내장형 음성합성 기술 동향 및 사례

The Recent Trends and Applications of Embedded TTS Technologies

임베디드 S/W 기술 동향 특집

김종진 (J.J. Kim) 자동통역연구팀 선임연구원
김정세 (J.S. Kim) 자동통역연구팀 선임연구원
김상훈 (S.H. Kim) 자동통역연구팀 책임연구원
박 준 (J. Park) 자동통역연구팀 팀장

목 차

* 본 연구는 정보통신부 및 정보통신연구진흥원의 "신성장동력산업용 대용량 대화형 내장처리 음성인터페이스 기술" 개발 사업의 일환으로 수행하였음.[2006-S-036-01]

음성합성 기술은 1990년대 중반 음편접합 방법론이 출현하면서 괄목한 만한 기술적 발전을 이루어, 2000년 전후에는 전화망을 이용한 ARS, VMS, UMS 서비스를 중심으로 폭넓게 사용되면서 일반 사용자들에게 매우 친숙한 서비스를 제공하여 왔다. 그러나 최근 텔레포니 기반의 음성 기술 시장은 기업고객 위주로 그 성장이 더딘 반면, 지능형 로봇, 텔레매틱스, 홈네트워크, 차세대 PC와 같은 전략적 국가 신성장동력 산업분야나 MP3 플레이어, 휴대폰, PMP 단말기, 휴대용 단말기와 같은 임베디드 분야가 음성 기술의 새로운 시장으로 주목을 받고 있다. 임베디드 분야에서 요구하는 음성 기술은 기존 서버급 시스템에서 운영되었던 기술과는 상당히 다른 기술 특성을 가지고 있다. 이에 본 고에서는 음성 기술 중 특히 음성합성 기술에 관한 임베디드 분야의 요구사항을 고찰하고, 이를 해결하기 위한 최근의 기술적 발전 동향 및 응용 사례에 대해서 기술하고자 한다.

I. 서론

음성합성이란 컴퓨터나 어떤 장치가 사용자에게 제공하고 싶은 정보를 사람들이 들을 수 있는 말의 형태로 제공할 수 있도록 기능을 제공하는 기술 또는 그 기술을 구현한 소프트웨어/하드웨어를 의미한다. 음성합성 기술은 자동차 운전 상황이나 매우 협소한 장소에서의 작업과 같이 시각을 운전이나 작업 외 다른 곳에 집중하기 어려운 상황이나, 유선전화망과 같이 디스플레이가 없는 정보통신 기기, 시각장애인이나 약시자와 같이 시각적인 핸디캡을 가진 사용자를 위한 매우 효율적인 정보 제공수단이다. 또한 최근에는 지능형 로봇과 같이 움직이는 로봇과 사용자 사이의 효율적인 정보 송수신 수단으로 음성인식 및 음성합성 기술이 사용되고 있다.

음성합성 기술은 (그림 1)과 같이 미리 특정한 안내 멘트를 특정 칩이나 저장장치에 저장하여 놓았다가 적절한 시점에서 녹음된 내용을 재생하여 사용자에게 음성 안내 기능을 제공하는 간단한 방법부터 휴대폰에서 단순메시지 읽기, 텔레매틱스 단말기에서 길안내 서비스 및 실시간 교통정보 등의 음성안내를 위한 임의의 텍스트가 주어지면 이를 합성해 낼 수 있는 무제한 음성합성 기술까지 그 기술적 스펙트럼이 매우 넓다.

음성합성 기술은 응용 서비스나 응용 제품에 따라 엘리베이터의 안내방송, 말하는 전기밥솥, 인사하는 인형과 같이 단순 녹음 및 재생 칩을 이용한 방법부터 시각장애인을 위한 전자책을 읽어주는 소프트웨어까지 고유의 영역에서 그 역할을 수행하고 있다. 그러나, 최근 각각의 방법론이 사용되고 있는 시장의 규모가 변하고 있다. 예를 들어, 음성합성 기술이 사람의 자연음과 유사한 합성음을 생성하기 요원했던—일부 제한된 범위를 제외하고—1990년대 초중반에는 음성출력 서비스는 대부분 미리 녹음된 안내 멘트를 재생하여 주는 서비스가 주를 이루었다. 그러나 2000년대 전후로 음편접합 기술을 이용한 무제한 음성합성 기술의 발달로 임의의 문장에 대한 상당히 자연스러운 음성합성 기술이 발달함에 따라 주로 전화망을 이용한 ARS, VMS, UMS 서비스 시장 등이 활성화되면서 서버형 음성합성 기술이 호황기를 누렸다. 이런 서버형 음성합성 기술은 원격지 서버의 고성능 컴퓨팅 장치 및 거대 메모리를 활용함으로써 매우 고품질의 합성음을 생성할 수 있었다.

그러다 2005년을 전후로 핸드폰, 텔레매틱스 단말기, MP3P, PMP, PDA 등과 같은 휴대용 단말기들이 대거 등장하고 일반인에게 빠른 속도로 보급되면서 이제는 전화망을 통한 서비스가 아닌 개인이 휴대한 핸드폰이나 텔레매틱스 단말기, PDA에 내장되어 합성음을 생성할 수 있는 다양한 서비스가 개발되면서 음성합성 기술은 임베디드 단말기로 표현되는 개인 휴대용 단말기 내부에 내장되어 음성합성을 수행할 수 있는 내장형 음성합성 기술이 시장에서 요구되고 있다. 그러나, 내장형 음성합성 기술은 서버형에 비해 상대적으로 매우 제한된 하드웨어

<자료>: http://www.voiceware.co.kr/

(그림 1) 음성합성 기술의 응용 분야

환경에서 동작하면서도 서버형과 같은 고품질의 합성음을 생성할 수 있는 기능을 시장에서 요구받고 있다.

본 고에서는 이러한 2005년 전후로 시작된 임베디드 응용 시장에서 요구하는 내장형 음성합성 기술에 대한 요구사항을 상술하고, 음성합성 기술별로 이러한 요구사항을 만족시킬 수 있는 방법에 대한 최신 경향을 소개하고자 한다. 또한 이러한 내장형 음성합성 기술의 일환으로 최근 본 연구팀에서 "신성장동력산업용 대용량 대화형 내장처리 음성인터페이스 기술" 개발 과제의 일환으로 개발중인 HMM 기반 소용량 대화체 음성합성 기술에 대해서 소개하고자 한다.

Ⅱ. 내장형 음성합성 요구사항

서론에서도 간략하게 소개한 바와 같이, 음성기술 시장의 가장 큰 변화는 텔레포니 기반 서버형 음성합성 기술을 응용한 시장의 성장은 더딘 반면 임베디드 단말기에 내장되어 사용되는 내장형 음성합성 수요는 폭발적으로 증가하고 있다는 점이다. 본 절에서는 내장형 음성합성 기술을 이용하고자 하는 응용분야별 요구사항의 특징을 기술한다.

1. 고품질 무제한 음성합성 기술

음성합성 기술에 대한 사용자의 요구사항은 항상 고품질의 합성음이다. 기술적 방법론이나 하드웨어 및 소프트웨어의 제약조건 등은 음성합성 시스템을 개발하는 기술적 측면이며 사용자는 이러한 하드웨어나 소프트웨어 조건, 또는 사용자가 구매하는 단말기의 가격에 무관하게 항상 고품질의 합성음을 요구한다는 점을 먼저 기억해야 한다.

최근 텔레매틱스 산업의 뚜렷한 변화는 기존 내비게이션 단말기로 대표되는 길안내 서비스에서 각종 재난정보 및 실시간 교통정보를 제공하는 지능형 교통 정보 제공 분야로 발전해가고 있는 점이다. 또한 제주도와 같은 관광 특구에서는 텔레매틱스 단말기를 통해 관광지의 안내, 관광지의 소개 및 상품안내 기능까지 서비스를 제공하고자 한다. 사용자들은 또한 텔레매틱스 단말기를 교통정보 제공 기능뿐 아니라 infortainment 수단으로 사용하고자 하는 요구가 뚜렷하다. 그래서 단말기에서 교통정보뿐만 아니라 음악 및 동영상 재생, DMB 시청 등 엔터테인먼트 기능과 WiBro와 같은 휴대 인터넷 기술과 연동하여 움직이는 사무실을 실현하고 any where, any time 정보통신 서비스를 제공해주기를 요구하고 있다.

기존 내비게이션 단말기에서 제공되는 길안내 서비스에서는 "좌회전 하십시오", "전방 사고 다발지역입니다"와 같은 고정된 수십에서 많게는 백여 개의 제한된 문장패턴만으로도 충분히 안내하고자 하는 정보를 음성으로 제공할 수 있었다. 그러나, 실시간 교통정보 및 재난정보, 관광지 안내, 휴대 인터넷을 통한 웹 서핑 및 뉴스 읽기, 이메일 읽기 등을 수행하기 위해서는 단말기 내에 임의의 어떤 문장이라도 합성할 수 있는 무제한 음성합성 기술의 탑재가 요구된다.

기술적인 측면에서는 무제한 음성합성의 경우 합성음의 음질과 메모리 및 CPU 연산량은 비례한다. 즉, 고품질의 합성음을 생성하기 위해서는 많은 양의 음성정보를 저장할 수 있는 메모리와 음성합성을 수행하기 위한 많은 연산량이 요구된다. 다시 역으로 제한된 메모리와 연산량을 사용하면서 고품질의 합성음을 생성하는 것은 기술적으로 매우 어려운 기술이다.

또한 사용자들은 기존 내비게이션 단말기에서 제공되는 길안내 음성멘트에 익숙해져 있어, 텔레매틱스 단말기에서 제공되는 재난정보나 실시간 교통정보, 관광지 안내 멘트 등도 동일한 수준의 고품질의 합성음을 요구한다. 그러나 내비게이션 단말기의 길안내 멘트 등은 대부분 녹음된 문장 또는 단어들의 조합을 이용하여 합성음을 생성하며, 그 품질은 거의 사람의 음성안내 목소리와 구별하기 어려울 정도로 고품질임을 고려하면, 제한된 하드웨어 규격에서 고품질의 무제한 음성합성 기술의 개발은 분명 도전적 과제이다.

2. 소용량 및 다국어 음성합성 기술

휴대폰 단말기의 경우 카메라폰, MP3폰 등의 등장으로 휴대폰 내 메모리 양이 증가되고 있으나 이는 사용자의 데이터를 저장하기 위한 메모리이며, 오히려 휴대폰의 기능이 더욱 복잡해지고 또한 최근 저가폰 시장의 활성화 등으로 휴대폰 내 응용프로그램에서 사용할 수 있는 메모리 및 CPU 점유율은 오히려 더 적어진 추세이다. 또한 휴대폰에서 요구하는 단문메시지 읽기용 내장형 음성합성 시스템의 경우 단일 언어뿐만 아니라 <표 1>에서 보는 바와 같이 다국어의 지원을 필수로 하고 있다. 특히, 유럽지역에서 판매되는 휴대폰의 경우에는 유럽지역의 지역적 특징으로 인해 하나의 단말기에 수 개에서 십여 개의 언어를 지원해야 하는 상황이다. 최근 뉘앙스의 경우에는 유럽피안 폰에 대해 10여 개 언어를 지원 가능한 음성합성 기능을 내장하고 있으며 엔진의 크기는 500k 바이트이고, 각 언어별 언어사전 500k 바이트, 음성데이터 500k 바이트 크기의 소용량 음성합성 시스템을 내장하였다고 밝히고 있다[1].

3. 표준의 지원

최근 음성기술은 VoiceXML과 SSML의 지원을 기본적으로 요구하고 있다. VoiceXML의 지원은 음성기술을 이용한 서비스나 응용제품의 개발자가 특정 음성기술 벤더에 종속되지 않고 더 좋은 품질의 음성기술 제품을 언제든 손쉽게 채택하게 할 수 있다는 장점이 있기 때문이며, SSML의 경우에는 응용 서비스 및 제품 개발자가 사용되는 음성합성 시스템의 성능을 극대화 할 수 있기 때문이다. 또한 다음 절에서 기술하겠지만, 최근의 음성합성 기술의

〈표 1〉 상용 내장형 TTS의 기술규격 예

Architecture	Single Client
Simultaneous Channels	Single Channel
Memory Requirements	RAM: from 2.5MB ROM: from 3.5MB
Type of Technology	Unit selection, concatenative
Sampling Rate	8/16/22kHz
CPU Requirements	Xscale, ARM9, Strongarm, X86, SH4, Motorola PowerPC
Platforms	Windows Mobile 5.0 PPC & SP, Pocket PC 2003, CE 5.0, CE.NET 4.2, Windows XP Embedded and TabletPC ed., VXWorks, Linux, Symbian OS Series 60
Interfaces	Loquendo API(C/C++ and compact .NET Framework) SAPI
Supported Languages	U.S. and U.K. English, Castilian, Catalan, Dutch, French, German, Greek, Italian, Polish, Portuguese, Swedish, American Spanish, Argentinean, Brazilian, Chilean, Mexican, and Mandarin Chinese
Standards Supported	SSML(Speech Synthesis Markup Language)
Key Features	• Pronunciation lexicon – for user definable pronunciation(acronyms, foreign names, etc.) • Mixed Language Capability • Audio Mixer • Dynamic switching between multiple voices • E-mail Preprocessing • Flexible voice control – for creating special effects, modifying speech rate and pitch • Customized voices – for extending corporate image and branding through unique voices • Support of the phonetic formats: PhonoMultiNet used by TeleAtlas®, NT-SAMPA used by Navteq™ • Custom lexicon and database for Automotive Quality/footprint trade-off flexibility according to user requirements • Expressive TTS

<자료:> http://www.loquendo.com/

활용은 이전과 달리 단순 뉴스 읽기와 같은 단방향 서비스가 아닌 지능형 로봇과 같은 양방향 의사교환 수단으로 활용되는 경향이 뚜렷하며, 이 때 지능형 로봇의 경우처럼 사용자의 의도나 상황에 맞는 적합한 모드의 음성합성을 수행하기 위해 상황에 맞는 합성음을 생성하기 위한 부가적인 정보를 합성기에 전달할 수 있고, 합성기에서는 이를 처리할 수 있기를 원하기 때문이다. 또한 소용량 내장형 음성합성 기술은 사용할 수 있는 메모리나 CPU 점유율이 제한되어 있으므로 모든 것을 합성엔진 내부에서 예측하여 생성하기 위해서는 많은 연산량이 필요하므로 이를 외부에서 제공할 수 있도록 하기 위함이다. (그림 2)는 합성할 텍스트를 SSML을 이용해 작성한 예이다.

또한 전자책과 같은 멀티미디어 토킹북을 위한 마크업 언어 표준 중에 하나인 DAISY 컨소시엄 표준과 같은 전자책의 정보나 텍스트를 태깅하는 방식이 표준화되고 있는 추세이므로, 전자책을 위한 음성합성 엔진 역시 위와 같은 표준 마크업언어 등을 파싱하고 지정된 특징의 합성음을 생성할 수 있는 기능이 요구된다.

```
<speak version= “1.0” xml:lang= “en-GB”>
        Hello, how are you?
        <prosody rate=“x-fast”>
            This sentence is spoken fast
        </prosody>
        <prosody pitch= “x-low”>
            This sentence is spoken low pitch
        </prosody>
        <prosody pitch= “medium”>
            This sentence is spoken medium pitch
        </prosody>
        <prosody pitch= “x-high”>
            This sentence is spoken high pitch
        </prosody>
        <prosody rate= “fast”>
            This sentence is spoken fast
        </prosody>
        <emphasis level= “strong”> This sentence is spoken with
                                   stress</emphasis>
</speak>
```

<자료>: http://www.voicexmltutorial.com/

(그림 2) SSML을 합성할 텍스트 표현 예

4. 대화체 음성합성 기술 요구

지능형로봇, 텔레매틱스, 홈 네트워크, 차세대 PC와 같은 신성장동력산업 분야에서 요구하는 음성합성 기술은 종전의 뉴스나 일기예보를 사용자에게 읽어주는 수준을 뛰어 넘어, 기기에게 음성으로 명령을 전달하고 가공된 정보는 음성으로 출력해 주는 대화형 음성 인터페이스를 매우 선호한다.

음성합성 기술측면에서는 이러한 용도에 맞는 합성음을 생성하기 위해서는 뉴스읽기와 같은 단조로운 낭독체 음성합성이 아닌 전달하고자 하는 메시지의 내용에 따라, 사용자의 의도에 따라 합성음의 분위기가 다르게 표현되는 대화체 음성합성 기술을 요구하고 있다. <표 2>는 IBM TTS에서 정의한 대화체의 유형을 나타낸다.

특히 이러한 요구사항은 지능형로봇 분야에서 두드러지게 나타난다. 일부 지능형 로봇을 개발하고 있는 기업에서는 로봇의 감정상태를 표현할 수 있도록 감정상태에 따라 로봇의 눈 모양이나 고개 모양 등을 제어할 수 있는 기술을 탑재하고 있다.

또한 대화체 음성합성 기술은 전자책 분야에서도 매우 중요하게 요구된다. 현재는 음성합성 기술을 이용한 전자책 읽기 기능을 구현하는 대부분이, 하나의 차분한 목소리로 전체 내용을 다 읽어준다. 그러나, 이러한 서비스는 사용자에게 매우 지루함을 느끼게 하여 장시간 이용이 어렵다. 이를 해결하기 위해서는 다음색과 합성하고자 하는 텍스트의 상황 문맥에 맞는 발화스타일을 합성할 수 있는 기술이 요구되고 있다.

〈표 2〉 IBM TTS에서 정의한 대화체 유형

Expression	Example
Good news	I have successfully reset your PIN.
Bad news	I am unable to verify your identity.
Confusion	I did not understand your request.
Contrast	This is a *round-trip* fare.
Apology	I cannot find you in my records.
Question	Do you confirm the sale of all shares?
Confidence	Your account balance is $8,432.50.
Greeting	Welcome to the IBM help desk.
Farewell	Thanks for calling. Goodbye.

<자료>: http://www.ssw5.org/

5. 음색변환 기술

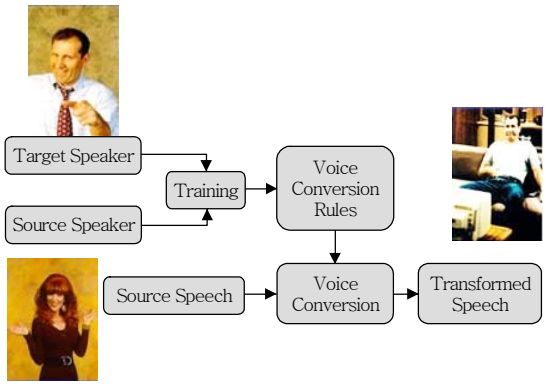
음색변환 기술이란 (그림 3)과 같이 단말기에 내장되어 있는 합성시스템의 음색을 특정 변환과정을 거쳐 사용자가 원하는 친구나 부모님, 선생님의 음색을 모방하여 합성음을 생성할 수 있도록 하는 기술이다.

휴대폰 단말기 제조 기업에서 요구하는 음색변환 기술의 서비스 시나리오는, 만일 여자친구의 메시지가 오면 여자친구의 목소리로 해당 단문 메시지를 읽어 주었으면 좋겠고, 자녀의 메시지가 도착하면 해당 자녀의 목소리로 메시지를 읽어주도록 하겠다는 것이다. 선생님의 메시지는 선생님 목소리로, 부모님의 메시지는 부모님의 목소리로, 친구의 메시지는 친구의 목소리로 메시지가 낭독된다면 기본적으로 메시지를 보낸 사람의 신원을 보다 쉽게 파악할 수 있으며, 또한 메시지의 내용을 보다 정감 있게 들을 수 있다는 점이다. 이는 최근 많은 관심을 가지고 있는 따뜻한 디지털 세상이란 키워드와 일맥상통하는 서비스 시나리오이다.

현재의 음색변환 기술은, 음색변환을 위해서는 변환될 사용자의 음성 데이터를 필요로 한다. 현재 널리 쓰이는 모델변환 기법을 이용하는 경우, 음성 데이터가 많을수록 모델변환에 용이하다. 그러나 사용자가 직접 단말기에서 많은 문장을 녹음하는 것은 사용자에게 불편을 초래할 수 있다.

또한 단말기에 내장된 합성 DB는 조용한 방음실에서 녹음된 성우의 음성으로 제작되나, 사용자는 단말기에서 직접 입력하거나 사용자의 개인 PC 환경에서 녹음을 하게 된다. 이때 녹음환경의 불일치 문제가 발생할 수 있다. 사용자들은 매우 다양한 수준의 녹음장치를 사용할 수 있으며, 또한 녹음환경 역시 매우 다양한 잡음 환경에서 수행되므로 환경의 불일치에 의한 변환상의 문제점이 발생할 수 있다.

현재까지의 음색변환 기술의 수준은 위에서 언급한 실제 환경의 문제점을 고민할 만큼 발전되어 있지 못하고 있다. 가장 큰 이유는 그만큼 음색변환이 어렵기도 하고, 또한 실제 이 기술이 적용된 사례가 아직 구축되어 있지 않기 때문이다.



<자료>: http://www.busim.ee.boun.edu.tr/

(그림 3) 음색변환 기술의 개념도

Ⅲ. 내장형 음성합성 방법론

본 장에서는 음성합성 기술의 변천사를 소개하고자 한다. 비록 각각의 기술들이 시간의 흐름에 따라 발전하여 왔지만 어느 방법 하나도 현재 완전히 시장에서 사라진 방법론은 없다. 그 이유는 각각의 방법론이 고유의 장점을 가지고 있기 때문이다. 본 장에서는 이러한 음성합성 기술들의 변천사와 각각의 방법론들이 내장형 음성합성 분야에서 어떻게 사용되고 있는지 고찰하고자 한다.

1. 음성 녹음 및 재생 칩

합성음을 생성하는 가장 간단한 방법은 사용자에게 들려줄 안내 멘트를 미리 녹음하여 가지고 있다가 필요한 시점에서 이를 재생하여 들려주는 기술로 (그림 4)와 같이 가전기기나 엘리베이터 등에 내장

<자료>: http://blog.paran.com/voiceland/

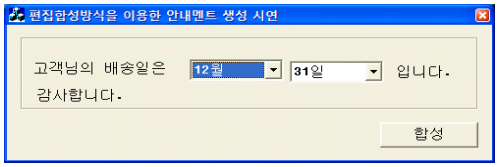
(그림 4) 녹음 및 재생방식의 IC 칩 예

시키기 위해 칩화되어 제작되는 경우가 많다.

이 방법은 음성안내뿐만 아니라 다양한 효과음 및 음악 등을 장치에 내장시켜 놓고 활용하고 있다. 간단한 예로 우리가 핸드폰에서 버튼을 누르면 버튼마다 재미있는 소리효과가 나오는 것이 대표적인 단순 녹음 및 재생 방식이다. 또 다른 예로는 소리가 나는 장난감 인형이다. 즉, 재미있는 몇 마디를 녹음하여 칩에 저장해놓고 특정한 상황이 되거나 버튼을 누르면 녹음된 소리를 들려주는 방식이다. 이 방법은 알고리듬적으로는 매우 간단하지만, 실제 구현 시에는 여러 가지를 고려해야 한다. 소리 나는 장난감 인형처럼 한번 만들어져 칩화되면 그 내용을 수정할 수 없다. 또한 녹음되는 문장 수와 압축방법도 잘 고려해야 한다. 장치에서 사용할 문장 수가 많은 경우 압축하지 않고 저장해 놓으면 많은 저장공간을 필요로 하므로 저장공간 비용이 높아진다. 반면에 음성 데이터를 고압축 방법을 사용하여 압축하여 집어 넣으면 저장공간은 줄일 수 있으나, 녹음된 내용을 재생해야 하는 경우 압축의 복원을 수행해야 한다. 일반적으로 고압축 방법은 압축을 풀 때에도 복잡한 연산을 수행하게 되므로 해당 칩이나 장치에 복잡한 코덱이 들어가야 하고 연산량도 많아지므로 그 만큼 전력 사용량도 증가하게 된다.

2. 편집합성

단순 녹음 및 재생 방식은 매번 정해진 내용만 반복적으로 들려준다. 그러나, 또 다른 응용 영역에서는 문장의 기본 틀은 고정되어 있어나 그 문장의 내용어는 변해야 하는 응용이 있다. 대표적인 예로써 전화번호를 알려주는 114서비스를 들 수 있다. 이 서비스에서는 기본적인 문장의 틀은 “문의하신 번호는 000국에 0000번입니다.”와 같이 고정되어 있으나, 그 내용어인 전화번호 부분 “000국에 0000번”은 항상 실시간으로 변경되어야 한다. 이와 같이 고정된 문틀에 변경되는 부분만 변경해서 합성음을 생성하는 방법을 편집합성이라 한다. (그림 5)는 본 절에서 기술하는 편집합성의 한 예이다. 즉 “고객님의 배송일은 00월 00일입니다. 감사합니다.”라는



(그림 5) 편집합성 예

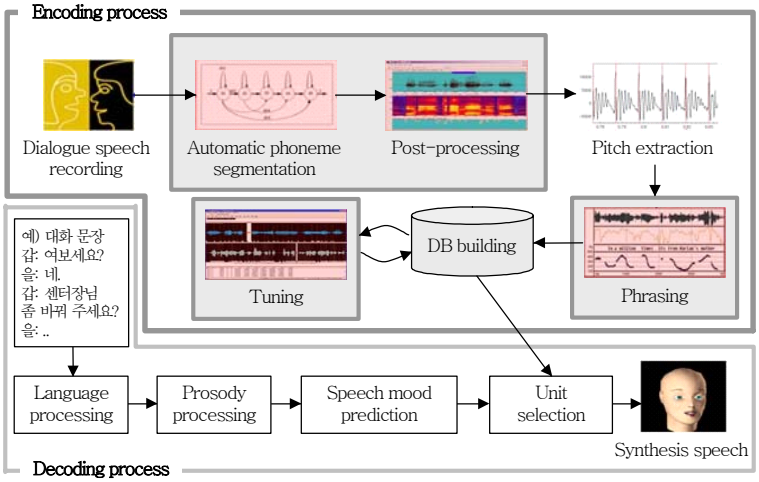
고정된 문장패턴에 날짜 부분만 교체하여 합성음을 만들 수 있다.

편집합성 방식은 위와 같은 제한된 영역에서는 매우 적은 연산량으로 고품질의 합성음을 생성할 수 있어 예금 조회, 거래내역 조회, 증권 조회, 간단한 길안내, 열차시간안내 등과 같은 고정된 문장 틀을 가진 안내분야에 활용되고 있다. 특히, 현재 길안내 서비스를 주로 하는 내비게이션용 단말기들은 주로 이와 같은 편집합성 방법을 활용한다.

그러나 이 방법은 기본적으로 무제한 음성합성이 어렵다는 한계로 인하여 주로 단문 형식의 안내멘트 합성용으로 사용되며, 지능형로봇과 같은 대화형 인터페이스를 요구하는 서비스나 실시간 교통방송 및 재난방송 안내, 이메일 읽기 등과 같은 무제한 음성합성이 요구되는 임베디드 분야의 서비스에서는 사용하기 어렵다.

3. 음편조합 방식의 음성합성

음편조합 방식을 이용한 음성합성 기술은 일반적으로 단어보다 훨씬 작은 소리 단위를 조합하여 합성음을 생성하며 개략적인 개발과정은 (그림 6)과 같다. 예를 들어, 한글의 경우 자음 19개, 모음 21개의 변이음으로 대표소리를 이루는데, 이를 19개 자음에 대한 음성과 21개 모음에 대한 소리만 가지고 있으면 음질은 매우 떨어지겠지만, 이론적으로 우리말의 모든 글자의 소리를 총 40개 변이음의 조합으로 생성해 낼 수 있다. 또한 음편조합을 이용한 음성합성 방식은 편집합성처럼 녹음된 음편을 그대로 결합하여 합성음을 생성하므로 음성신호 생성을 위한 음성 신호 조작을 거의 이용하지 않아 신호처리에 의한 왜곡이 없고 빠른 속도의 음성합성을 수행할 수 있다.



(그림 6) 음편조합 방식을 이용한 음성합성 개념도

음편접합 방식은 위에서 기술한 특징에 의해 대용량 서버형 음성합성 응용분야에서 매우 성공적으로 활용되고 있으나 내장형 소용량 음성합성 기술로 사용하기에는 다음과 같은 문제점을 가지고 있다.

첫째는, 음소와 같은 작은 단위의 음편을 사용하므로 적은 수의 음편만 있으면 무제한 합성을 수행할 수 있을 것 같지만 실제로는 전혀 그렇지 않다. 그 이유는 조음효과 때문이다. 조음효과를 예로 들어 설명하면, /ㅏ/ 라는 소리는 앞뒤에 오는 소리가 무엇이냐에 따라 상당히 다른 신호특성을 가지게 된다. 즉, /ㄲ/+/ㅏ/+/ㄱ/ 음소환경에서 사용된 /ㅏ/ 소리와 /ㄳ/+/ㅏ/+/ㅎ/ 음소환경에서 사용된 /ㅏ/ 소리의 음향적 특징은 매우 다르다. 그러므로 "우아한"이란 단어의 가운데 음절인 /아/ 소리를 생성함에 있어서, /ㄳ/+/ㅏ/+/ㅎ/ 환경의 /아/가 아닌 /ㄲ/+/ㅏ/+/ㄱ/ 음소환경의 /아/가 사용된다면 합성된 소리 "우아한"은 전혀 우아하지 않게 들린다. 이러한 현상을 문맥불일치로 인한 음질열화라고 한다. 그러므로, 음편조합 음성합성 방식에서는 임의의 문장을 합성할 수 있는 매우 다양한 문맥의 음편들을 미리 녹음하여 가지고 있어야 한다. 즉, /ㅏ/ 소리의 앞뒤 음소를 고려하면 /?/+ ㅏ/+/?/와 같은 음소문맥을 형성하게 되고, /?/ 부분에는 다른 모든 음소가 올 수 있다고 가정하자. 그러면 우리말은 40개 음소가 존재하므로 /?/+/ㅏ/+/?/는 서로 다른 문맥환경을 가지는 총 1600개의 음편이 있어야 한다. 또한 이를 확장하여 가운데 음편이 /ㅏ/ 대신에 다시 40개의 우리말 음소가 올 수 있으므로(즉, /?/+/?/+/?/ 형태의 음소열, 이를 3음소열 또는 트라이폰(tri-phone)이라고 함), 우리말에 대한 음편은 총 64000개가 필요하게 된다. 그러나 이것은 앞뒤 결합음소 문맥만 고려한 것이고, 이 외에 음성의 음향적 특성에 영향을 미치는 다른 많은 요인들이 있는데 이들을 고려하면 실로 엄청난 수의 음편을 가지고 있어야 무제한 음성을 합성할 수 있다. 또한 이런 많은 수의 음편을 확보하기 위해서는 다양한 영역의 텍스트 데이터를 기반으로 한 발성목록을 설계하고 장시간의 합성 DB 녹음을 수행해야 한다. 이는 결국, 대량의 데이터를 제작하기 위한 긴 준비과정, 긴 녹음과정, 녹음과정에서의 일관적인 녹음품질의 유지의 어려움, 녹음된 데이터의 장시간의 가공과정이 필요하며 개발비용이 매우 높다.

둘째는, 음편조합 방식의 장점에 기인한 것으로 음성에 대한 상세한 디지털 신호처리를 수행하지 않으므로 원음의 음질을 유지할 수 있다는 장점이 있지만, 역으로 음성신호에 대한 조작이 어려워 합성음의 운율조절이나 음색조절이 어렵다는 단점을 가진다. 이로 인하여 첫번째 단점에서 기술한 바와 같이 막대한 개발비용을 들여 개발한 합성 DB에 대한 활용률이 떨어지고, 응용 제품이 다음색을 요구하는

경우에는 요구되는 음색 수에 비례한 개발비용의 증가를 초래한다.

셋째는, 저장공간의 문제이다. 즉 첫번째 단점에서 기술한 바와 같이 본 방법은 보유하고 있는 음편의 수가 많을수록 고품질의 합성음을 생성할 수 있으며, 통상적으로 수백 MB~수 GB급의 저장공간을 사용하게 된다. 그러므로 상대적으로 저장공간의 크기가 제한된 분야에서는 고품질을 생성하기 어렵다.

그러나, 최근 연구에서는 위에서 기술한 단점들을 해결하여 소용량 보이스폰트를 생성할 수 있는 다양한 방법들이 연구되고 있으며, 이 중에서 합성음의 음질저하 없이 소용량의 보이스폰트를 생성할 수 있는 방법으로 널리 사용되고 있는 방법은 pre-selection을 이용한 음편의 수를 감소시키는 방법이다[2]. 원리는 다음과 같다. 즉, 오프라인에서 대량의 임의의 텍스트를 합성하여 각각의 음편의 사용빈도를 조사하여 사용빈도가 낮은 음편을 보이스폰트에서 제거하는 기술이다. 그렇게 함으로써 최종적으로 보이스폰트의 수를 매우 많이 줄일 수 있고, 음편수가 줄어듦에 따라 실행시간 연산량도 동시에 줄이는 방법이다.

그러나, 이 방법 역시 줄일 수 있는 음편의 수는 어느 정도 한계가 있으며 통상적으로 기존 수백 MB급 보이스폰트를 수십 MB급으로 줄이는 데 주로 사용된다. 그 이상의 음편 수를 제거하게 되면 음편의 접점에서 불연속이 발생하거나 또는 합성음이 단조로워지는 경향이 있다. 그러므로 이 방법은 수십 MB급 메모리를 제공할 수 있는 고사양 단말기(예를 들어 PDA 수준)에서 주로 많이 사용되는 방법이며, 수 MB급 메모리를 제공하는 단말기에서는 사용이 어렵다.

4. 통계적 모델링을 이용한 파라미터릭 방식의 음성합성

통계적 모델링을 이용한 파라미터릭 음성합성 방식은 음편조합을 이용한 음성합성 방식과 마찬가지로 무제한 음성합성을 하는 방법으로, 주로 음편조합을 이용한 음성합성 방식과 같이 대용량 저장공간을 활용할 수 없는 소용량의 합성 시스템을 요구하는 분야에서 주로 활용되는 방법이다. 본 방법은 그 이름에서 알 수 있듯이 먼저 음성신호의 음편을 그대로 사용하지 않고 특징 파라미터 형태로 변환하여 사용하는 것과, 그 특징 파라미터들을 통계적으로 모델링하여 어떤 소리의 대표값을 생성하고, 이를 보코더를 통과시켜 합성을 생성하는 방식이다.

음성신호를 파라미터 형태로 변환하는 것은 주로 음성코딩이나 음성인식 등에서 많이 사용되는 방법이다. 많이 사용되는 파라미터 형태는 성도전달함수 모형을 이용해 선형예측계수를 이용하는 방법이다. 선형예측계수는 우리의 발성기관인 성도와 성대의 관계를 모델링한 것으로, 성도에서 어떤 소리의 원시 신호(떨림)를 발생시키고, 이 원시신호는 사람마다 가지는 고유의 성대를 지나면서 그 사람의 고유의 음색에 맞게 신호가 가공되어 음성신호로 생성된다는 원리를 이용한다. 성도의 원시신호의 떨림의 값을 하나의 숫자 값으로 표현하고, 성대를 넓이와 길이가 다른 몇 개의 관이 연결된 형태로 이해하고, 이렇게 연결된 여러 개의 관이 가지는 신호전달 특성을 마치 디지털 신호 처리의 필터로 표현하여 해당 음성신호의 주파수 특성 모양을 가장 잘 표현하는 필터의 계수를 구하여 몇 개의 수치 값으로 표현하게 된다.

통계적 모델링을 이용한 파라미터릭 음성합성에서는 위와 같은 방법으로 구한 어떤 특정 소리의 파라미터 계수들을 모아서 통계적으로 처리하여 몇 개의 대표적인 파라미터 계수의 집합으로 표현하게 된다. 예를 들어 /ㅏ/ 음소에 해당하는 0.5초 길이의 음성신호가 있다고 하면, 이 신호를 0.025초 단위로 파라미터를 추출하면 총 20개의 계수열(즉, 1개의 떨림 계수+ 19개의 필터 계수, 총 20개의 수치로 표현되는 성도모델계수가 다시 20개가 연속하여 있는 형태)이 생성되고, 이를 통계적 모델링을 이용하여 다시 20개의 연속을 대표할 수 있는 3~5개의 대표 계수열을 구하게 된다. 이렇게 되면 0.5초 길이를 가지는 /ㅏ/음은 파형으로 표현하면 16bit, 16kHz를 기준으로 총 8000개의 파형 값으로 표현되나, 위와

같이 통계적 파라미터 모델로 표현하면 20차×3개의 계수, 총 60개의 수치 값으로 표현할 수 있게 된다. 이를 수치 값의 수로 표현하면 8000개의 숫자를 60개의 숫자로 대표하게 되는 것이므로, 약 133.3배의 압축효과가 있게 된다.

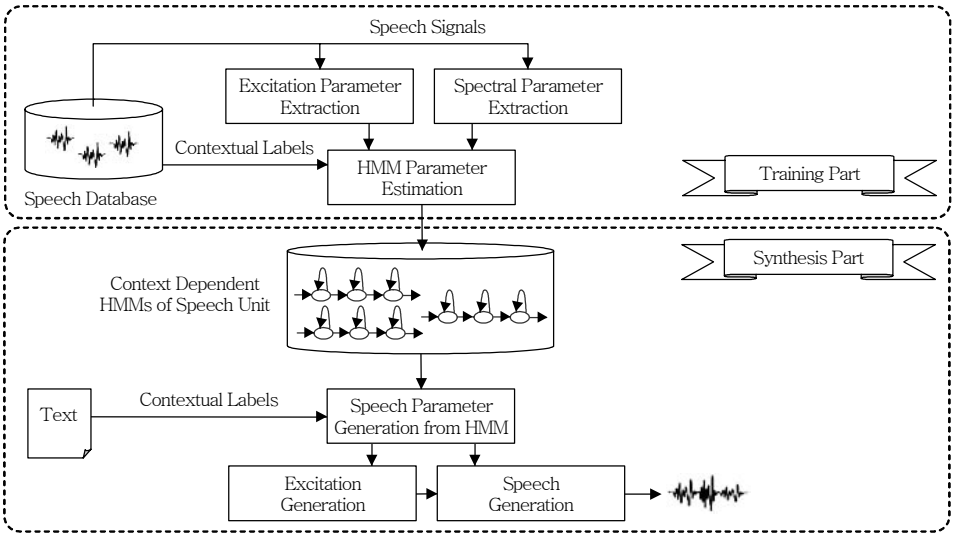
음편조합을 이용한 음성합성에서는 각각의 음편에 대한 파형을 직접 저장하여 가지고 있지만, 통계적 모델링을 이용한 파라미터릭 방식에서는 위에서 기술한 최종 대표치만을 가지고 있게 되므로 훨씬 적은 저장공간에 저장시킬 수 있다. 즉, 휴대폰이나 전자사전 등과 같이 매우 작은 메모리만 가진 장치에도 저장시킬 수 있게 된다.

그러나, 8000개의 숫자를 60개의 숫자를 이용하여 대표치를 구하는 과정에서 일정부분 정보의 손실이 발생하며, 합성 시에는 다시 60개의 숫자를 8000개의 파형에 대한 숫자 값으로 복원하게 되는데, 이렇게 복원된 8000개의 음성신호에 대한 파형의 숫자 값은 원래의 8000개의 숫자 값과 동일하지 않게 된다. 그래서 합성된 신호는 원본의 신호보다 훨씬 품질이 떨어지는 문제점을 가지게 된다. 비록, 합성된 8000개의 파형 값이 원본과는 다르더라도 이 합성된 소리를 사람이 들었을 때 합성된 소리가 어떤 내용인지를 파악하는 것은 전혀 문제가 없다. 즉, 비록 음질은 음편조합 방식에 비해 떨어지기는 하지만 그 내용을 사람이 듣고 이해하는 데에는 전혀 문제가 없는 수준은 되고, 상대적으로 적은 저장공간을 사용하게 되므로 음편조합 방식을 이용한 음성합성 기술을 이용할 수 없는 응용영역에 사용된다.

Ⅳ. 응용사례

본 연구팀에서는 “신성장동력산업용 대용량 대화형 내장처리 음성인터페이스 기술” 개발 연구의 일환으로 임베디드 단말기에서 사용할 수 있는 내장형 소용량 음성합성 기능을 가지며, 대화체에 특화되어 있고, 음색변환에 용이한 HMM 기반 내장형 소용량 대화체 음성합성 기술을 개발하고 있다. 본 장에서는 현재까지 개발된 HMM 기반 음성합성 기술에 대하여 기술하고자 한다[3]-[6].

HMM 기반 음성합성 기술[3],[7],[8]은 (그림 7)과 같이 음편접합 방식과 달리 음성신호의 스펙트럼 정보, 피치 정보, 지속시간 정보를 각각 독립된 Gaussian 확률분포를 가지는 HMM 모델로 훈련하여 합성용 보이스폰트를 생성한다. 합성 시에는 훈련된 HMM 모델 파라미터로부터 합성음 생성을 위한 음성 특징 파라미터를 생성하고, 이를 적당한 방법



(그림 7) HMM 기반 음성합성 개요

으로 보간하여 합성음 궤적을 생성한다.

HMM 기반 음성합성 방식은 파형접합 방식을 이용한 소용량 합성시스템에서 나타나는 접점에서의 불연속 등이 없다는 장점이 있으나, 과도하게 스무딩된 음성 특징 파라미터 궤적과 보코딩 방식의 합성음 생성 과정에서 나타나는 둔탁한 음질 등 전반적으로 파형접합 방식에 비해 명료성이 떨어진다는 것이 단점으로 지적되고 있다. 그러나 최근 이러한 문제점을 해결하기 위해 정교한 음향모델링, global variance를 이용한 다이내믹한 음성파라미터 궤적의 생성, mixed excitation을 이용한 보코더의 명료성 개선 등을 통해 상당부분 음질개선이 이루어지고 있다.

음향모델 훈련단계에서는 전처리 단계에서 구축된 문맥정보를 이용하여 문맥종속 HMM 모델을 훈련시키는 과정으로, 음성인식시스템의 훈련과정과 유사하게 진행하였다. 훈련에 사용된 HMM의 구조는 left-to-right 5 state HMM을 사용하였다. 음향모델 훈련을 위한 특징 파라미터는 프레임 크기 20ms, 프레임 중첩 5ms, blackman windowing을 통한 프레임단위로 25차 Mel-cepstrum을 추출하고, 이들의 delta, delta-delta 총 75차를 구하였으며, 여기신호모델 훈련을 위해 Praat를 이용해 F0를 구하여 log를 취한 후 이들의 delta, delta-delta를 구하여 총 78차의 다중 스트림 특징 파라미터 벡터를 구성하였다. 스펙트럼 정보에 대해서는 대각 공분산 행렬을 가지는 Gaussian 분포로 모델링하였으며, F0 정보는 유성음에 대해서는 대각 공분산 행렬을 가지는 Gaussian 분포로 모델링하고, 무성음 구간은 이산분포로 모델링하는 다중공간 확률분포를 가지는 Gaussian 모델로 모델링하였다.

기본적인 프로토타입 버전은 PC 환경에서 초벌을 개발하고, 개발된 HMM 기반 합성엔진을 저사양 프로세스와 스피커를 가진 임베디드 환경에서 문제점 및 성능개선 부분 등을 고찰하기 위하여 60MHz clocks/second 처리속도의 ARM720T 프로세서, 16MB NAND 메모리, 32MB SDRAM으로 구성된 ITS 단말용 OBE 보드에 정수형 버전 및 일부 모듈은 연산의 고속화를 위해 ARM 어셈블리어로 구현하였다. 현재 ITS 단말기에 탑재한 HMM 기반 음성합성 시스템의 메모리 사용량은 <표 3>과 같다.

또한 본 방법은 합성음 샘플단위로 MLSA 역필터를 통한 음성신호 생성 부분이 합성시간의 대부분을 차지한다. ITS 단말기에서 실시간 음성합성을 수행하기 위해 <표 4>와 같이 실수연산 MLSA 역필터링 모듈을 정수형 연산 버전을 어셈블리어로 각각 구현하여 실시간 음성합성을 할 수 있도록 하였다.

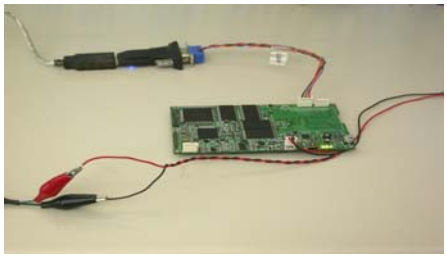
(그림 8)은 현재까지 개발된 HMM 기반 내장형 소용량 대화체 음성합성 엔진이 탑재된 ITS 테스트 보드이다.

〈표 3〉 내장형 음성합성 시스템의 메모리 사용량

	NAND	SDRAM
언어처리사전	1.7	2.1
보이스폰트	0.93	1.3
합성엔진	0.47	2.2
총합	3.19	5.6

〈표 4〉 내장형 음성합성 시스템의 연산속도

구현방법	실행시간(초)	실시간율
실수형 버전	12.0	6.85
정수형 버전	2.40	1.39
어셈블리어 버전	0.92	0.52



(그림 8) 개발된 소용량 음성합성엔진을 내장한 ITS 테스트 보드

V. 결론

본 고에서는 최근 점차 커져가고 있는 임베디드 응용분야에서 요구하는 음성합성 기술에 대한 요구

사항과 소용량 음성합성 기술을 개발하는 방법을 살펴보았다. 또한 본 연구팀에서 "신성장동력산업용 대용량 대화형 내장처리 음성인터페이스 기술" 개발 연구의 일환으로 개발중인 HMM 기반 내장형 소용량 대화체 음성합성 기술에 대해서도 개략적으로 기술하였다.

서론에서도 기술한 바와 같이, 일반사용자들은 하드웨어 규격이나 기술적 방법론에 상관없이 항상 고품질의 합성음을 요구한다. 이에 본 연구팀에서는 현재 개발된 내장형 소용량 음성합성 시스템의 음질을 보다 향상시키고 보다 자연스러운 대화체 합성음을 생성할 수 있도록 지속적으로 연구 개발을 수행할 예정이다.

● 용 어 해 설 ●

SSML(Speech Synthesis Markup Language): 2004년부터 W3C 권고안으로 제정되었으며 XML 형식을 웹 및 기타 응용에서 음성합성 시스템을 보조하기 위해 제정된 마크업언어 규격

VOCODER Voice Decoder: 음성신호의 내재된 특성을 이용, 모델링하여 이 모델의 각종 매개변수의 집합으로 변환하고, 복원에는 이 매개변수를 이용하여 다시 음성신호를 복원하는 기술

약 어 정 리

DAISY	Digital Accessible Information System
MLSA	Mel Log Spectrum Approximation
MSD	Multi-Space Distribution
TTS	Text To Speech

참 고 문 헌

[1] http://www.nuance.com/realspeak/vocalizer/

[2] Mark Beutnagel, Mehryar Mohri, and Michael Riley, "Rapid Unit Selection from a Large Speech Corpus for Concatenative Speech Synthesis," *In EUROSPEECH'99*, 1999, pp.607-610.

[3] S.J. Kim, J.J. Kim, and M.S. Hahn, "Implementation and Evaluation of an HMM-based Korean Speech Synthesis System," *IEICE Trans. Inf. & Syst.*, Vol. E89-D, No.3, 2006, pp.1116-1119.

[4] 김종진, 오승신, 최문옥, 김상훈, 박준, 이영직, "ETRI 대화체 음성합성시스템 소개," 대한음성학회 봄 학술대회, 2003.

[5] 김종진, 김정세, 김상훈, 박준, 이윤근, "ETRI 소용량 대화체 음성합성 시스템," 대한음성학회/한국음성과학회 공동학술대회, 2007.

[6] 김종진, 김상훈, 김정세, 박준, 이윤근, "혼합 여기 모델을 이용한 HMM 기반 소용량 대화체 음성합성 시스템," 대한음성학회 추계 학술대회, 2007.

[7] K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, "Multi-space Probability Distribution HMM," *IEICE Trans. Inf. & Syst.*, Vol.E85-D, No.3, Mar. 2002, pp.455-464.

[8] http://hts.sp.nitech.ac.jp/