특 집

고품질 다채널 한국어 음성합성 시스템 개발 동향

강 동 규*, 한 민 수**

*(주)코아보이스, **한국정보통신대학원대학교

I. 서 론

음성기술은 기계와의 대화를 가능하게 할 수 있는 가장 편리한 기술로 알려져 왔지만 아직까지 음성기술은 대부분 서비스 시스템 내에서 주로 보조적 기능을 담당해 왔었다. 최근 들어 음성기술을 이용하여 상용화에 성공한 사례가 나타나면서 각 분야에서 서비스의 다양화, 운용비 절감이라는 장점이 더욱 부각되어 적극적인 도입이 시도되고 있다.

상용화 예로서 최근 수년 전부터 준비해 왔던 서비스 시스템 구축을 완료하고 본격적인 서비스를 시작하고 있는 텔레매틱스 서비스가 있다. 이 서비스는 자동차 내에서 생활하는 시간이 증가하면서, 자동차 내에서 보다 편리한 서비스를 받고자 하는 욕구에서 시작된 서비스이다. 대도시의 복잡한 도심지에서 원하는 목적지까지 혼잡한 구간을 피하여 최단시간에 차량으로 이동할 수 있는 지능형 위치 안내 서비스, 차량고장 진단과 동시에 가장 가까운 서비스 센터까지 안내해주는 서비스, e-mail과 각종 공공정보 안내 서비스 등을 비롯한 갖가지 새로운 서비스들이 준비될 전망이다.

이 서비스에서 음성기술은 서비스의 시작과 처리된 결과를 고객에게 전달하는 필수적인 역할을 담당하고 있다. 또한 대부분 유료 서비스로 시행되고 있으며 그 가입자 수는 높은 증가 추세를 나타내고 있다.

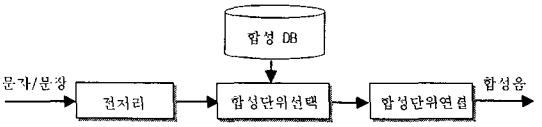
텔레매틱스에 필요한 음성기술은 자동차 환경에서의 음성인식과 고품질의 음성합성기술이다. 특히 음성합성의 경우 자동차 잡음과 운전에 집중한 상태에서도 정확하고도 편안하게 알아 들을 수 있을 정도의 자연성과 명료도가 요구된다.

음성합성은 사용자를 서비스하고자 하는 범주로 유도하거나 사용자가 원하는 정보를 전달하고 이해시키는 역할을 담당하고 있다. 본 고에서는 음성합성의 원리와 기술동향에 대해 논하고 최근 가장 많이 활용되고 있는 코퍼스기반의 음성합성 방법에 대하여 기술하고자 한다.

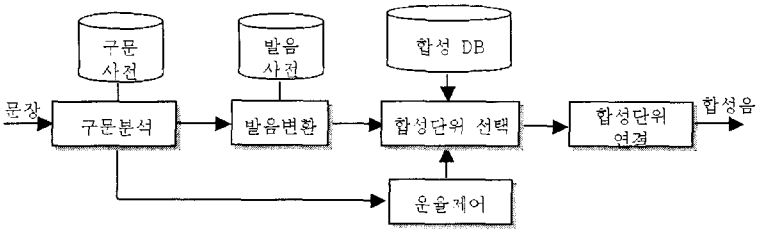
II. 음성합성의 개요

음성합성이란 글자, 문장, 숫자, 기호 등을 일반적으로 발성하는 형태의 음성으로 변환하는 것을 말한다. 글자를 표기할 때 한글의 경우 자모음을 이용하여 모든 글자를 구성할 수 있듯이 〈그림 1〉과 같이 음성합성에서도 자모음에 대한 음성을 미리 작성하여 음성사전을 구성한 다음, 입력되는 문자에 해당하는 음성을 음성사전에서 가져와 연결하는 것이 기본적인 방법이다. 일반적으로 음성사전에 등록된 음성을 합성단위라 하고 음성사전을 합성 DB라 한다.

간단한 음성합성기의 예로서, 일반적인 한글을



〈그림 1〉 음성 합성기의 기본 구조



〈그림 2〉 일반적인 합성기의 구조

표현하는 2,350여 가지의 단음절을 개별적으로 녹음하여 사전화 한 다음 입력되는 문자에 해당하는 단음절을 출력시키면 기본적인 음절단위의 음성합성기를 구성할 수 있다.

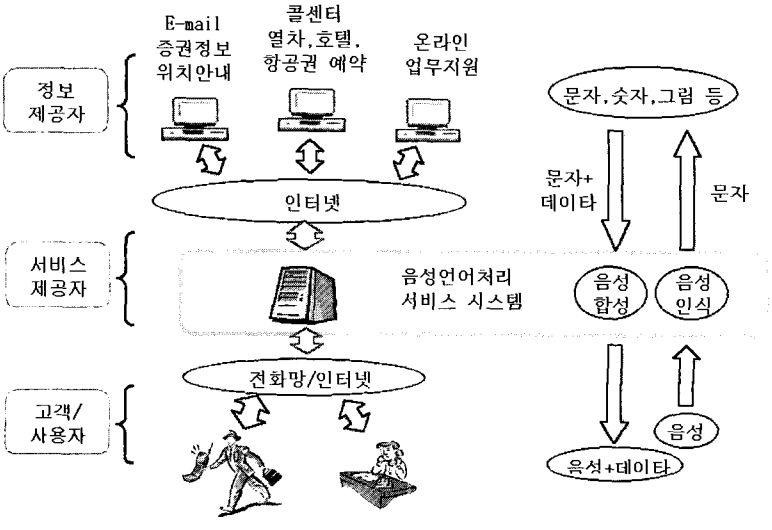
음절단위의 음성합성기는 텔레뱅킹과 같이 고유명사나 숫자 만을 발성하는 간단한 분야에서는 아직도 널리 이용되고 있는 방법이다. 음절단위 합성방법에서는 다음과 같은 부분에 대한 처리가 어려워 자연스런 문장단위의 합성음이 필요한 분야에서는 사용되지 못하고 있다.

- 음절간 음운현상
- 띄어쓰기, 끊어 읽기
- 문장에 따른 자연스런 억양

위의 사항을 처리할 수 있도록 하기 위한 방법으로서 합성단위를 음절단위에서 반음절, 음소, 반음소 등으로 세분화하고, 〈그림 2〉에서와 같이 입력문장에 대한 품사, 형태소 등을 분석하여 보다 정확한 발음변환을 수행하고, 문장에 따른 띄어쓰기, 끊어 읽기, 억양 등을 제어해야 보다 자연스런 합성음을 생성할 수 있다. 합성단위를 세분화하면 생성 가능한 종류가 기하급수적으로 증가하게 되고 이에 따라 합성 DB의 크기도 증가하게 된다.

III. 음성합성을 이용한 서비스

음성합성은 서비스 시스템에서 기존에 성우가 녹음한 음성데이타를 이용하거나 상담원이 응답하던 부분을 대신하는 것으로 서비스 내용이 수



〈그림 3〉 음성기술을 이용한 서비스 모델

시로 변경되거나 녹음할 내용이 많을 경우보다 효율적이다.

응용분야로서는 다음과 같은 분야가 있다.

- 위치안내, 교통상황 안내와 같은 텔레매틱스 서비스
- 쇼핑몰의 자동주문, 방송시간 안내
- 구내전화 자동교환, ARS, CTI, ITI, UMS
- E-book/audio-book, 각종 교육용 저작도구
- 고객안내, 지원, 관리 시스템의 자동화
- 휴대용 단말기에서 음성 인터페이스

음성합성을 이용한 일반적인 서비스 모델은 〈그림 3〉과 같이 정보 제공자가 공급한 내용을 서비스 업체가 전화망이나 인터넷을 통해 사용자에게 제공하는 형태이다.

Ⅳ. 국내 음성합성의 기술개발 동향

국내의 음성기술은 1990년대 말까지는 학계 및 연구기관을 중심으로 기술개발이 이루어졌으나 그 이후부터는 업계에서도 활발한 기술개발이 진행되었다. 1990년대 말에 고품질 음성합성 제품이 출시되면서 국내외 업체들의 기술개발 경쟁이 가속화되어 최근에는 특정 서비스영역에서 자연음 수준의 합성품질을 낼 수 있는 수준까지 도달하였다.

최근에 많이 사용되고 있는 코퍼스기반 음성합성기는 국내에서 상용화에 성공하여 세계 정상급의 기술수준을 확보하게 되자 이제는 국외의 대부분 업체에서도 이 방식을 채택하고 있다.

국내에서 상용화되었던 음성합성기술 변화는 합성 DB를 구성하는 방식에 따라 1세대인 LPC 계열 방식, 2세대로는 PSOLA 방식, 3세대의 코퍼스 방식의 세 가지로 구분할 수 있으며, 초기 기술개발은 모두 한국전자통신연구원에서 이루어졌으며 그 후 각계에서 여러가지 변형된 방법들이 개발되었다.

1. 제 1 세대 합성기

1980년 말경에는 메모리의 한계점을 극복하기 어려운 관계로 합성 DB를 압축하여 구축한 다음 합성 최종단계에서 복원하여 합성음을 생성하는 방법을 이용하였다. 합성 DB의 크기를 줄이기 위한 방법으로서는 음성압축방법이 활용되었으며 압축방법에 따라 다음과 같은 방법으로 분류된다.

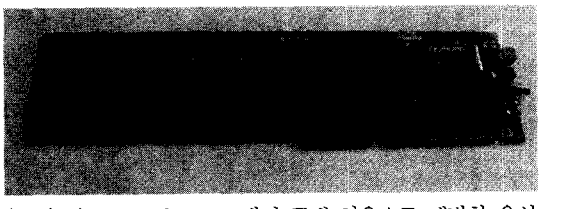
- LPC 계열 합성방법
- Formant 합성방법

국내에서는 주로 LPC 계열 기반의 합성기가 개발되었으며, 소형 합성기에서는 아직도 널리 이용되고 있는 방법이다. Formant 합성방법은 학계나 연구기관에서 연구되었으나 합성품질이 낮아 상용화는 이루어지지 않았다.

2. 제 2 세대 합성기

음성을 압축하여 복원할 경우 일반적인 연속음성에서는 고품질이 유지 되지만 합성에 사용되는 단위는 매우 짧으므로 선행되는 음성에 대한 정보가 부족하여 비교적 품질이 낮은 것이 일반적이다. 따라서 1990년대 중반에는 원래 음성에 포함된 운율을 PSOLA 기법에 의해 변경하여 합성하는 방법을 연구되었다. 이 방법에 의한 합성음은 화자의 음색이 살아 있을 정도의 명료도를 확보할 수 있는 방법으로써 파라미터 합성방법에 비하여 훨씬 높은 품질의 합성음을 얻을 수 있었다.

이 방법은 각 합성단위에 대해 복수개의 후보를 확보하면 품질을 높일 수 있어서 여러가지의 변형된 제품들이 출시되었다. PDA용 합성기와



〈그림 4〉 1991년 ETRI에서 국내 처음으로 개발한 음성합성기(LSP방식, 한국종합전시장(COEX)에서 91'장애인 재활용품전에 출품)

같이 중형의 합성 DB크기를 요구하는 분야에서는 지금도 많이 활용되는 방법이다.

3. 제3세대 합성기

1, 2세대의 합성 방식은 메모리 크기를 줄이기 위해 일반적으로 합성단위별 후보가 1개 이거나 수개에 불과하므로 합성시 운율변경이 필연적으로 수반된다. 인간의 음성은 억양의 변화에 따라 성도길이가 변화하여 같은 음소라 할지라도 억양의 높낮이에 따라 스펙트럼 구조가 다르게 나타나므로 일정한 음성에 대해 스펙트럼의 변경없이 억양을 과도하게 변경하면 매우 거북스런 음성이 생성되어 합성음의 자연성이 저하되며, 억양의 높 낮이에 따라 발성되는 지속시간 또한 다르므로 자연성은 점점 저하되는 단점이 있다. 이러한 문제점을 극복하기 위해 연구된 방법이 코퍼스기반 합성방법이며 합성단위별로 다양한 운율을 포함한 후보를 수십개에서 수백개 이상 확보하여 이들 후보간의 상호연결이 적절한 경로를 선택하면 고품질의 합성음을 얻을 수 있다.

코퍼스기반 합성방법은 최근에 가장 많이 활용되고 있는 방법으로서 다음과 같은 장단점이 있다.

⊙ 장점

- 고품질의 합성음 생성 가능
- 다양한 운율에 의한 자연 운율 재현 가능
- 서비스 영역별 학습 및 튜닝에 의해 자연음 수준의 합성음 생성 가능

⊙ 단점

- 합성 DB에 포함되지 않은 어휘를 합성할 경우 음질 저하
- 자연운율 제어 및 안정성 유지가 어려움
- 피치 및 속도 제어가 어려움.
- 합성 DB의 크기가 크고 처리 속도가 느림
- 음색 변경이 어려움
- 합성 DB 제작에 많은 시간과 비용이 소요됨

V. 코퍼스 기반 음성 합성기의 개발 단계

1. 음성 DB 구축 단계

음성 DB구축은 〈그림 5〉와 같이 문장설계, 화자선정, 녹음, 전사오류수정 과정으로 구분할 수 있다.

⊙ 음소균형 문장 설계

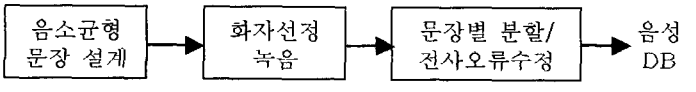
고품질의 합성 DB 구축을 위한 문장은 먼저 풍부한 음운환경이 포함되어야 하고 각 음소별로는 다양한 운율이 담긴 음성데이터가 필요하다. 이를 위해서는 대용량의 말뭉치 데이터를 발음변환하여 음운환경의 분포를 고려한 문장을 추출해야 한다.[2] 문장 DB는 크게 나누어 일반영역에서 추출한 문장과 서비스 영역별로 추출된 문장으로 구분된다. 서비스 영역에는 일기예보, 교통정보, 증권정보, 위치정보, 관광정보, 의료정보, 법률정보, 첨단 과학 기술정보, 부동산정보 등이 있으며 이는 특정영역의 서비스에서 보다 자연스런 합성음을 얻기 위한 것이다. 일반영역의 경우 다양한 음운환경을 고려하여 추출하며 서비스 영역별 문장은 다양한 운율에 초점을 맞추어 추출하는 것이 바람직하다.

하나의 합성단위는 비교적 많은 후보로 구성되는 것이 바람직하므로 청취에 의해 구분이 어려운 유사음의 경우에는 통합하여 보다 많은 후보를 확보하는 것이 합성음의 안정성을 유지할 수 있다.

⊙ 화자선정 및 녹음

구축된 음소균형 문장을 이용하여 음성합성에 필요한 문장을 녹음하기 위해서는 비교적 발음훈련이 잘되어 있고, 서비스에 적합한 음색을 내는 전문 성우나 아나운서의 도움으로 얻을 수 있다.

먼저 개략적인 발음과 서비스에 적합한 선호도



〈그림 5〉 음성 DB 구축 단계

를 갖는 화자를 선택하기 위해 수십 명의 후보 화자가 낭독한 음성데이터를 이용하여 다수의 청취자가 후보를 선정한다. 선정 시 고려사항은 운율의 안정감, 발음의 정확성, 음성의 명료도, 음색의 선호도 등이다.

운율의 안정감은 화자가 유사한 문장을 다시 읽었을 경우 유사한 운율을 낼 수 있는 화자가 안정된 운율특성 재현에 매우 유리하며, 발성속도가 일정해야 합성음이 안정될 수 있다.

발음의 정확성은 몇 가지 문장으로 판단하기 어렵기 때문에 최소한 난이도 있는 문장으로 수백 문장을 녹음해야 보다 정확한 판단이 가능하다. 음성의 명료도는 화자가 발성시 활기있고 힘차게 발음할 수 있어야 많은 문장을 녹음할 경우 음색이 변하지 않고 명료한 음성을 얻을 수 있다. 또한 전화망에서 사용할 경우 전화선로에 의한 손실로 인하여 명료도가 낮아지지 않는 음성이 적합하다.

음색의 경우, 서비스 용도에 따라 약간씩 다를 수 있으며 일반적으로 상냥한 말씨를 선호하는 경향이 있다. 그러나 이 경우에는 운율의 변화폭이 크므로 합성음의 안정성이 저하될 수 있다.

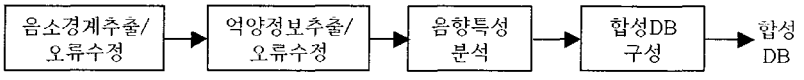
2. 합성 DB 구축 단계

합성 DB 구축 단계는 〈그림 6〉과 같이 음소경계추출 및 오류수정, 억양정보 추출 및 오류수정, 각 합성단위 후보들에 대한 음향특성 분석, 합성 DB 구성 과정으로 구분될 수 있다.

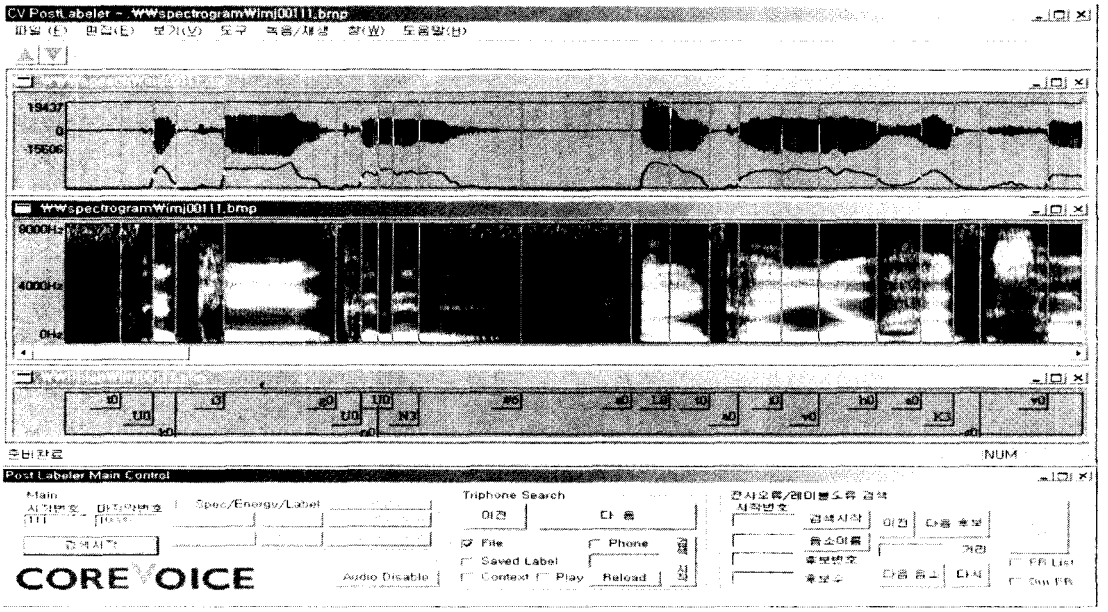
⊙ 음소경계 추출 및 수정

정확한 음소경계 추출은 합성단위 간 스펙트럼의 연속성, 지속시간, 강세정보 모델링에 직접적인 영향을 줌으로 합성 DB 구축에서 가장 중요한 부분 중에 하나이다.

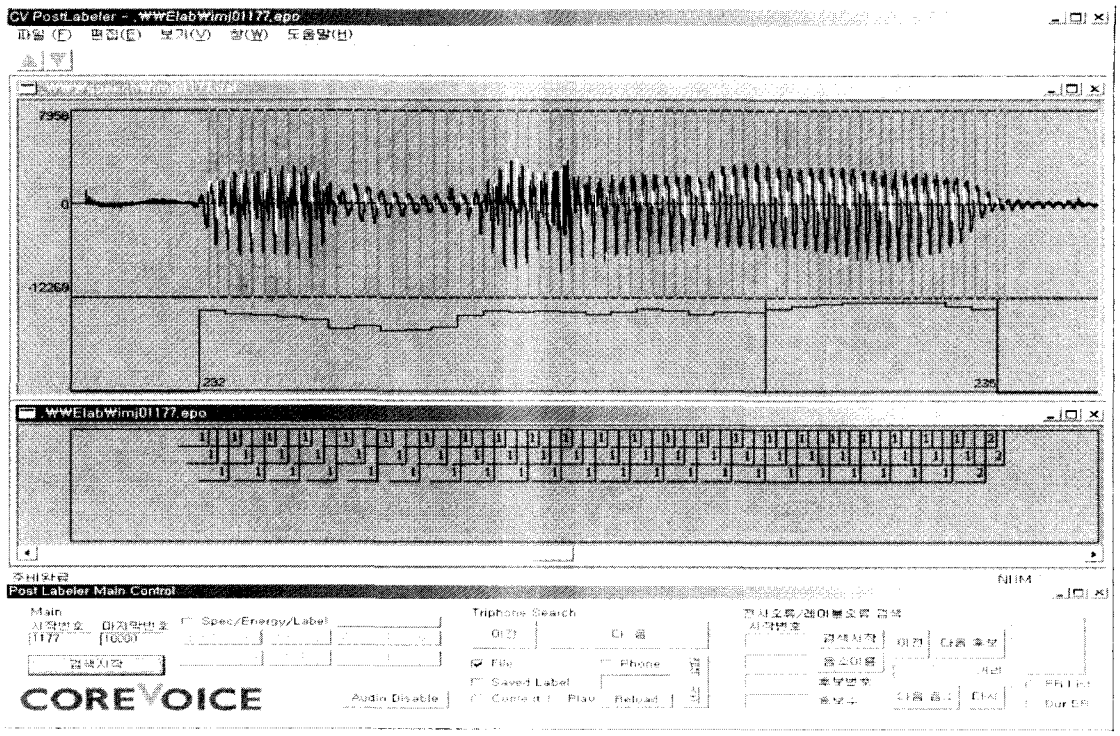
일반적으로 코퍼스 기반의 합성 DB를 구축하는 경우에는 대용량이므로 먼저 음성인식기를 이



〈그림 6〉 합성 DB구축 단계



〈그림 7〉 음소경계 수정 도구의 예



〈그림 8〉 억양정보 추출 도구의 예

용하여 개략적인 음소경계를 추출한 다음 수동으로 보다 정밀한 음소경계를 추출하는 것이 일반적인 방법이다. 수동 음소경계 수정은 많은 시간이 소요되므로 제작기간을 단축하기 위하여 훈련된 여러 사람들이 동시에 음소경계 수정작업을 진행한다. 이러한 경우에는 음소분할 경계에 대한 기준이 서로 다르므로 일관성을 유지하기 어려워 합성 품질이 저하된다. 이를 극복하기 위해서는 음소분할을 수행하는 사람들이 일정한 기준을 가질 수 있도록 주기적인 훈련과 교육이 필수적으로 병행되어야 한다.

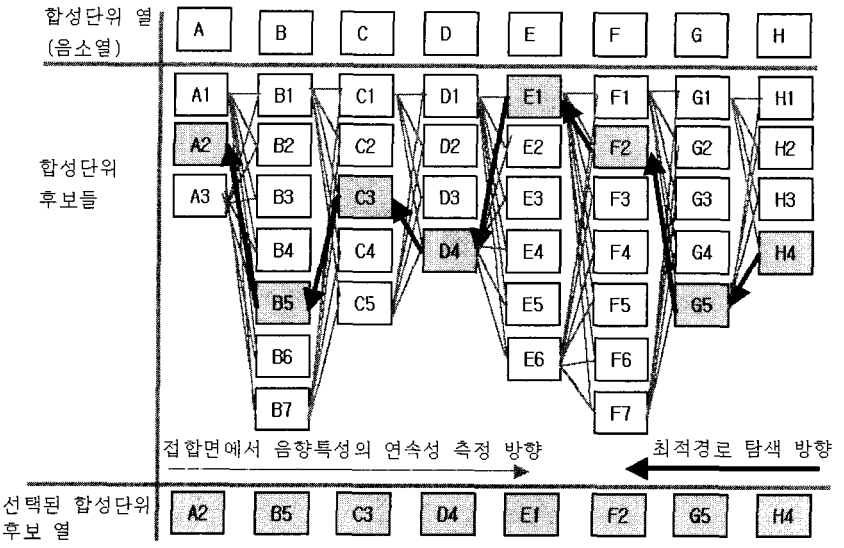
또한 음소분할 도구 역시 일관성을 유지할 수 있도록 설계되어야 하고, 음성청취에 의해 음소경계를 판단할 경우 다양한 방법으로 청취할 수 있어야 보다 정확한 경계를 추출할 수 있다.

⊙ 억양정보 추출 및 오류 수정

코퍼스 기반의 합성음 품질은 다양한 음운환경에 대한 합성단위의 종류와 각 후보들에 대한 풍부한 운율후보가 갖추어질 경우 높은 품질의 합성음을 얻을 수 있다. 그러나 이러한 풍부한 운율후보가 존재하더라도 이에 대한 정확한 특징 추출이 선행되지 않는다면 원래화자의 자연운율을 재현할 수 없으므로 합성음의 자연성과 안정성은 크게 저하되게 된다.[3]

운율특성은 그 중요도에 따라 억양, 지속시간, 강세 특성으로 나타나며 지속시간과 강세특성은 음소경계 분할시 정보를 추출할 수 있으나 억양정보의 경우 별도의 처리가 필요하다.[4] 억양정보를 보다 편리하게 추출하기 위해서는 문장 녹음시 EGG 신호를 동시에 녹음하면 보다 정확하면서도 편리하게 억양정보를 추출할 수 있다. 그러나 이 경우 화자의 목 부분에 센서를 부착해야 하므로 화자가 발성시 약간의 불편이 수반되어 발성에 정확도가 저하될 수 있다.

EGG를 이용하더라도 실제 운율과 추출된 운율정보와 일치하지 않는 경우가 있다. 즉 EGG에 의해 추출된 억양정보의 정확도는 90~95% 정



〈그림 9〉 최적 합성단위 후보 경로 선택 방법

도이며 나머지 5~10%의 오류는 훈련된 사람에 의해 오류를 수정해야 한다. 대부분의 운율에 있어서 억양정보는 매우 중요한 정보로서 약간의 오류가 발생하여도 부자연스럽게 느껴지므로 반드시 검증을 해야 할 부분이다.

3. 음성합성기 개발 단계

음성합성기는 〈그림 2〉에서와 같이 구문분석, 발음변환, 최적합성단위 후보 선택, 합성음 연결 과정으로 구성된다. 구문분석은 품사 및 형태소를 분석하여 보다 정확한 발음변환, 띄어쓰기, 끊어읽기, 운율추정 등에 필요한 정보를 추출한다. 발음변환에 의해 생성된 문자열의 자모에 대한 최적의 합성단위후보를 선택하여 연결하면 입력된 문장에 대한 음성을 생성할 수 있다.

최적 합성단위 후보 선택과정은 일반적으로 〈그림 9〉와 같이 합성단위 후보들이 포함하고 있는 음향특성의 연속성을 비터비 탐색(Viterbi search)에 의해 최적후보를 선택하고 있다. 그러므로 합성단위에 대한 후보수가 증가하면 계산량이 기하급수적으로 증가하게 된다. 합성음질을 높이기 위해서는 후보수를 늘려야 하지만 동시에 많은 채널을 지원하기 위해서는 최대 후보수를 제한해야 한다.

일반적으로 합성기에서 음성을 합성할 때 최적 합성단위를 선택하는 과정에서 소요되는 계산량은 전체 계산량의 90% 이상을 차지하므로 고속 탐색 기술 및 합성 DB의 최적화가 필수적으로 요구된다.

구문분석, 발음변환과정은 합성방식에 관계없이 적용할 수 있으나 코퍼스기반 합성방식의 합성기는 합성단위에 포함된 운율을 가급적 변경하지 않고 입력문장에 가장 적절한 운율을 포함한 합성단위 후보를 선택하는 최적 합성단위 후보 선택과정이 포함된 것이 다른 방식과 구별되는 점이다.

VI. 음성합성의 향후과제

음성합성을 이용하는 서비스에서 요구하는 합성품질은 자연음 수준이지만 대부분의 합성기술이 아직 이에 미치지는 못하고 있으나 특정 서비스 분야에 대해서는 자연음 수준에 도달하고 있다. 합성품질을 높이기 위해서는 다양한 음운환경과 운율이 포함되도록 해야 하고 이를 위해서는 합성 DB의 크기를 확대해야 한다. 합성 DB

의 크기를 늘리는 것은 많은 인력, 시간, 비용이 수반되므로 어느 정도의 한계점을 가지고 있다.

향후 음성합성에 있어서 가장 시급한 것은 합성DB 구축시 음소분할과 억양정보 추출의 오류 수정을 최소화할 수 있는 기술개발, 한정된 합성 DB에서 필요한 합성단위를 자동 생성할 수 있는 기술 개발이 요구되고 있다.

또한 합성 DB 구축에 대한 문제를 해결하기 위한 또 다른 방법으로써 기존에 구축된 화자의 음색을 변경하여 새로운 화자를 생성하여 개발기간과 비용을 획기적으로 절감할 수 있는 기술개발이 요구되고 있으며 이는 비용절감에 따른 음성시장 활성화에도 큰 도움이 될 것으로 기대된다.

음성시장에서는 대화체 합성기에 대한 필요성이 대두되고 있어서 이에 대한 운율모델링 및 제어기술 개발이 요구되고 있다. 대화체 운율은 낭독체에 비하여 변화폭이 매우 크므로 보다 정교한 처리 기술이 필요한 분야이다.

Ⅶ. 결 론

본 고에서는 음성합성의 기본원리, 활용 가능한 서비스, 기술개발 동향 그리고 최근에 가장 많이 이용되고 있는 코퍼스기반 음성합성 시스템의 개발에 대하여 기술하였다.

최근 콜센터를 중심으로 음성합성 시스템의 도입이 활발히 진행되고 있는 것은 합성품질이 비교적 높아서 고객들로부터 불평의 소지가 적을 것이라는 판단이 있기 때문이다. 합성품질을 모든 분야에 대해 자연음 수준을 유지하기는 어렵지만 최근의 상용화가 적극적으로 추진되고 있는 것은 특정 서비스 분야에 대해 합성품질을 최적화하여 자연음 수준까지 끌어 올릴 수 있기 때문인 것으로 분석된다. 이는 현재 해결하기 어려운 기술부분에 대한 적극적인 대처로 보이며 다음 단계의 기술로 발전하기 위한 계기가 될 수 있을 것이다.

과거에 음성기술이 인간을 대신하여 많은 것을 해결할 수 있는 기술로 알려졌다가 고객들이 음성기술로부터 눈길을 돌렸던 것은 음성기술이 실질적으로 사용자에게 편리함이나 이익을 주지 못했기 때문이다. 다른 기술들과 마찬가지로 음성기술도 관련시장이 활성화되면 재투자로 인하여 기술개발이 활발해지고, 이로 인해 현재의 기술적인 문제점들은 단기간 내에 대부분이 해결될 수 있을 것이다. 그러므로, 현재의 기술을 이용하여 제품화하고 문제점에 대해서는 적극적인 대응방안으로 상용화하여 보다 많은 성공사례가 나올 때 음성기술은 새로운 시대를 맞이할 수 있을 것이다.

참 고 문 헌

〔1〕 Nick Campbell, Alan W. Black, *Progress in Speech Synthesis*, pp.279-292, springer, 1996

〔2〕 Steve Young and Gerrit Bloothooft, *Corpus-based methods in Language and Speech Processing*, Kluwer Academic Publishers, 1997

〔3〕 P. Price et al., "The use of prosody in syntatic disambiguatin," J.Acoust. Soc. Amer., vol. 90, pp.2956-2970, 1991

〔4〕 Allen J., Hunnicutt S., and Klatt D. *From text to speech : the MITalk system*, MIT Press, Cambridge, Massachusetts, 1987

〔5〕 Hsin-min Wang, *Statistical Analysis of Mandarin Acoustic Units and Automatic Extraction of Phonetically Rich Sentences Based Upon a Very Large Chinese Text Corpus*, Computational Linguistics and Chinese Language Processing, vol. 3 no. 2, pp.93-114, August 1998.

〔6〕 Black, A., and Taylor, P., "Automatically clustering similar units for unit selection in speech synthesis," *Proc.*

Eurospeech, pp.601~604, Sep 1997
〔7〕 Donovan, R.E., *Trainable Speech Synthesis*, PhD. Thesis, Cambridge University Engineering Department, 1996

저 자 소 개

강 동 규

1989년 2월 호서대학교 전자공학과(학사), 1991년 2월 호서대학교 대학원 전자공학과(석사), 1991년 2월~2000년 3월 : 한국전자통신연구원 음성언어팀 선임연구원, 2000년 5월~현재 : (주)코아보이스 대표이사, 〈주관심 분야 : 음성합성, 음성분석〉

한 민 수

1979년 2월 서울대학교 전기공학과(학사), 1981년 2월 서울대학교 대학원 전기공학과(석사), 1989년 12월 Univ. of Florida 전기전자공학과(박사), 1990년 2월~1997년 12월 : 한국전자통신연구원 책임연구원, 1998년 1월~현재 : 한국정보통신대학원대학교 부교수, 〈주관심 분야 : 음성합성, 음성분석, 잡음제거, 3-D 음향〉