특집

음성정보처리 기술 개발 현황 및 전망

김 회 린*

목 차

1. 서 론

미국 MIT의 기술혁신 잡지인 테크놀로지 리뷰(Technology Review)는 2004년 2월호에서 미래를 변화시킬 10대 기술(10 emerging technologies that will change your world) 중 그 첫 번째로 만국어 번역(Universal Translation)을 선정하였다. 만국어 번역은 통역자가 중간에서 어떤 음성언어를 상대방 언어로 즉시 통역해 주는 일을 컴퓨터가 자동으로 처리해 주는 것이다. 이 기술이 상용화되면 그 사회적 파급효과가 실로 지대한 꿈같은 기술로서 이 기술의 핵심에는 음성정보처리 기술 및 언어정보처리 기술이 자리 잡고 있다.

최근의 음성정보처리 기술은 지난 20세기 후반의 지속적인 기술 개발에 힘입어 개별 요소기술들이 다양한 분야에서 실생활에 이용될 수 있는 수준으로 발전되어 왔지만, 아직 만국어 번역과 같은 궁극적인 응용 분야에 적극적으로 이용되기에는 아직 해결해야 할 기술적 과제가 산적해 있는 실정이다. 음성정보처리 기술 중 음성인식 기술은 원래 음성통신 기술의 개발과 더불어 연구가 시작되었다. 통신 기술은 기본적으로 사람과 사람 사이의 통신의 자유도를 향상시키는 방향으로 발전되어 왔는데, 최근에는 이를 뛰어 넘어 사람과 기계 사이에 음성을 이용한 대화를 가능케 하는 기술에 대한 연구 및 개발에 큰 관심이 집중되고 있다. 이와 같은 voice-activated man-machine interface 기술에 대한 요구는 원래 음성을 전송할 때 전송 효율을 최대한 향상시키기 위해서는 음성을 신호 자체가 아니라 그 음성을 문자로 표현한 symbol 들로 변환하여 전송할 때 가장 효율적으로 전송할 수 있다는 발상에서 출발하였다. 그러나 만일 사람이 발성한 음성을 기계 즉, 컴퓨터가 문자로 변환하고 이를 이해할 수 있다면, 이는 우리가 공상과학 소설이나 영화에서 무수히 접해 왔던 대로 다양한 분야에서 그 파급효과가 대단히 클 것으로 쉽게 예상할 수 있다.

그러면, voice-activated man-machine interface 기술의 구성 요소는 무엇일까? 이는 사람과 사람 사이의 대화 과정을 보면 쉽게 이해할 수 있다. 즉, 사람의 두뇌에서 생성된 개념이 일정한 규칙을 가지는 특정 언어 형태로 완성되어 입의 조음기관에서 공기의 진동으로 전파되어 나오는 과

* 한국정보통신대학교(ICU) 공학부 조교수

정이 음성언어의 발생 과정이며, 이를 우리는 "음성합성" 과정이라고 한다. 다음으로 공기 중에서 전파되는 과정을 바로 옆의 사람이 아니라 멀리 떨어진 사람에게 전달되도록 할 때 이를 우리는 음성통신, 특히 "음성부호화" 과정이라고 한다. 한편, 전달되어 온 음성 신호를 청각기관인 귀를 통해 받아들여 모종의 신호변환 과정을 거쳐 두뇌로 전달하여 음성신호 내에 포함되어 있는 음성언어를 인지하는 과정을 "음성인식" 과정이라고 한다. 또한, 상대방의 음성만을 듣고도 그 사람이 누구인지를 판별해 내는 과정을 "화자인식" 과정이라고 한다. 이렇게 음성언어의 발생/전달/인지 과정을 음성합성/음성부호화/음성인식 및 화자인식의 3단계로 나누어 볼 때, 음성합성 및 음성부호화 기술은 아직 미흡한 측면이 있지만 그런대로 상용화되어 비교적 널리 이용되고 있는 반면, 음성인식 기술은 아직 크게 대중화되지 못하고 있는 실정이다. 이는 근본적으로 음성합성 및 음성부호화 시스템은 최종 수신측이 매우 지능적인 인간이기 때문에 비록 자연스럽지는 못해도 청취한 음성이 무슨 내용인지를 사람이 판단할 수 있는 수준까지 기술이 발전한 때문이라고 볼 수 있지만, 음성인식이나 화자인식의 경우는 수신측이 훨씬 덜 지능적인 컴퓨터이므로 잘못 인식된 결과에 대한 지능적 재처리가 매우 힘들어서 아직 널리 이용되고 있지 못한 실정이다.

본 고에서는 음성정보처리 기술의 주요 연구 분야들에 대하여 최근의 기술 동향을 분석하고, 향후 기술 발전을 조망해 보기로 한다. 이를 위해 2장에서 사람의 말 즉, 음성언어를 컴퓨터가 알아듣는 음성인식 기술의 최근 연구동향을 살펴보고, 3장에서는 화자인식의 연구 동향에 대하여 살펴본다. 다음으로 4장에서는 음성합성의 최근 기술을 요약한다. 5장에서는 이러한 음성정보처리 기술들의 향후 연구를 예측해 보며, 마지막으로 6장에서 결론을 맺기로 한다.

2. 음성인식 연구 동향

음성 인식의 궁극적인 목표 즉, 잡음이 있는 실제적인 환경에서 임의의 화자가 어휘에 제한 없이 자연스럽게 발음한 연속 음성을 실시간에 인식 및 이해하는 수준을 만족시키는 시스템은 아직 개발되지 못하고 있으며, 현재의 음성 인식 시스템들은 몇 가지 제약 조건 하에서 운용됨을 전제로 하고 있다. 음성 인식의 난이도를 결정하는 대표적인 요인들은 다음과 같다.

- 인식 대상 어휘 규모, 발성 형태, 문법 구조 및 주제: 일반적으로 인식하고자 하는 어휘 수가 많아질수록 난이도가 높아진다. 물론 동일한 어휘 수라고 할지라도 단어들의 음성학적 유사성에 따라 난이도는 다르다. 예를 들면, '불 켜'와 '불 꺼'의 구별이 '예'와 '아니오'의 구별보다 어렵다. 한편, 각 단어와 단어를 또박또박 띄어 발음하는 고립 단어의 인식에 비해, 자연스럽게 연결시켜 발음한 연속 음성의 인식이 어려우며, 말하는 속도가 빨라질수록 더 어려워진다. 자유 발화 대화 음성은 입력 음성에 문법적 제약을 많이 두는 경우에 비해 인식하기 매우 어렵다. 또한 다수의 주제를 대상으로 하는 경우가 호텔 예약 등 특정한 단일 주제로 내용을 한정할 때보다 인식하기 어렵다.

- 화자 독립성: 자신의 목소리를 미리 등록시킨 특정인의 음성을 인식하는 화자 종속 인식에 비해, 훈련에 참여하지 않은 임의의 다수 화자의 음성을 인식하는 화자 독립 인식이 어렵다. 이러한 화자 독립성을 보완하기 위한 방법으로 화자 적응 기법을 적용하면 특정 화자에 보다 적합한 음성 인식이 가능하게 된다.

- 인식 대상 음성의 왜곡: 자동차 소음, 주변 사람의 음성 등 배경 잡음이 부가되어 신호대 잡음비(SNR)가 낮을 경우나 사용하는 마이크 및 전화망과 같이 미지의 채널 왜곡이 있을 경우 인식 성능

이 현저하게 떨어진다. 기본적으로 훈련 환경과 인식 환경을 일치시키는 경우 인식 성능이 향상되나, 근본적으로 모든 환경을 포함하는 훈련은 불가능하고, 또한 SNR이 극히 낮은 경우에는 이러한 방법도 효과가 저하된다.

- 시스템 제약: 음성 인식이 적용될 응용 분야에 따라 최종 제품의 연산 능력, 메모리량, OS 등의 제약 조건에 의해서 이용 가능한 음성 인식 방식에 제약이 따른다.

2.1 전처리 기술

음성 인식 기술이 절실히 요구되는 응용 분야에서는 일반적으로 사용 환경에 큰 배경소음이 존재하여 입력음성을 왜곡하거나, 입력음성이 상이한 특성을 가지는 채널을 통하여 전달되어 왜곡되는 현상으로 인하여 음성 인식기의 성능이 크게 저하된다. 이러한 additive noise 및 convolutive noise에 강인하거나 이들을 보상하는 방법들이 상용인식기 개발의 관건으로 부각되어 다음과 같은 여러 가지 해결방안이 제안되어 적용되고 있다.

- 적응신호 처리 기법을 이용한 반향제거 (Adaptive echo cancellation)
- 스펙트럴 차감(Spectral subtraction)
- 스테레오 신호를 이용한 주파수 영역 MMSE
- RASTA 필터링
- 켑스트럴 평균 정규화(Cepstral mean normalization)
- 1차 혹은 2차 dynamic features 추출
- PCA(Principal component analysis)를 적용한 특징벡터 차원 수 감축
- LDA(Linear discriminant analysis)를 적용한 특징벡터의 class간 변별력 향상
- 잡음에 강인한 음성구간 검출

이러한 노력의 일환으로 최근 유럽에서는 분산처리 음성 인식(distributed speech recognition: DSR) 방식의 음성 인식 처리구조에서 음성 전처리 규격을 표준화 하려는 움직임이 활발히 진행되고 있다. DSR이란 어떤 종류의 통신 단말에서나 전처리부만 수행하고 인식은 중앙처리 방식을 사용하는 시스템을 의미하며, 이 때 단말기 주변의 잡음에 강인한 전처리부를 표준화하기 위하여 ETSI(European Telecommunication Standards Institute)의 STQ(Speech processing, Trans-mission and Quality aspects)-Aurora group에서는 일명 Aurora Project를 수행하여 여러 가지 표준화 방식들을 제안하고 있다.

2.2 음향 모델링 기술

HMM을 기반으로 한 음성 인식에 있어서 최종적인 음성 인식의 성능을 향상시키기 위한 노력이 지난 20여년에 걸쳐 꾸준히 진행되어서 다양한 입력환경 및 응용 분야에 대처하기 위한 방법들이 제안되어 왔다. 특히, 음향모델을 개선하기 위한 방법들로 대표적인 진전은 다음과 같은 방법들로부터 성취되고 있다.

- 화자 및 환경의 변화를 능동적으로 반영하는 모델 파라미터 선택 및 적응 기법들
- Maximum a posteriori (MAP) estimation
- Maximum likelihood linear regression (MLLR) estimation
- Clustered models: gender-dependent model, cluster adaptive training, eigenvoice modeling
- Parallel model combination (PMC)
- Vector Taylor series (VTS)
- 보통의 음향모델로 모델링하기에는 부적합한 배경잡음, 입술소리와 같은 비언어적인 잡음 등을 별도의 모델로 패턴화하는 방법: non-stationary noise modeling
- 인간두뇌의 신경망 구조를 단순한 수학적 구조로 근사화하여 정적 패턴의 변별능력을 향상

시키기 위한 음향모델링: neural networks
- HMM이 가지고 있는 근본적인 가정들 중 관측벡터의 독립 가정을 보완하기 위한 방법으로 보다 긴 구간의 특징벡터의 시간축에서의 궤적을 HMM framework에 통합하는 방법: segment models
- Parametric trajectory model
- Unified frame- and segment-based model

2.3 언어 모델링 기술

소규모 영역의 시스템을 위하여 FSN이나 CFG와 같은 언어모델이 사용되었고, 이러한 언어모델은 어휘간의 관계를 네트웍으로 표현하므로 대어휘로의 확장이 어렵다. 그러므로 대어휘를 위한 언어모델링 기술은 통계적 언어모델 기법을 따르고 있다.

- FSN (Finite State Network): 인식하고자 하는 어휘의 연결관계를 네트웍으로 표현하여 영역이 제한적일 때 성능이 우수
- CFG (Context Free Grammar): 문법을 FSN과 같은 네트웍으로 표현하여 FSN보다 자유도가 크지만, 모든 가능 문법을 표현하기 어려움
- Smoothing: deleted interpolation, back-off (Good-Turing, Katz, Kneser-Ney), class N-grams
- Long distance language model: cache language model, trigger language model, whole sentence language model
- Adaptive language model: topic-based language model, maximum entropy model
- 의미문법(semantic grammar): LSA (Latent Semantic Analysis)
- Structured language model: 어휘들의 구조정보를 반영한 언어모델로 기존 N-gram과 결합하여 성능 향상

2.4 탐색 기술

최근 수년간의 탐색기술 개발은 대어휘 연속음성 인식을 고속으로 처리하는 다양한 기법들의 개발에 집중되어 왔으며, 특히 보다 정밀한 음향모델의 수용, 고성능 언어모델과의 효율적인 통합, 다중 인식 후보 문장들의 출력 등의 분야에서 아래와 같은 여러 가지 방법들이 제안되고 성공적으로 적용되어 실제 인식시스템에 사용되고 있다.

- 단어 경계에서 context-dependent acoustic model을 사용하는 방법
- 탐색공간을 효율적으로 축소하기 위한 tree lexicon
- 언어모델 확률의 factorization
- 다단계 beam search
- State level, phone level, word level
- Multi-pass search
- Forward-backward search algorithm using tree-trellis search
- 다중 인식결과 후보 추출 방법
- N-best lists: exact N-best algorithm, word-dependent N-best algorithm
- Word lattices: word-lattice algorithm, word-graph algorithm

2.5 발화검증 기술

HMM과 같은 확률모델을 이용한 음향모델을 사용할 경우 가장 견고한 발화검증 기법은 log-likelihood ratio를 비교척도로 사용하는 것이다. 이 방법은 입력된 음성발화를 등록어휘로 구성된 탐색공간에서 검색하여 최대 likelihood를 가지는 단어열을 찾고, 다음에 각 단어의 얻어진 likelihood를 그 단어와 가장 혼동하기 쉬운 어휘들(cohort set)에 대한 평균 likelihood와 log 영역에서 비교하여 그 비율이 어떤 임계값보다 크면 인식결과를 인

정하고, 그렇지 않으면 거절하는 방식이다. 이 때 각 likelihood를 구하고 정규화하는 방식들에 따라 여러 가지 알고리즘이 제안되어 왔다. 한편, 발화검증 기법을 인식과정의 후처리 과정으로 수행하지 않고 주 탐색과정과 통합하여 탐색공간을 줄이고 인식 성능을 향상시키는 방식에 대한 연구도 수행된 바 있다. 이러한 화자검증 기법들은 주로 화자인식 (speaker recognition) 및 핵심어 인식 (keyword recognition) 분야들에서 개발되기 시작하였으며, 현재는 음성 인식이기의 상용화 주요 모듈로서 그 중요성이 강조되고 있다.

3. 화자인식 연구 동향

화자인증 시스템의 개략적인 과정은 다음과 같다. 입력된 음성 신호는 잡음신호 구간을 제거하는 끝점 검출 과정을 거쳐 특징추출 단계에서 특징 벡터로 변환된다. 등록단계에서는 사용자가 발성한 음성에서 추출된 특징 벡터열을 이용한 학습과정에 의하여 화자모델이 만들어진다. 사용단계에서는 사용자가 사용자 번호를 제시하고 음성을 발성하면, 이 음성에서 추출된 특징 벡터열과 화자모델을 비교하여 유사도를 측정한다. 이 유사도와 배경화자와의 유사도와의 비율 등을 이용하여 사용자의 음성이 제시한 사용자 번호와 일치하는 지를 결정한다. 이때, 유사도가 미리 정해놓은 임계값 보다 높으면 승인하고, 그렇지 않으면 거부하게 된다.

3.1 전처리 기술

화자인증에서는 음성 인식과 반대로 화자내 변이보다 화자간 변이가 큰 특징을 사용해야 한다. 화자인증을 위한 특징에 관하여 많은 연구가 진행되어 왔으나, 대부분의 경우에는 음성 인식에서 좋은 효과를 내는 특징이 화자 인식에서도 높은 성능을 보이고 있다. 주로 사용되는 특징으로는 MFCC(Mel-frequency cepstral coefficients)와 이의 1차 및 2차 미분이 사용되고, 잡음을 감소시키기 위하여 Cepstral mean subtraction (CMS) 방법을 사용한다.

3.2 화자 모델링 기술

화자모델링은 화자의 특성을 표현하는 방법을 말하는 것으로, 최근에는 통계적인 방법이 주로 사용된다. 아래에는 대표적인 모델링 방법들을 요약하였다.

- 동적정합법(DTW, Dynamic Time Warping): 발화속도가 다른 두 음성을 정합하는 데 가장 일반적으로 사용되는 방법
- VQ source modeling: 음성의 모든 프레임을 표현하기 위해서 몇 개의 표준패턴을 만들어 놓는 방법
- Nearest Neighbor: DTW와 VQ 방법의 장점을 모두 살린 방법
- Gaussian Mixture Modeling: 각 화자를 특징 벡터를 출력하는 랜덤 소스로 가정하고, 이를 바탕으로 다차원 가우시안 확률분포 함수에 의해 화자를 모델링하는 방법
- UBM(Universal Background Model): 배경화자를 표현하기 위하여 하나의 화자독립 배경화자 모델을 사용한다. 이 모델은 특정 화자에 상관없이 모든 화자의 공통적이고 일반적인 특징을 표현하도록 학습하여 화자확인 시 거절 기준에 사용한다.
- 화자 모델 적응을 이용한 화자모델 생성: UBM으로부터 적응을 통하여 화자모델을 학습하여 각각의 화자모델을 생성하는 방법

3.3 인증 기술

화자 식별 시스템에서 최종 결과의 결정은 비교적 간단해 진다. 고립단어 음성 인식 시스템에서와 같이 여러 후보 단어들 중에서 가장 유사도가 높은

단어를 선택하면 되기 때문이다. 화자 확인 시스템에서는 입력된 음성이 제시된 화자의 음성인지 아닌지만 판별하면 되므로 매우 단순해 보이지만 실제로는 화자 식별 시스템 보다 훨씬 복잡해진다. 그것은 제시된 화자에 해당하는 모델은 등록과정에서 잘 모델링 되지만 이와 반대되는 모델이 정확히 정의되기 어렵기 때문이다. 화자인증 시스템은 이와 같이 잘 정의된 모델과 잘 정의되지 않은 모델간의 구별을 해야 한다. 테스트 과정에서는 입력된 음성의 화자가 제시된 화자와 같을 때의 선택은 H0이 되고, 다를 때는 H1이 된다. 따라서 최종적인 인증과정은 이 H0와 H1의 유사도비를 이용하여 결정하게 된다. 가장 일반적으로 적용되는 방법으로는 log likelihood ratio가 있다.

4. 음성합성 연구 동향

음성합성에 대한 연구는 1970년대에 들어서 MIT 및 AT&T Bell 연구소 등을 중심으로 활발한 연구가 이루어져 1980년대부터 상품화된 음성 합성기들이 출시되기 시작했다. 1980년대 초반에 개발된 대표적인 TTS 시스템으로 Digital Equipment 사의 DECTalk 시스템을 들 수 있다. 이 시스템은 MIT 대학의 Klatt 교수 등이 연구용으로 개발한 Klattalk 시스템 및 MITalk 시스템을 상용화시킨 것이다. 이 시스템은 합성 방식으로 Formant 방식을 채용하고, 음성 기본단위로는 음소를 사용한다.

최근 음성 합성 분야의 연구동향은 대용량의 음성 데이터베이스로부터 합성단위를 연결하여 합성하는 코퍼스 기반의 연결 합성방식이 주류를 이루고 있다. 1995년 영국 Cambridge 대학교는 인식기의 향상된 성능을 합성에 이용하고자 HMM 기반의 trainable 합성기를 처음으로 구현하였다. 합성 DB 구축에서부터 합성음 생성까지 자동으로 이루어지도록 한 이 방식은 다른 화자, 다른 언어로의 전환이 매우 용이하며, 미국 Microsoft의 음성 합성기를 비롯한 다수의 상품화 제품에 많은 영향을 끼쳤다.

1995년 일본의 ATR 연구소에서는 대용량 코퍼스로부터 합성단위를 온라인으로 선정하는 CHATR 방식을 개발하였다. 이 방식에서는 음성 기본 단위들의 운율환경이 최적화되는 target cost와 인접 음성 단위들 사이의 연결의 자연스러움을 판단하는 continuity cost를 고려한 동적 프로그래밍 방법을 사용하여 신호 처리 과정에서의 음성왜곡을 최소화할 수 있다. 이 방식은 코퍼스 기반 합성방식의 주요 수단으로 널리 사용되고 있다.

국내의 음성 합성 연구는 한국과학기술원, 한국전자통신연구원, 한국통신, LG전자기술원, 삼성종합기술원 등에서 1980년대 후반에 시작되었으며, 1993년 디지콤에서 개발한 '가라사대'가 최초의 상용화 제품이다. 1990년대에 들어서서는 음성파형상에서 운율조절이 가능한 PSOLA(Pitch Synchronous Overlap and Add) 방식이 도입되어 합성음의 명료성이 많이 향상되었다. 그러나 음성파형을 그대로 이용하는 PSOLA방식에서는 음소경계, 피치 등의 정보가 매우 정밀하게 분절되어야 하기 때문에 합성 데이터베이스를 제작하는데 많은 시간이 소요되는 문제점이 있었다. 1990년대 중반에 들어서 하드웨어의 발전에 따라 합성 DB 크기의 제약이 점점 줄어들게 되었고, 합성단위의 측면에서도 음운환경뿐 아니라 피치, 지속시간, 에너지 등 운율적 요소까지 고려된 합성단위를 사용하려는 코퍼스 기반의 합성방식이 등장하게 되었다. 이에 따라 국내에서도 최근의 합성방식은 코퍼스 기반의 합성방식이 주류를 이루고 있으며, 음성 DB의 크기도 수 GB 규모로 증가되었다. 이 방식의 경우 합성단위의 분절 과정을 과거의 방식대로 몇 사람의 전문가가 수동으로 하기에는 오랜 기간이 소요되므로, 음성 DB의 구축에 자동화 과정이 절실히 요구되고 있다.

5. 향후 연구 전망

이상에서 살펴 본 바와 같이 음성인식, 화자인식, 음성합성 등의 음성정보처리 기술에 있어서 수많은 기법들이 개발되어 왔으나 아직 이를 상용화하는 데에 있어서 극복해야 할 많은 문제들이 남아있는 실정이다. 음성인식에 있어서는 최근 다양한 잡음 환경 및 채널 왜곡 환경에서 강인한 음성인식을 위하여 여러 가지 기법들이 연구되고 있다. 특히 마이크 어레이를 이용하여 잡음을 제거하거나 특정 음성을 검출해내는 기법과 운율특징을 효과적으로 추출해 내는 기법 등의 연구가 전처리 기술로서 향후 주목을 받을 것으로 예상된다. 음향 모델링 분야에서는 기존의 HMM이 가지는 여러 가지 제약을 극복하기 위한 방법으로 machine learning 분야에서 최근 주목받고 있는 Bayesian network 기법을 모델링에 적용하는 방법 및 다언어 음향 모델링 기법 등이 주요한 연구분야로 부상할 것으로 예상된다. 언어 모델링 분야에서는 다영역/대어휘 대화체 음성인식을 위한 언어모델이나 멀티모달 상호작용을 위한 모델링 기법 등이 계속해서 주요 연구 주제로 지속될 것이다. 음성인식의 탐색 기술 중에서는 소형 단말기에서의 실시간 대어휘 음성인식을 위한 고속 탐색 기법, 다중 인식결과 후보로부터 semantic 및 pragmatic 정보를 이용하여 인식 성능을 향상시키는 방법, 운율정보를 탐색 공간에 통합하여 표현하고 탐색하는 기법 등이 주요 연구 분야로 계속 연구될 것이다. 음성인식의 후처리 과정으로서의 발화검증 기술 분야에서는 이질적인 특징벡터 및 운율정보 등 다른 knowledge source를 이용하여 검증에 적용하는 방식, 언어모델 등 high level 정보를 재이용하여 검증에 적용하는 방식, 가변어휘 음성인식에 적용 가능한 고정밀 발화검증 기법, 환경 및 화자 적응에 따른 발화검증 파라미터 동적 적응화 등에 대한 연구가 필요할 것으로 예상된다.

한편, 화자인식 분야에서는 기존의 모델링 제약을 극복하기 위한 기본적인 연구 주제로부터 초대규모 화자수에 대한 화자식별이나 화자확인을 위한 고속/고정밀 인증 기술, 잡음 및 채널 왜곡에 강인한 화자인식 기술, 문장 독립형 화자인식 기술, 고속 화자등록 기술 등에 대한 연구가 지속될 것으로 예상된다. 음성합성 분야에서는 현재 상용화되어 널리 활용되고 있는 코퍼스 기반의 대용량 DB 음성합성기를 발전시켜서 소형 단말기에서도 이와 같은 성능의 합성음질을 출력할 수 있는 합성기에 대한 연구와 다양한 음색을 실현할 수 있는 합성기술 및 감정표현까지도 가능한 음성합성기에 대한 연구가 향후 주요 연구 주제가 될 것이다.

6. 결 론

궁극적인 의미에서의 음성정보처리 즉, 사람의 능력에 버금가는 음성처리 기술을 개발하기 위해서는 결국 사람의 두뇌 속에서 음성언어가 어떻게 생성되고, 이것이 어떻게 발성기관을 동작시키는지를 발견해 내어야 하며, 또한 청각기관이 전달되어 온 음성신호를 어떻게 분석하고 이를 두뇌가 어떤 방법으로 이해하는지를 분석해야 한다. 이와 더불어 사람과 사람 사이의 대화에서는 음성뿐만 아니라 다른 정보 즉, 동작이나 표정 등과 같은 다른 modality를 함께 사용하므로 이를 이용하는 multi-modal interface에 대한 연구도 수행되어야 한다. 또한, 음성인식기가 어떤 특정 언어에만 잘 동작하지 않고 다국어 음성입력에 대해서도 우수한 성능을 갖도록 하기 위해서는 multi-lingual capability도 가져야 하므로 이에 대한 연구도 향후 지속적으로 수행되어야 한다.

본 고에서는 음성정보처리 기술의 동향 및 향후 과제를 개괄적으로 살펴보았다. 요약하면, 지난 30여년간 지속적인 기술의 발전이 이루어져 와서 현

재는 비록 제한된 분야이기는 하지만 상용화 시스템 및 서비스가 우리 주변에서 급속히 확산되고 있다고 할 수 있다. 하지만, 아직 극복해야 될 수많은 과제를 안고 있어서 상용 시스템 개발과 병행하여 핵심 요소기술에 대한 지속적인 연구가 절실한 형편이다. 또한, 이제 막 출시되고 있는 음성인식 기반 서비스를 이용할 때에도 상기한 여러 가지 제약들을 고려하여 이용하면 나름대로 큰 편리성을 얻을 수 있을 것으로 기대한다.

참고문헌

[1] 김회린, 이영직, "Voice Interface 및 인식," 정보처리학회지, 제5권, 제1호, pp. 42-48, 1998.

[2] 김형순, "음성정보처리기술의 현황과 전망," 전자공학회지, 제30권, 제7호, 2003.

[3] ETSI standard document, "Speech Processing, Transmission and Quality aspects(STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm," ETSI ES 201 108 V1.1.1(2000-02), Feb. 2000.

[4] ETSI standard document, "Speech Processing Transmission and Quality aspects; Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," ETSI ES 202 050 V1.1.1, 2002.

[5] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, 1993.

[6] B. H. Juang, S. Furui, et al., "Special Issue on Spoken Language Processing," in Proceedings of the IEEE, Aug., 2000.

[7] X. Huang, A. Acero and H. Hon, Spoken language Processing, Prentice Hall PTR, 2001.

저자약력

김 회 린

1984년 2월 한양대학교 전자공학과 (공학사)
1987년 2월 KAIST 전기및전자공학과 (공학석사)
1992년 2월 KAIST 전기및전자공학과 (공학박사)
1987년 10월~1999년 12월 한국전자통신연구원(ETRI) 네트워크연구소 선임연구원
1994년 6월~1995년 5월 일본 ATR연구소 방문연구원
2000년 1월~현재 한국정보통신대학교(ICU) 공학부 조교수
관심분야 : 음성인식, 화자인식, 음성/오디오 부호화, 오디오 인덱싱
이 메 일: hrkim@icu.ac.kr