성 질의들은 아예 처리하지 못하고 있는 실정이다. 본 논문에서는 이러한 현실적 어려움을 개선할 수 있도록 다목적 전처리 프로세서를 제안하였다. 이 처리기는 음성인식 엔진에 독립적이며, 잡음 제거 기능, 규칙에 따르지 않은 음성 질의도 처리 할 수 있는 핵심어 추출 기능, 그 핵심어를 수식하는 전술부 및 그 해당 음성 질의로부터 수행하기를 원하는 후술부 까지도 추출할 수 있는 기능을 갖추도록 하였다. 실험을 통해, 잡음 제거 효과 평가, 핵심어 인식 성공률, 연속음 인식 성공률을 측정하여 제안한 방법의 타당성을 확인하였다.

주제어 : 음성인식, 전처리기, 핵심어 추출, 소음 제거, 연속 음성

Abstract The voice recognition has been made continuously. Now, this technology could support even natural language beyond recognition of isolated words. Interests for the voice recognition was boosting after the Siri, I-phone based voice recognition software, was presented in 2010. There are some occasions implemented voice enabled services using Korean voice recognition softwares, but their accuracy isn't accurate enough, because of background noise and lack of control on voice related features. In this paper, we propose a sort of multi-purpose preprocessor to improve this situation. This supports Keyword spotting in the continuous speech in addition to noise filtering function. This should be independent of any voice recognition software and it can extend its functionality to support continuous speech by additionally identifying the pre-predicate and the post-predicate in relative to the spotted keyword. We get validation about noise filter effectiveness, keyword recognition rate, continuous speech recognition rate by experiments.

Key Words : Voice Recognition, Preprocessor, Keyword Spotting, Noise Filter, Continuous Speech

1. 서 론

음성인식 기술을 사용하여 기계와 사람간의 인터페이스를 보다 편리하고 자연스럽게 만들고자 하는 노력이 국내외에서 꾸준히 진행 되어 오고 있으며, 그 결과 단순한 단어 인식 수준을 넘어 자연스럽게 발성한 음성 도 처리 할 수 있는 수준으로 발전 되어 왔다[3][14[16]. 음성인식 기술은 지난 20세기 후반의 지속적인 기술개발에 힘입어 다양한 분야에서 실생활에 이용될 수 있는 수준으로 발전되어 왔지만, 아직 우리가 상상하는 수많은 음용

*(주)지앤넷
**전남대학교 경영학부(교신저자)
논문접수: 2012년 11월 18일, 1차 수정을 거쳐, 심사완료: 2012년 12월 5일

분야에 적극적으로 이용되기에는 아직 해결해야 할 기술적 과제가 산적해 있는 실정이다[1].

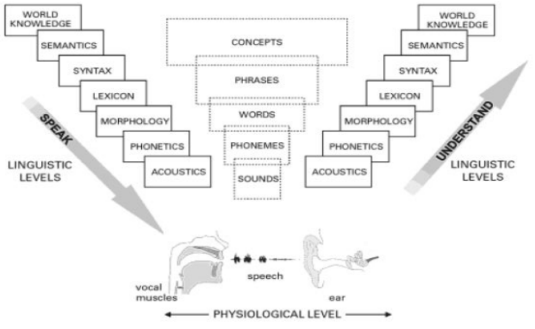
최근에 이러한 자연어 음성 인식 처리 기술을 활용하여 사용자들에게 보다 편리한 서비스를 제공하려는 노력들이 국내에서도 통신사 및 금융기관을 중심으로 일어나고 있다[7]. 스마트 폰에서도, 제한된 키보드를 통한 입력의 어려움을 개선하기 위해 음성인식 지원기능이 무선 모바일 기기 등에서 필수 기능이 되고 있다[8]. 이에 대한 솔루션으로 현재 애플은 음성인식 서비스인 시리(Siri)를, 안드로이드 진영에서는 구글 보이스를 사용하고 있다[4]. 구글에서는 아직까지 비영어의 인식에는 크게 영향을 못 미치고 있어서, 각 언어별로 연구되고 있다[13]. 음성 인식은 일반적인 비즈니스 영역 뿐만 아니라, 시각 장애인들을 위한 여러 서비스들에서도 매우 유용하게 사용 될 수 있는 기술이다.

본 논문은 이처럼 앞으로 광범위하게 사용되어 질 것으로 기대 되는 음성인식 서비스 시장에 보다 적은 비용으로 해당서비스를 구현 할 수 있도록 도와주는 지원 도구를 제안하였다. 제안하는 도구는 다음과 같은 기대효과를 얻을 수 있도록 설계되었다. 첫째, 음성인식 엔진과는 독립적으로 설계한다. 둘째, 소음 제거 기술 및 음질 개선 효과를 제공한다. 셋째, 음절 구분을 기반으로 핵심어를 추출할 수 있는 기능을 지원한다. 넷째, 연속 음성인식 서비스를 제공할 수 있도록 한다. 다섯째, 발음 보정 데이터베이스를 구축하고 이를 이용하여 인식률을 제고시킬 수 있는 후처리를 제공한다. 여섯째, 음성인식 성공률 최소 85% 이상을 지원한다.

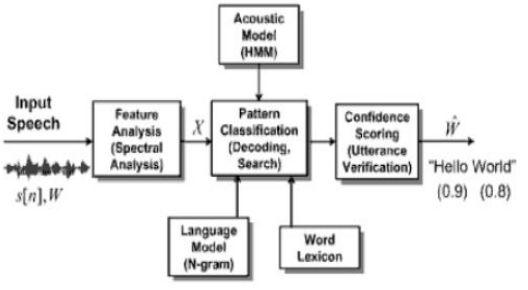
본 논문의 구성은 다음과 같다. 2장에서는 음성인식 기술 개요, 3장에서는 전체 시스템 설계, 4장에서는 본 연구 결과물에 대한 성능평가, 5장에서는 연구 결과의 의의를 논의하였다.

2. 관련연구

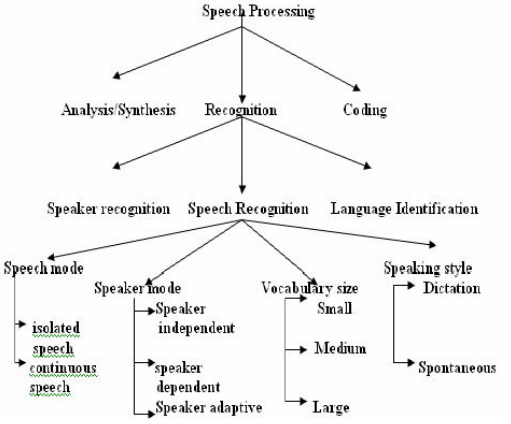
실생활에서 음성인식은 [그림 1]과 같이 말하는 사람의 지식, 의미, 문법, 문장요소, 단어등이 음성근육의 움직임에 따로 생리학적으로 말이 듣는 사람 귀에 전달되어 의미와 지식이 전달되는 과정이다[14]. 정보과학 분야에서는 음성에 포함된 음향학적 정보로부터 음운 및 언어적 정보를 취득하여 이를 사람이 아닌 기기가 인지하고 반응하게 만드는 일련의 과정으로 음성인식의 전체 다이어그램은 [그림 2]와 같다[16]. 음성인식의 구체적 구현은 1950년대에 있었던 Bell Lab의 화자 종속 숫자 음 인식이 그 시작이었다. 그 후 1970년대 미국 국방성에서 연속어 인식에 대한 연구를 시작했으며 1980년대에 이르러 인공지능중 하나의 방법인 패턴 인식에 기초한 음성 인식의 연구가 진행되어 왔다[2].



[그림 1] 음성 전달



[그림 2] 음성 인식 블록 다이어그램



[그림 3] 음성 인식 시스템 분류

음성처리는 대화 모드(단독, 연속), 화자 모드(화자 독립, 화자의존, 화자적응), 단어크기(소,중,대), 대화 스타일(구술식, 즉흥식)에 따라 분류되어 이 기준에 따라 음성인식 시스템은 [그림 3]과 같이 분류될 수 있다[9].

음성인식 시스템은 분석, 특징 추출, 모델링, 테스팅이라는 4단계로 동작된다[12]. 모델링 단계에서 음성 인식 기법은 음향음성학적 기법, 패턴인식 기법,뉴럴네트워크 기법, SVM(Support Vector Machine) 기법, 인공지능 기법 등이 있다. 이중 패턴 인식 기법에는 템플릿, DTW(Dynamic Time Warping), VQ(Vector Quantification), 통계적 기법[17]등이 있다. 최근에는 HMM(Hidden Markov Model)을 사용하는 통계적 기법이 사용된 연구들이 많이 있다[5][6][15][19]. HMM 기법은 인식할 수 있는 단어가 개수가 대용량 규모인 경우에도 적용할 수 있어서 연속음성 인식에도 많이 사용되고 있다[10][18][20][21].

음성은 같은 언어라 할지라도 발음하는 사람의 성별, 나이, 발음 하는 그 때 당시의 상태 등에 의해 매우 다르게 변할 뿐만 아니라, 단독으로 사용 되는지 또는 문장 내에서 사용되는 지에 따라서도 그 성질이 변한다. 이러한 음성의 특성을 잘 표현할 수 있는 특징 벡터들이 음성인식에 있어 가장 기본적인 요소가 된다. 현재 켑스트럼(Cepstrum), 라인 스펙트럼 주파수(Line Spectrum Frequency), 주파수 대역별 에너지(Filter Bank Energy) 및 웨이블렛(Wavelet) 등에 의한 특징 벡터들이 사용되고 있다[12].

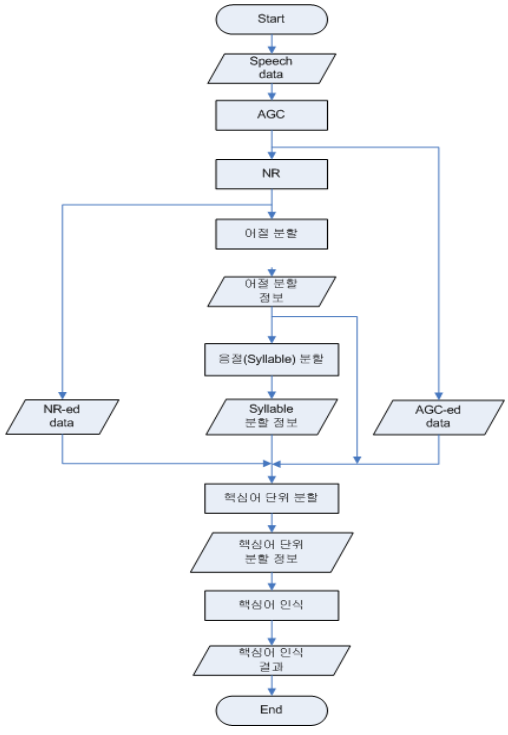
음성인식 기법들로 아무리 잘 구현된 인식기의 경우에도, 실제 활용에 있어서의 가장 큰 문제점 중의 하나가 인식기의 학습 환경과 다른 환경, 특히 배경 소음이 존재하는 환경에서 입력된 음성에 대해서 인식률이 현저하게 떨어지는 어려움이다. 더불어, 음성 인식기의 학습 시 고려할 수 있는 배경 소음은 한계가 있으므로, 음성인식 전처리 단계에서 배경 소음을 제거하는 과정이 반드시 필요하다[11][17].

소음 제거 기술은 음성 왜곡과 잔재 소음을 해결하고, 소음 추정의 지연 시간을 단축시키도록 요구되고 있다. 연속된 음성 신호들로부터 핵심어를 추출하기 위한 음절 기반 분할방법이 필요하다. 연속된 음성신호를 동일한 음운 특성을 갖는 소구간으로 나누는 것을 분할이라 한다. 정확한 음성 분할을 위해서는 음소에 대한 정확한 정보와 지식이 필요하다. 그러나, 발화자의 발음 습관 혹은 심리 상태등과 같은 발화자간의 개인성 때문에 정확한 음소의 경계점을 찾는다는 것은 매우 어려운 작업이다[12].

3. 전처리기 설계

본 연구를 통해 설계 제안 하고자 하는 전처리기는 크게 세 가지로, 소음제거 기술 및 음절 분리에 의한 핵심어 추출 기술, 인식된 핵심어를 기반으로 전술부와 후술부의 의미 구를 다시 추출하는 기능, 음절 구분을 통한 세그멘테이션이 그 경계 구분이 잘못 되었을 시 발생할 수 있는 오인식의 위험으로부터 핵심어를 보정하기 위한 발음 보정 데이터베이스 구축으로 되어 있다.

3.1 소음 제거 기술 및 핵심어 추출 기술 설계



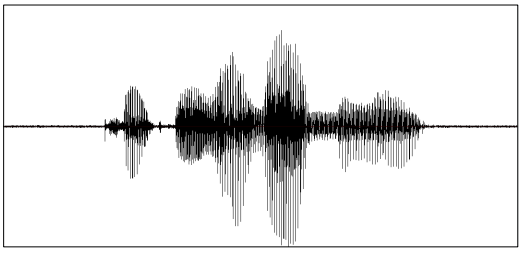
[그림 4] 노이즈 제거 및 음절 분리에 의한 핵심어 추출 과정 흐름도

[그림 4]는 소음제거 기술 및 음절 분리에 의한 핵심어 추출 과정의 흐름도로, AGC(Automatic Gain Control), 소음제거(Noise Reduction)), 어절분할, 음절분할, 핵심어 단위 분할, 핵심어 인식 모듈의 전반적인 흐름도이다. 이 전체 흐름도를 구성하고 있는 개별 기능들에 대한 설명

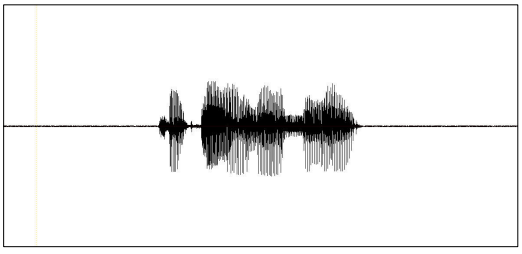
은 다음과 같다.

3.1.1 AGC(Automatic Gain Control)에 의한 입력된 음성데이터의 크기 조절

AGC를 거친 입력신호는 정해진 일정한 범위의 값으로 표현되어져 마이크로폰이나 시스템, 사용자 환경에 영향을 줄여 오 인식 방지에 도움이 된다. AGC를 이용하여 [그림 5]의 음성데이터를 [그림 6]과 같이 음성인식에 적합한 크기로 입력 음성데이터의 크기를 조절해 준다.

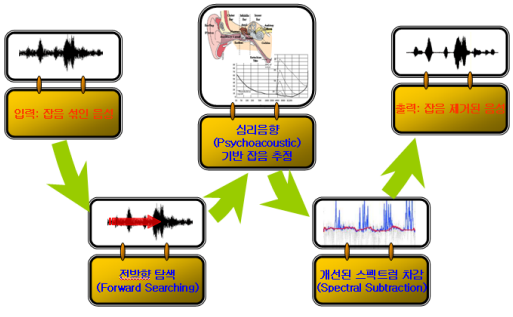


[그림 5] 남성화자가 발화한 "저기 영어안내요"에 대한 음성파형



[그림 6] AGC에 의해 크기가 조절된 그림 5의 음성파형

3.1.2 소음이 포함된 음성이 입력되었을 경우 소음을 감쇄/제거



[그림 7] 음성왜곡과 잔재소음 및 소음추정의 지연문제 극복을 위한 기술

발화자의 주변 환경에서 발생하는 소음이 음성신호에 섞여 마이크로 유입되는 경우, 오 인식을 유발하는 원인이 된다. 본 논문에서는 기존의 잡음 제거 기술이 가진 음성 왜곡과 추정 시간 지연 문제를 극복하기 위해 [그림 7]와 같은 방식을 사용했다.

(1) 전반향 탐색(forward searching)에 의한 판별기술

소음의 정보는 음성구간 검출기에 의해 얻어진 여러 소음 프레임에서 제시하는 통계적 정보를 이용하거나 1~1.5초에 해당하는 긴 과거 프레임에서 추적된 최소값(minima value)에 의존한 방식을 취한다. 음성 에너지가 약한 구간 또는 낮은 신호 대 소음비(SNR)에서 올바른 소음의 정보를 판독하기 어렵우며, 특히 소음이 갑자기 커지는 높은 비정상인 소음환경에서는 올바른 소음의 정보를 판독하는 것은 더욱 어렵다.

이를 해결하기 위해 과거 프레임으로부터 갱신된 식별자(indicator)를 지닌 효과적인 전방향 탐색 기술을 이용한다. 제안하는 식별자는 소음의 변화에 따라 적응 속도를 차등적으로 수행하기에 다양하고 높은 비정적인 소음환경에서 조차 소음의 정보를 1초 이내로 빠르고 정확하게 판독한다.

(2) 심리음향(psychoacoustic)기반의 추정기술

소음을 추정하기 위해서, 일반적으로 회귀 평균화(recursive averaging)를 기반으로 하는 가중된 평균화(weighted averaging) 방법은 현재 프레임에서 오염된 음성의 스펙트럼 크기와 이전 프레임에서 추정된 소음의 스펙트럼 크기 사이에 고정된 망각요소(fixed forgetting factor)를 이용하여 수행한다. 잘못 추정된 소음을 이용하여 개선된 음성은 잔재 소음 또는 음성왜곡이 일어날 수 있다.

이를 해결하기 위해 음향심리를 응용하여 계산된 SNR기반의 적응적 망각요소를 도입하여, 잔재소음과 음성왜곡을 거의 발생시키지 않고 정확한 소음추정을 수행한다.

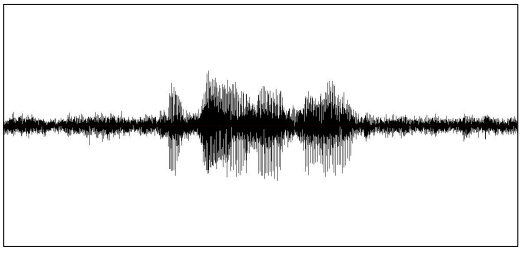
(3) 개선된 스펙트럼 차감에 의한 제거기술

스펙트럼 차감(SS, Spectral Subtraction)는 소음제거를 위한 여러 방법 중에서 적은 계산비용과 구현의 용이성 때문에 널리 사용되는 방법이지만, SS 방법에 의해 개선된 음성은 새로운 인공음(artifact)인 잔재소음을 유발

하는 주요 단점을 가진다. 이 문제점을 해결하기 위해 이득함수(gain function)를 기반으로 하는 여러 SS 방법이 사용되지만, 제시된 많은 방법들은 낮은 SNR을 가진 비정상인 소음환경에서 음질개선을 효율적으로 수행하지 못하는 것으로 알려져 있다.

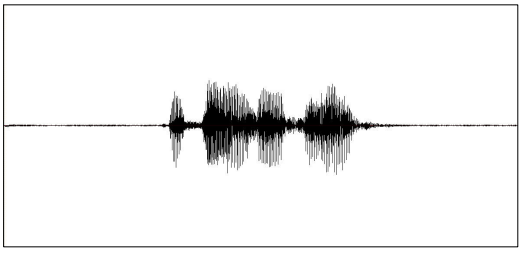
본 논문에서는 이러한 문제를 해결하기 위해 상태 크기 차이(RMD, Relative Magnitude Difference)를 기반으로 하는 비선형(nonlinear) 구조의 과중이득 함수를 가진 개선된 SS를 이용하여, 잔재소음의 유발을 효율적으로 억제할 수 있을 뿐만 아니라 음성명도를 신뢰적으로 제시함으로써 소음제거를 할 수 있다.

[그림 8]은 많은 사람들이 운집한 곳(광장, 식당 등)에서 여러 사람들이 웅성거리는 상태에서 발화된 음성인식 대상 어휘에 대한 파형으로, 배경잡음 신호성분에 의해 오인식이 유발될 수 있다.



[그림 8] Public place 잡음이 섞인 [그림 6]의 음성파형

[그림 9]는 본 논문에서 개발된 노이즈 필터에 의해 [그림 8]의 음성데이터에서 주변잡음 성분을 제거한 것으로, 다양한 배경잡음들에 대하여 SNR이 평균 18dB 이상 향상되는 효과를 얻었다.



[그림 9] 배경잡음이 제거된 [그림 8]의 음성파형

3.1.3 어절단위 분할

[그림 10]에서 인식대상 어휘가 '영어안내'이지만, 규칙에 따르지 않은 문인 '저기 영어안내 요'로 발화하여 인식에 실패하게 되는 것을 보여 준다.

[그림 10] AGC에 의해 크기가 조절된 "저기 영어안내 요" 에 대한 음성파형

이런 경우를 해결하고자 AGC와 소음제거 모듈을 통해 처리된 음성구간에 전체에 대하여 어절단위 분할 과정을 수행한다. 일련의 음성데이터를 대상으로 발화자에 의한 음성이 일시적으로 발화되지 않는 순간들을 1차 어절단위 분할 대상으로 설정한다.

다음으로, 1차 분할된 어절구간의 길이가 너무 짧거나, 어절구간 사이의 간격이 너무 짧을 경우, 인접한 두 어절구간을 하나의 어절구간으로 병합시킬 수 있다.

(1) 어절단위 분할 절차

① 5~20msec 중의 특정 값으로 동일하게 분할된 프레임 단위의 개선된 음성데이터를 이용하여 특징벡터(feature vector)를 구한다.

② 특징벡터는 음성데이터의 시간 또는 주파수영역의 파워(power) 또는 에너지(energy), 포만트(formant) 정보 등이 활용된다.

③ 각 프레임의 특징벡터들을 이용하여 인접한 3개 프레임의 평균 특징벡터를 구하여 프레임 특징벡터 평활화 과정을 수행한다.

④ 어절단위 분할을 위하여 평활화된 각 프레임의 특징벡터(Esmt(n))와 비교할 문턱값_1(Eth1), 문턱값_2(Eth2), 문턱값_3(Eth3), 문턱값_4(Eth4)를 설정한다.

⑤ n-1번 프레임의 평활화된 특징벡터가 문턱값_1 이하이고, n번, n+1번, n+3번 프레임 각각이 문턱값_2, 문턱값_3, 문턱값_4 이상이면 1차 어절 시작점 후보로 설정한다.

⑥ n+i (i=1,…,N-n)번 프레임이 문턱값_5(Eth5) 미만이거나, n+i번 프레임의 평활화된 특징벡터가 앞, 뒤로 인접한 2개 프레임들보다 모두 작을 때까지

프레임 특징벡터를 순차적으로 검사한다.

⑦ 프레임 특징벡터가 문턱값_5 이하인 n+i번 프레임을 1차 어절 종료점 후보로 설정한다.

⑧ 또한, Esmt(n-(i+1))>=Esmt(n-i) 이고 Esmt(n+i)<=Esmt(n+(i+1)) 인 경우에도 1차 어절 종료점 후보가 된다.

⑨ 1차 어절 시작점 후보에서부터 어절 종료점 후보 사이의 길이가 문턱값_6 이상이면 2차 어절구간 후보로 결정한다. 실험을 통해 문턱값_6은 150~170msec를 권장한다.

⑩ 2차 어절구간 후보 K개에 대하여 다음의 과정을 통해 3차 어절구간 후보를 결정한다.

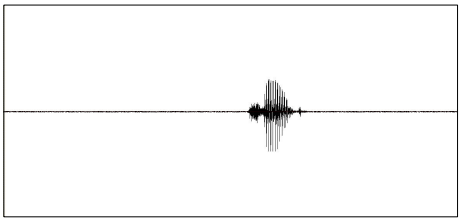
⑪ k번 2차 어절구간 후보의 끝점과 k+1번 2차 어절구간 후보의 시작점 사이의 길이가 문턱값_7 미만이면 k번과 k+1번 2차 어절구간 후보는 하나의 어절구간으로 병합한다. 실험을 통해 문턱값_7은 200~250msec로 설정한다.

⑫ 병합 과정을 수행한 후, 얻어진 2차 어절구간 후보들 중 길이가 400msec 이상이면 3차 어절구간 후보로 결정한다.

⑬ 3차 어절구간 후보들에 대하여 최종적인 어절구간으로 결정하기 위하여 다음의 과정을 수행한다.

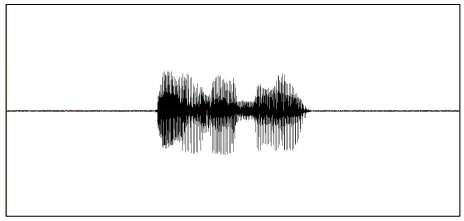
⑭ 2차 어절구간 후보들 각각에 대하여 각 어절구간 안에 포함된 프레임들의 평균 특징벡터를 구한 다음, 이 어절구간들의 평균 특징벡터 중의 최대값인 최대평균 특징벡터를 구한다.

⑮ 2차 어절구간 후보의 평균 특징벡터보다 최대평균 특징벡터가 M배 이상일 경우, 그 구간의 시작점과 종료점이 3차 어절구간 후보와 같거나, 3차 어절구간 안에 포함되어 있으면 최종 어절구간으로 결정한다.



[그림 11] "저기 영어안내요"에서 분할된 "저기"의 음성파형

[그림 11]과 [그림 12]는 발화된 "저기 영어안내요" 음성에 대하여 어절단위로 분할된 결과로, "저기"와 "영어안내요"로 분할됨을 보여준다.



[그림 12] "저기 영어안내요"에서 분할된 "영어안내요"의 음성파형

3.1.4 음절단위 분할

발화자의 특성에 따라 다른 어절로 분할될 것으로 예상되는 구간들이 하나의 어절이 되거나 하나의 어절이어야 할 구간이 두 개의 어절 형태로 나타나는 경우가 발생한다. 유성음과 무성음의 특성, 포만트(formant), 프레임 및 서브밴드 에너지 등을 이용하여 음절단위 분할을 수행한다.

(1) 음절단위 분할 절차

음절단위 분할을 위해서 최종 어절구간 분할 정보, 각 프레임의 평활화 특징벡터 및 서브밴드별 평활화 특징벡터(Esub(k,l))를 기반으로 아래와 같은 방법들을 이용한다. 여기서, k는 프레임 인덱스, l은 프레임 내 서브밴드의 인덱스를 의미한다.

각 방법들은 순서에 상관없이 병렬로 처리가 가능하다.

① 각 어절구간에 포함된 프레임들 각각에 대하여 N-포인터 DFT(Discret Fourier Transform)를 행하여 주파수영역에서의 각 DFT-포인터별로 파워를 구한 후, 이를 다시 L개의 서브밴드로 grouping하여 각 서브밴드별 특징벡터를 구한다.

② 음절구간 검출을 위해 문턱값_8, 문턱값_9, 문턱값_10, 문턱값_11, 문턱값_12를 설정한다.

(가) 방법-1

최종 어절구간들의 각 시작점은 해당 어절구간의 시작 음절 후보의 시작점으로 설정한다. 최종 어절구간들의 각 끝점은 해당 어절구간의 마지막 음절 후보의 끝점으로 설정한다.

(나) 방법-2

k-1번 프레임 1번째 서브밴드의 특징벡터가 0이 아니면서, k번 프레임의 1번째 서브밴드의 특징벡터가 0인 경우를 셈(counting)하여, 이 값이 문턱값_8보다 크면 k번 프레임을 임의 음절의 끝점 후보를 결정한다. 서브밴드의 개수는 청각특성을 고려하여 log scale로 6이 적당하며, 문턱값_8은 2 또는 3으로 하는 것이 적절하다.

(다) 방법-3

각 프레임이 가지고 있는 L개 서브밴드별 평활화 특징벡터의 크기를 비교하여, 내림차순으로 크기 순서 정보를 구한다. 이전 프레임과 현재 프레임의 서브밴드별 평활화 특징벡터 크기 순서를 비교하여 순서가 바뀐 서브밴드의 개수가 문턱값_9보다 크면 음절이 변경되는 후보 구분점으로 설정한다.

(라) 방법-4

각 프레임별 평활화 특징벡터 크기 변화의 추이(증가 또는 감소) 정보 FEgrd(k)를 추출하여, 다음과 같은 절차에 따라 음절 변경 후보 구분점을 설정한다.

① 만약 FEgrd(k)>0 이면서, Esmt(k-1) > Esmt(k) 이면 음절 끝점 후보로 설정하고, FEgrd(k)은 -1로 설정한다.

② 만약 FEgrd(k)<0 이면서, Esmt(k-1) < Esmt(k)와 Esmt(k) > 문턱값_10을 만족하면 음절 변경 시작 후보점으로 설정하고, FEgrd(k)은 1로 설정한다.

③ 만약 FEgrd(k)=0 이면서, Esmt(k) > 0와 Esmt(k) > 문턱값_10을 만족하면 음절 변경 시작 후보점으로 설정하고, FEgrd(k)은 1로 설정한다.

④ 만약 FEgrd(k)이 0이 아니면서 Esmt(k)=0이면 각각 음절 변경 끝 후보점으로 설정하고, FEgrd(k)은 0으로 설정한다.

앞에서 구한 어절 및 음절 구간 후보 점들에 대하여 다음의 과정을 거쳐 최종적인 음절단위 분할 정보를 결정한다.

① 순차적으로 첫 번째 음절 시작 후보점 이전의 음절 끝점 및 변경점 후보는 탈락시킨다.

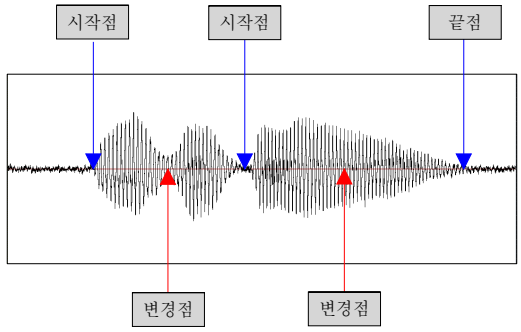
② 맨 마지막 음절 끝 후보점 이후의 음절 시작점 및 변경점 후보는 탈락시킨다.

③ 각 음절의 시작 후보점 전,후로 문턱값_11 이내에 있는 음절 구간 후보점들은 탈락시킨다.

④ 각 음절의 끝 후보점 전,후로 문턱값_12 이내에 있는 음절 변경 후보점들은 탈락시킨다.

[그림 13]은 음절분할에 있어 음절의 시작점, 끝점, 변경점을 나타낸 것으로, 시작점과 끝점은 한 음절의 발화가 시작되거나 마무리되는 시점으로 구분이 대체로 용이하다. 복수개의 음절이 존재하는 경우에 음절 사이에서는 이전 음절의 끝점과 다음 음절의 시작점이 같을 수 있으며, 이러한 경우에는 시작점으로 설정한다.

음절 변경점의 경우는 발성 시 유성음이 연속되어, 음절 시작점과 끝점보다 찾아내기가 어려우며, 분할시 불연속점이 발생하는 특징을 가진다.



[그림 13] "여보세요"에 대한 음절 분할 예시

3.1.5 핵심어 인식단위로 분할

위의 방법으로 구해진 어절 및 음절 구간 정보를 기반으로, AGC를 거친 음성데이터와 잡음이 제거된 음성데이터 각각에 대하여, 다음과 같은 과정을 통해 음성인식 대상인 핵심어 인식 단위로 분할한다.

① 핵심어 인식 단위 분할에 사용할 문턱값_13, 문턱값_14를 설정한다.

② 각각의 최종 어절구간 내에 존재하는 음절 구간의 개수가 문턱값_13 이하이면, 이 어절구간은 하나의 핵심어 인식 대상 단위로 결정한다.

③ 각각의 최종 어절구간 내에 존재하는 음절의 개수가 문턱값_13을 초과하면, 분할된 음절구간을 S개씩 이동시키며 W개의 분할된 음절구간을 병합한 후, 이를 하나의 핵심어 인식 대상 단위로 결정한다.

④ 음성인식엔진이 적용될 시스템의 특성에 따라, S는 1~3, W는 2~어절구간 내의 음절 개수까지 적용할 수 있다.

⑤ 인식 대상 핵심어의 길이 정보를 파악해, 핵심어 인식 대상 단위 결정에 사용되는 W 값 설정에 반영한다.

⑥ 한 어절구간 내에 존재하는 핵심어 인식 대상 단위의 개수가 문턱값_14보다 많으면, 문턱값_14번째에 해당하는 핵심어 인식 대상 단위에는 어절구간 내의 나머지 모든 음절구간들을 포함시킨다.

⑦ 인식률 향상을 위해 각각의 핵심어 인식 대상 구간의 전/후에 길이 P의 묵음구간을 추가해 인식 대상 데이터들이 중앙에 위치하도록 보정한다.

⑧ 묵음구간에 사용되는 데이터는 실제 입력된 음성신호에 포함된 묵음구간의 데이터를 활용하여 적용함으로써, 인식 대상 음성신호 성분과 조화를 이루게 된다.

⑨ 핵심어 인식 대상 구간의 시작점과 끝점이 음절의 시작점과 끝점이 아닌 음절 변경점일 경우, 가장자리에 신호의 불연속점이 존재해, 추출한 특징벡터가 왜곡되는 문제가 발생할 수 있어, 이를 해결하기 위하여 음절 변경 점에 해당하는 구간에는 해닝창(hanning window)을 적용해 신호의 불연속점을 없애준다.

3.1.6 핵심어 음성인식 예

AGC를 거친 음성데이터로 만든 각각의 핵심어 인식 단위 데이터를 음성인식엔진의 입력으로 넣어 인식결과를 구한다. 인식결과로 구해진 R개의 핵심어 후보들 중에 다음의 조건에 따라 최종적으로 인식된 핵심어를 결정한다.

① 가장 큰 신뢰도(confidence level)를 가진 핵심어 후보의 값이 일정값 이상이면 최종적으로 인식된 핵심어로 결정한다.

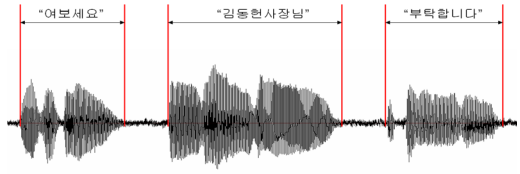
② 가장 큰 신뢰도를 가진 핵심어 후보의 값이 일정값 미만인 경우, 최소 값 보다 큰 모든 핵심어 후보들에 대하여 사용자가 최종 핵심어를 선택하도록 제시해 준다.

③ 가장 큰 신뢰도를 가진 핵심어 후보의 값이 문턱값_16 미만인 경우, 주변잡음신호에 의해 특징벡터가 영향을 받은 것으로 가정해, 잡음이 제거된 음성데이터로 만든 각각의 핵심어 인식 단위 데이터를 음성인식엔진의 입력으로 넣어 ①과 ②의 과정을 반복한다.

④ 잡음이 제거된 음성데이터로도 가장 큰 신뢰도를 가진 핵심어 후보의 값이 문턱값_16 미만인 경우는 인식 실패로 결정한다.

인식률 및 처리속도를 높이기 위해 AGC를 거친 음성데이터와 잡음이 제거된 음성데이터 각각의 핵심어 인식 단위를 동시에 병렬 음성인식엔진의 입력으로 사용하여 처리할 수도 있다.

[그림 14]는 핵심어 인식단위 분할에 의한 음성인식의 예로, 두 번째 어절에서 핵심어 인식단위로 분할된 4개에 대하여 첫 번째 핵심어 인식단위에서 인식대상어인 '김동헌'이 검출되어 인식에 성공하였다.

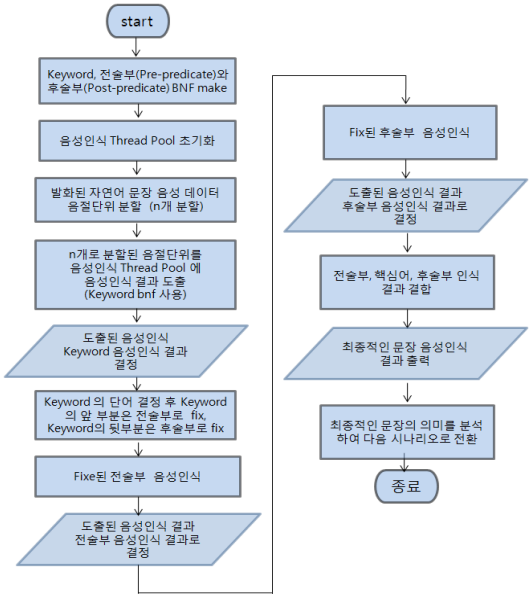


• 핵심어 인식 단위 분할 결과: S=1, W=3, 문턱값_13=4 인 경우
– "부탁합니다"는 "부"와 "탁합니다"가 병합된 형태임

어절: 3개	어절별 음절수	어절별 핵심어 인식 단위수
여보세요	4개	1개: 여보세요
김동헌사장님	6개	4개: 김동헌, 동헌사, 헌사장, 사장님
부탁합니다	5개	3개: 부탁합, 탁합니, 합니다

[그림 14] 핵심어 인식 단위 분할에 의한 음성인식 예

3.2 전술부와 후술부 추출 방안 설계



[그림 15] Keyword, 전술부, 후술부 음성 인식 과정

핵심어 추출 수행시와 마찬가지로, 전술부와 후술부의 의미구 추출 기능 수행할 때도 병렬 알고리즘을 이용한다. [그림 15]는 핵심어, 전술부(Pre-predicate)와 후술부(Post-predicate) 의미구를 추출 인식하는 핵심 흐름도이다.

3.3 핵심어 보정을 위한 발음 보정 데이터 베이스 구축 설계

음절 구분을 통한 세그멘테이션이 단어와 단어의 경계를 놓쳐 실제 발화자의 의도와는 상이한 핵심어가 인식될 수 있는 위험을 사후적으로 보정할 수 있는 보정 데이터베이스 구축을 통해 최소화 할 수 있다. 추출된 핵심어가 잘못된 음절 분할로 인해 또는 어절 내에 인식 후보가 2개 이상 존재하여 오 인식 될 수 있는 위험이 있다. 이러한 오인식을 사후에 보정하고자 발음 보정 데이터베이스를 구축한다.

<표 1>은 발음보정 데이터베이스 예를 보여준다. 소장님의 경우 사장님을 발음 보정 데이터베이스에 등록한다. 이러한 발음 보정 데이터베이스를 사용 '연구소장'을 연결하기를 원하는 발신자에게 사장이라는 발음 유사어를 같이 인식 후보 군으로 제시하여 발화자가 소장 과 사장중에서 선택할 수 있게 처리 할 수 있다. 발음 보정 데이터베이스에 등록 될 수 있는 단어들은 인식하고자 하는 핵심어들을 바탕으로 하여, 이와 유사하게 발음 될 수 있는 단어들로서 인식하고자 하는 핵심어 집합에 속해 있는 것들로 제한된다.

〈표 1〉 발음보정 데이터베이스 예

인식 핵심어	전체 핵심어 군중 발음유사 후보어
소장님	사장님
사장님	소장님

4. 성능 평가

본 연구 결과물인 전처리기의 성능을 측정코자, 샘플 데이터 와 미국 뉴앙스 사의 단어 인식 소프트웨어를 사용하여 그 인식 성능을 평가하고, 본 전처리기 이용으로 인한 잡음 제거 향상정도, 그리고 요구된 규칙에 따르지 않고 자연스럽게 발성한 질의어의 경우들에 대한 처리 결과들을 제시한다.

4.1 잡음 제거 효과 평가

ETSI(European Telecommunications Standards Institue)의 평가를 위한 노이즈는 Pink, Pub, Outside traffic road 3 가지가 있으며, 그 각 noise의 기준은 아래와 같다.

① Pink : 단위 주파수 대역에 포함된 성분의 세기가 주파수에 반비례되는 성질을 가진 잡음으로. flicker 잡음 또는 1/f 잡음이라고도 한다. 일반적으로 백색 잡음의 동일한 음량을, 1옥타브 올라갈 때마다 -3dB씩 감소시킨 것으로 시험용 대역 잡음원으로 사용하고 있다. 백색 잡음은 주로 음향 장비들 간의 주파수 응답 특성을 알아보고자 할 때 사용하며, 분홍색 잡음은 사람의 귀로 판단해야 하는 공연장이나 녹음 스튜디오에서 많이 사용한다.

② Pub: 선술집과 같이 많은 사람들이 모여서 얘기하는 소리와 접시나 잔 부딪히는 소리 등이 혼재하는 비정적인 잡음이다.

③ Outside traffic road: 자동차들이 지나다니는 도로에서 발생하는 잡음이다.

본 논문에서는 측정된 잡음 제거 평가 항목을 ITU-T G.160에규정되어 있는 SNRI(Signal-to-Noise Ratio Improvement)를 사용하여 개선 정도를 측정하였다. <표 2>의 Input SNR은 잡음 제거 효과를 평가하기 위해 사용되는 입력신호의 신호 대 잡음비를 나타낸 것으로 5dB, 10dB, 15dB에 대하여 본 노이즈 제거 필터를 사용한 결과 SNRI 개선 정도가 Pink noise 에서는 평균 24, Public noise 에서는 평균 21, Outside traffic road 에서는 평균 22로 개선 되었음을 나타낸다.

〈표 2〉 배경잡음에 따른 잡음제거 및 음질 평가 결과

Noise	Input SNR	SNRI[dB]
Pink	5dB	21.3
	10dB	25
	15dB	26.8
	Average	24.4
Public	5dB	16.8
	10dB	20
	15dB	21.4
	Average	19.4
Outside traffic road	5dB	20.9
	10dB	23.6
	15dB	21.9
	Average	22.1

4.2 핵심어 인식 성공률

<표 3>은 G사의 전화번호 안내 서비스에 핵심어 추출 전처리기를 사용하여 인식률이 높아진 실 사례 평가 결과를 보여 준다. ARS로 전화를 걸어 전화 받고 싶은 상대를 자연스럽게 발성한다. 예를 들어 “저기 김동헌 사장님 바꿔주세요” 라고 발성 시에도 “김동헌 사장님”을 인식하고 해당 사용자에게 전화를 연결하는 시나리오이다. "저기“, ”바꿔주세요“ 라는 인식 단어는 인식단어 테이블에 데이터가 없으므로, 해당 내용은 무시하고, ”김동헌 사장님“ 부분만 전 처리기가 추출 인식하여 작동된다. 단어 인식 엔진만 사용 시에는 자연스런 문장은 전혀 인식하지 못하였으나, 단어인식 엔진 + 전처리 엔진의 경우 92%를 인식하는 성과를 거두었다.

〈표 3〉 핵심어 인식 성공률 결과

NO	발성내용	단어인식엔진 + 핵심어 추출 사용
1	저기 김동헌 사장님 바꿔주세요	o
2	저기 김동헌 사장님 바꿔주세요	o
3	저기 원명숙 과장님 바꿔주세요	o
4	저기 김주라 팀장님 바꿔주세요	o
5	저기 채송화 대리 바꿔주세요	o
6	저기 이영삼 상무님 바꿔주세요	o
7	저기 김학수 이사님 바꿔주세요	o
8	저기 이덕규씨 바꿔주세요	o
9	저기 양선희 소장님 바꿔주세요	o
10	저기 신동진씨 바꿔주세요	o
11	저기 류정훈씨 바꿔주세요	o
12	저기 이득주 부사장님 바꿔주세요	o
13	저기 이병택 이사님 바꿔주세요	o
14	저기 안재현 본부장님 바꿔주세요	o
15	저기 양희준 매니저님 바꿔주세요	x
16	저기 차지혜 팀장님 바꿔주세요	o
17	저기 이지우 강사님 바꿔주세요	o
18	저기 류혜림 총무님 바꿔주세요	o
19	저기 양영한 부장님 바꿔주세요	o
20	저기 김민기 대리님 바꿔주세요	o
음성인식률(%)		95

■ 매개변수 설정값

〈표 4〉 매개변수 설정값

매개변수	값
SyllablesPerPartLimit	7
SyllablesPerWindow	A3
SlidingStep	1
MaxMeaningUnit	4

■ 인식단어 리스트 일부

〈표 5〉 인식단어 리스트 예

1	김동헌
2	김동헌사
4	사장님
6	비서실
7	양선희
8	양선희소

4.3 전술부와 후술부 추출에 따른 연속음 인식 성공률

<표 6>은 국내 K 통신사의 고객 질의 중 주요한 질문 20개를 사용하여 전술부, 후술부추출에 의한 연속 음성 인식 결과를 보여 준다. 단어 인식 엔진을 사용한 경우와 단어인식엔진 + 핵심어 추출 + 전술부 + 후술부 사용의 경우 그 인식율이 16%에서 96%로 향상되는 결과를 가져왔음을 알 수 있다.

〈표 6〉 전술부와 후술부 추출에 따른 연속음 인식 성공률

NO	발성내용	단어인식엔진 + 핵심어 추출 + 전술부 + 후술부 사용
1	요금 확인 부탁 드립니다.	o
2	이 번호로 나올 대금 알려주세요	o
3	이번 달에 현재까지 사용한 요금이 총 얼마예요	o
4	현재까지 제가 사용한 요금이 얼만가여?	x
5	요금제 변경이요	x
6	제 요금상품 젤 싼걸로 바꿔주세요	o
7	기본요금	o
8	요금 납부 문의 하려구요	o
9	우리 아들 휴대폰 요금 낼려고 그랍니다.	o
10	지금까지 밀려서 못낸 요금 얼마예요	o
11	이번달 요금이요	o
12	이달에 낼 게 얼만가요?	o
13	제가 자동이체를 하는데 요금 언제 빼가나요?	o
14	핸드폰요금내역에대해문의할라그랍니다.	o
15	납부방법 변경	o
16	저번달 요금	o
17	청구요금 확인해 주세요.	o
18	요금제 확인	o
19	단말기 대금 얼마 남았어요?	o
20	당월 청구요금이 이상해요	o
음성인식률(%)		96

6. 결론 및 의의

본 논문에서는 음성 구간 추출, 잡음 제거 등이 음성 인식 소프트웨어와 밀 결합 되어 있는 종래의 설계방법에 대해 그 음성인식에 필요로 되는 핵심 요소들을 논리적으로 구분 설계하였다. 이러한 논리적인 구분으로부터 각 요소 기술들이 개선되어 지면 그 개선 효과를 기존 음성인식 시스템의 환경의 변화를 최소화 하며 수용할 수 있는 설계상 이점을 제공 할 수 있다. 즉 미래에 더 향상된 잡음 개선 기술이나 한글 음절 분리기술이 가용해 질 때 그와 같은 기법들을 전처리기에 손쉽게 수용함으로써 향상된 음성 인식 성공률을 제공할 수 있다.

동적 소음 제거 기술을 제시함으로써 현재 음성인식 기술이 환경적 제약에 의한 인식률 저하에 대한 개선안을 제시 하였다.

본 논문에서는 음성인식에 전처리 부분에서 처리 할 수 없는 인식 결과를 보정하기 위한 보정 데이터베이스를 제시 함으로써 음절 분할 에러 혹은 잡음에 의한 오인식을 최소화 할 수 있는 실용적인 방법을 제안하였다. 특히 음성 인식 소프트웨어와 별도로 설계되었다는 점으로부터, 이미 음성인식 서비스를 하고 있는 경우에도 기존 음성 인식 소프트웨어를 그대로 활용하면서 전처리기만을 추가로 설치하는 것으로 핵심어 인식 기능의 보강 및 더 나아가 자연스럽게 발성된 연속음도 처리할 수 있는 기능을 추가 할 수 있다.

본 논문의 결과는 전출부 및 후출부 추출 기술을 통해 대화형 질의 응답 시스템을 가능케 함으로써 로봇, 텔레매틱스, 홈오토메이션 등의 많은 응용분야에서 기존 제품들의 부가가치를 높이고 새로운 시장을 창출할 수 있을 것으로 기대된다. 또한, 발음 보정 데이터베이스를 통한 후처리기 기술은 현재 무상으로 제공되고 있는 딕테이션 엔진들과 함께 활용됨으로써 세멘틱 의미를 획득할 수 있는 기술로도 발전 될 수 있을 것으로 기대된다.

참 고 문 헌

[1] 김회린 (2003). 음성인식 기술, 한국멀티미디어학회지, 7(2), 16-22.

[2] 신옥근 (2000). 음절핵의 위치정보를 이용한 우리말의 음소 경계 추출, 한국 음향학회지, 19(5). 13-19.

[3] 이건상, 양성일, 권영헌 (2004). 음성인식, 한양대학교 출판부.

[4] 이윤근 (2012). 음성인터페이스 기술 개요 및 스마트폰 환경에서의 서비스 동향, 한국통신학회, 29(4), 3-9.

[5] 정용주 (2012). 음성특징 보상을 이용한 멀티모델 기반 음성인식기의 성능향상, 한국정보기술학회논문지, 10(7), 179-184.

[6] 한학용, 고시영, 허강인 (2001). 우리말 연속 음성의 음절 분할법, 한국 음향학회지, 20(3). 70-75.

[7] 현대카드, 음성인식 ARS 서비스실시, 디지털 타임스, 2010, 05, 19.

[8] KTF, 음성인식 ARS 서비스, 파이넌셜 뉴스, 2010, 05, 19.

[9] Anusuya, M. A. & Katti, S. K. (2009). Speech Recognition by Machine: A Review, International Journal of Computer Science and Information Security, 6(3), 181-205.

[10] Dahl, G. E., Dong Yu, Li Deng & Acero, A. (2011). Large vocabulary continuous speech recognition with context-dependent DBN-HMMS, IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 4688-4691.

[11] Delcroix, M., Watanabe, S., Nakatani, T. & Nakamura, A. (2013). Cluster-based dynamic variance adaptation for interconnecting speech enhancement pre-processor and speech recognizer, Computer Speech & Language, 27(1), 350-368.

[12] Gaikwad, S. K., Gawali, B. W. & Yannawa, P. (2010). A Review on Speech Recognition Technique, International Journal of Computer Applications, 20(3), 16-24.

[13] Pelemans J., Demuynck K. & Wambacq P. (2012). Dutch automatic speech recognition on the web: Towards a general purpose system, 13th annual conference of the International Speech Communication Association (ISCA).

[14] Pieraccini, R. (2012). The Voice in the Machine: Building Computers That Understand Speech, The MIT Press.

[15] Poveya, D., Burgetb, L., Agarwalc, M., Akyazid, P., Kaie, F., Ghoshalf, A., Glembekb, O., Goelg, N., Karafiátb, M., Rastrowh, A., Rosei, R. C., Schwarzb,

P. & Thomash, S. (2011). The subspace Gaussian mixture model－A structured model for speech recognition, Computer Speech & Language, 25(2), 404-439.

[16] Rabiner, L. R. & Schafer, R. W. (2010). Theory and Applications of Digital Speech Processing, Prentice Hall.

[17] Ragni, A. & Gales, M. J. F. (2011). Structured discriminative models for noise robust continuous speech recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4788-4791.

[18] Sainath, T.N., Kingsbury, B., Ramabhadran, B., Fousek, P., Novak, P. & Mohamed, A.-R. (2011). Making Deep Belief Networks effective for large vocabulary continuous speech recognition, IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 30-35.

[19] Singh, B., Kapur, N. & Kaur, P. (2012). Speech Recognition with Hidden Markov Model : A Review, International Journal of Advanced Research in Computer Science and Software Engineering, 2(3), 400-403.

[20] Triefenbach, F., Demuynck, K. & Martens, J. (2012). Improving large vocabulary continuous speech recognition by combining GMM-based and reservoir-based acoustic modeling, Proceedings of IEEE Workshop on Spoken Language Technology.

[21] Van Segbroeck, M. & Van Hamme, H. (2011). Advances in Missing Feature Techniques for Robust Large-Vocabulary Continuous Speech Recognition, IEEE Transactions on Audio, Speech, and Language Processin, 19(1), 123-137.

김 동 헌

- 1983년 2월: 서울대학교 사범대학 수학과(이학사)
- 1996년 2월: 서강대학교 공공 정책대학원 정보처리학과(이학석사)
- 2012년 2월: 전남대학교 전자상거래 협동과정 박사과정 수료
- 현재: (주)지앤넷 대표이사 및 한국과학기술대학교 산업정보시스템학과 겸임교수
- 관심분야: 전자상거래, 음성인식, 경영정보시스템 등
- E-Mail: dhkim@gnnet.co.kr

이 상 준

- 1991년 2월: 전남대학교 전산통계학과(이학사)
- 1993년 2월: 전남대학교 전산통계학과(이학석사)
- 1999년 8월: 전남대학교 전산통계학과(이학박사)
- 1995년 3월~2005년 2월: 서남대학교 경영전산정보학과 조교수
- 2005년 3월~2007년 2월: 신경대학교 인터넷정보통신학과 조교수
- 2007년 2월~현재: 전남대학교 경영학과 부교수
- 관심분야: 경영정보시스템, 스마트컴퓨팅, 소프트웨어공학
- E-Mail: s-lee@chonnam.ac.kr