# Voice Browser를 위한 음성 인식 웹서비스 환경에 관한 연구

홍인숙, 김윤중
한밭대학교 컴퓨터 공학과
e-mail : ishong@hanbat.ac.kr

# A Study of Speech Recognition Web Services Environment for Voice Browser

In-Suk Hong, Yoon-Joong Kim
Dept of Computer Engineering, Hanbat University

요 약

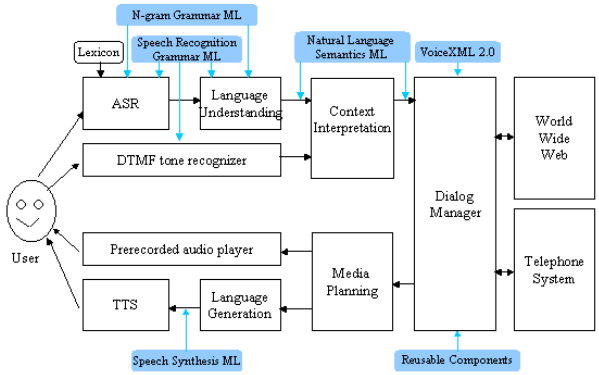
음성인터페이스 관련 표준화는 음성 대화, 음성인식/합성, 전화망 등의 접속망을 상호 분리하여 음성정보시스템 구성요소들 각각의 상호 독립적인 개발을 보장해 주며, 각 요소의 이해가 없이도 음성정보시스템을 개발할 수 있도록 함으로써 음성정보기술의 보급 및 확산에 크게 기여하고 있다. 이에 W3C에서는 Voice Browser에 대한 표준화를 현재 진행 중에 있으며 Vocie Browser WG에서 Voice Browser를 위한 SIF(Speech Interface Framework)를 제안하였다. 제안된 SIF에서 Voice Browser가 음성인식을 실행하기 위해서는 많은 자원의 소요와 부하가 생길 수 있다. 이러한 문제점을 해결하기 위해 본 논문에서는 음성인식 웹 서비스를 기존의 SIF에 추가한 새로운 형태의 SIF를 제안하고자 한다. 음성인식은 원격 시스템에서 수행하고 그 결과를 Voice Browser가 사용할 수 있도록 음성인식 웹서비스 환경을 구축하였다. 그리고, XML-SRGS 포멧의 grammar를 음성인식기가 사용하는 EBNF 포멧의 grammar로 변환시키는 변환기를 구현하였다.

## 1. 서론

음성 합성 및 인식 기술의 발전으로 음성 기술을 이용한 웹 브라우저의 개발 연구가 가속화되고 있으며 W3C에서는 그 중요성을 인식하여 Voice Browser에 대한 표준화를 진행하고 있다[1][2]. Voice Browser Working Group에서는 Voice Browser를 위한 SIF (Speech Interface Framework)를 정의하였다. 음성인터페이스 관련 표준화는 음성 정보시스템 구성요소들 각각의 상호 독립적인 개발을 보장해 주며, 각 요소의 이해가 없이도 음성 대화만을 설계 기술하여 음성정보시스템을 개발할 수 있도록 하였다. 음성 인터페이스 표준을 지원하는 음성/합성 시스템이라면, 개인이 개발한 또는 어떤 기관이 개발한 것이라도 별다른 수정 없이 시스템에 그대로 적용할 수 있다[3][4].

(그림 1)은 W3C에서 제안한 SIF의 구조를 보여주고 있다. ASR(Automatic Speech Recognizer)은 사용자로부터 음성을 받아들여 텍스트를 생성한다. 입력 받은 음성으로부터 단어를 인식하기 위해 음성인식문법을 사용한다.[5] SRGS(Speech Recognition Grammar Specification)[6]는 음성 인식 문법 명세서이며 음성과 DTMF 입력 문법을 모두 기술하며 음성인식문법을 코딩하기 위한 XML 포맷(XML-SRGS)과 텍스트 포맷(ABNF-SRGS)을 포함하고 있다.

W3C SIF는 Dialog Script를 포함한 웹 페이지를 입력받아 Dialog Script에 포함된 명령에 따라 application을 실행한다. 음성인식을 Voice Browser에서 수행하므로 많은 자원이 소요되며, 또한 Voice Browser에 부하를 줄 수 있다. 이러한 문제를 해결하기 위해서는 음성인식을 원격 시스템에서 수행하고 Voice Browser는 이를 호출하여 인식 결과만을 이용하는 웹서비스 환경으로 구축해야 한다.




(그림 1) W3C Speech Interface Framework

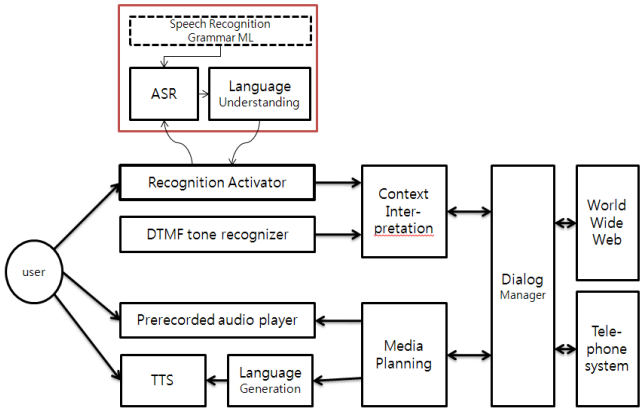
이에 본 논문에서는 음성인식 웹 서비스를 W3C SIF에 추가한 새로운 형태의 SIF를 제안하고자 한다. 그리고, XML-SRGS 포멧의 grammar를 음성인식기가 사용하는 EBNF 포멧의 grammar로 변환시키는 변환기를 구현하고자 한다.

## 2. 제안한 Speech Interface Framework

SIF는 사용자의 음성과 다른 인터페이스를 이용하여 응

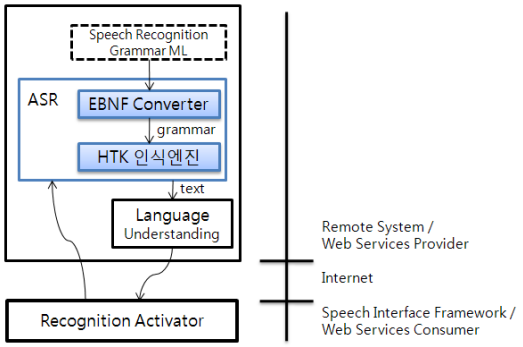
용프로그램과 상호작용할 수 있는 마크업 명세이다. 그 중 Voice Browser의 중요한 인터페이스인 음성인터페이스를 위한 음성인식은 Voice Browser에서가 아닌 원격 시스템에서 수행하고 그 결과를 Voice Browser가 사용할 수 있는 음성인식 웹서비스 환경을 구축하고자 한다. [7][8]

(그림 2)는 본 논문에서 제안한 SIF 구조이다.




(그림 2) 제안한 Speech Interface Framework

음성 인식 웹 서비스는 (그림 3)과 같이 웹서비스 소비자(Web Services Consumer)와 웹서비스 제공자(Web Services Provider) 로 구성되어 있다.




(그림 3) 음성인식 웹 서비스

웹 서비스 소비자는 Recognition Activator로, 사용자로부터 음성을 입력받아 웹서비스 제공자를 호출하고 수신받은 결과를 Dialog Manager에게 전송한다. 웹 서비스 제공자는 웹 서비스 소비자로부터 사용자의 음성을 입력 받아 인식한 후 의미를 해석하고 인식된 결과를 텍스트로 출력한다. 이때, Dialog Manager로부터 받은 XML-SRGS로 기술된 grammar를 가지고 인식하는데 인식기가 사용하는 형태의 문법으로 변환하기 위해 본 논문에서 구현한 EBNF Converter를 사용한다. 웹 서비스 제공자는 웹서비스 소비자로부터 입력받은 음성을 인식하는 Speech_Recognition( )함수와 XML-SRGS(XML-SRGS로 기술된 grammar)을 음성인식기가 사용하는 EBNF(EBNF로 기술된 grammar)로 변환하는 Grammar_Component( ) 함수를 제공한다.

## 3. EBNF Converter

EBNF Converter는 음성인식문법 표준인 XML-SRGS를 EBNF로 변환해주는 변환 컴포넌트이다.

다음은 "예/아니오"를 인식하기 위한 Dialog Script의 일부 샘플예제이다. (그림 4)은 EBNF Converter의 입력인 XML-SRGS이며, (그림 5)는 출력인 EBNF이다.

본 논문에서는 번역기(그림 6)를 이용한 방법과 XSLT Style Sheet (그림 7)를 이용한 두 가지 방법으로 EBNF Conveter를 구현하였다.

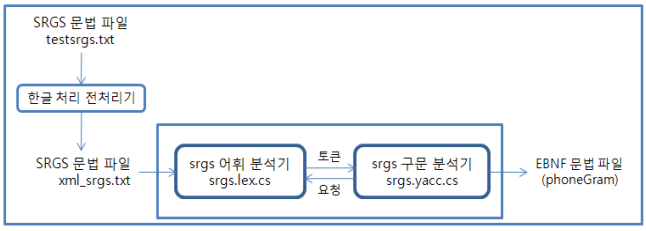
```
<?xml version="1.0" encoding="UTF-8"?>

<!DOCTYPE grammar PUBLIC "-//W3C//DTD GRAMMAR 1.0//EN"
          "http://www.w3.org/TR/speech-grammar/grammar.dtd">
<grammar xml:lang="ko" version="1.0" mode="voice"
         xmlns="http://www.w3.org/2001/06/grammar"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://www.w3.org/2001/06/grammar
         http://www.w3.org/TR/speech-grammar/grammar.xsd"
         root="yes_no_ko">
  <!-- yes/no grammar -->
  <rule id="yes_no_ko" scope="public">
    <example>예</example>
    <one-of>
      <item>예</item>
      <item>아니오</item>
    </one-of>
  </rule>
</grammar>
```

(그림 4) Grammar Component 입력 : XML-SRGS

```
$yes_no_ko = 예 | 아니오 ;
( sil $yes_no_ko sil )
```

(그림 5) Grammar Component 출력 : ABNF

### (1) 번역기를 이용한 EBNF Converter




(그림 6) 번역기를 이용한 EBNF Converter

번역기는 본 연구실에서 개발한 컴파일러 생성기(IISPLGemerator)[9]를 사용하여 생성하였다.

번역기를 이용한 EBNF Converter는 한글처리를 위한 전처리기와 어휘 분석기, 구문 분석기로 구성되어 있다.

한글처리 전처리기는 XML-SRGS 문법을 입력 받아 컴파일러에서 추출 불가능한 한글을 유니코드화로 변환하여 파일로 출력하여 어휘 분석기의 입력으로 보내진다.

SRGS 어휘 분석기는 전처리기에 의해 출력된 XML-SRGS을 입력으로 받아 각각의 태그들을 분석하고, 태그와 텍스트를 분리한 후 불필요한 태그들과 주석 등을 제거 한 후 토큰을 반환한다. 인식대상에 관련된 태그를 찾아내기 위해 <표 1>과 같이 토큰을 정의하였다. <표 1>은 정의한 토큰의 일부를 나타낸다.

<표 1> XML-SRGS 토큰

| 정규표현식 | Token | 렉심의 예 |
|---|---|---|
| grammar | GRAMMAR | grammar |
| rule | RULE | rule |
| item | ITEM | item |
| "one-of" | ONEOF | one-of |
| ruleref | RULEREF | ruleref |
| root | ROOT | root |
| id | ID | id |
| uri | URI | uri |

SRGS 구문 분석기는 토큰을 입력 받아 각 태그에 해당하는 속성들과 문법을 검사한다. 파싱이 완료되면, 생성규칙에 의해 대응되는 EBNF 문법 코드를 생성한다. (그림 6)은 구문 중심 정의법( SDD : Syntax-Directed Definition)을 이용하여 XML-SRGS의 생성 규칙을 표기한 것으로 총 12개의 규칙을 정의하였다.

```
① grammar → ST GRAMMAR attrs ET rulelist ST SL GRAMMAR ET
          | error ST {}
② attrs   →  attrs ID EQ STR
          | attrs  ROOT EQ STR
          | attrs URI EQ STR
          | attrs  IDL EQ STR
          |
③ rulelist →  rulelist  rule
          |
④ rule →  ST RULE attrs ET rulecontent   ST SL RULE ET
⑤ rulecontent →  rulecontent oneof
          | rulecontent item
          | rulecontent ruleref
          | rulecontent example
          |
   (중략)
```

(그림 7) XML-SRGS 생성 규칙

생성된 구문 분석기의 파싱테이블은 74개의 상태수와 15개의 입력 심볼 및 12개의 비단말 기호로 구성하였다. 또한 의미규칙에 의해 요청되는 번역함수는 인식 대상을 배열에 저장하고 그 결과를 출력하는 4개의 함수로 구성하였으며 GeneratorGrammar클래스에 정의하였다.

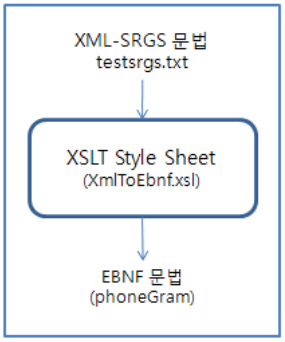
XML-SRGS를 EBNF로 변환함에 있어 다음과 같은 제한조건을 두었다.

1) 주석과 DTD는 무시한다.

2) 문법에 사용하는 규칙은 내부 규칙에 한한다. 즉, 외부에서 사용되는 규칙은 적용하지 않는다.

### (2) XSLT를 이용한 EBNF Converter

XSLT(Extensible Stylesheet Language Transformations ) 은 W3C에서 제정한 표준으로 XML 문서를 다른 형태의 문서로 변환하는 마크업 언어이다. SRGS Spec에서는 XML-SRGS를 ABNF-SRGS로 변환하는 표준으로 XSLT를 사용할 것을 제시하고 있다.




(그림 8) XSLT Style Sheet를 이용한 EBNF Converter

본 논문에서는 두 번째 방법으로 EBNF Converter를 표준인 XSLT을 이용하여 구현하였다.

XML-SRGS를 EBNF로 변환하기 위한 Sytle Sheet는 SRGS 표준에서 제공하는 grammar-transformer.xsl[10] 파일을 사용하였다. 변환 결과가 EBNF Grammar 형태에 맞게 출력될 수 있도록 수정하여 사용하였다.

(그림 9)은 본 변환기에서 사용한 XSLT Style Sheet의 일부이다.

```
- <xsl:template match="srgs:rule">
    <xsl:call-template name="addexamples" />
    <xsl:value-of select="@scope" />
    $
    <xsl:value-of select="@id" />
    =
    <xsl:apply-templates />
    ;
  </xsl:template>


- <xsl:template match="srgs:one-of">
    (
    <xsl:apply-templates />
    )
    <xsl:call-template name="addlang" />
  </xsl:template>
- <xsl:template match="srgs:one-of/srgs:item">
    <xsl:call-template name="addweight" />
    (
    <xsl:apply-templates />
    )
    <xsl:call-template name="addlang" />
    <xsl:call-template name="addrepeat" />
    <xsl:if test="not(position()=last())">|</xsl:if>
  </xsl:template>
- <xsl:template match="srgs:item">
    (
    <xsl:apply-templates />
    )
    <xsl:call-template name="addlang" />
    <xsl:call-template name="addrepeat" />
  </xsl:template>
```

(그림 9) XSLT Style Sheet : xmlToEbnf.xsl

XML-SRGS를 EBNF로 변환함에 있어 역시 다음과 같은

제한조건[6]을 두었다.

1) 주석은 변환되지 않는다.
2) ENBF에서 유효하지 않은 토큰들은 다루지 않았다.
3) 참조하는 외부 파일이나 media type에 대한 토큰은 그대로 저장하였다.

3)번에 대한 조건은 본 시스템에서는 무시하였다. 즉 문법에 사용하는 규칙은 내부 규칙에 한해서 작성하였다.

## 4. 실험 및 결과

본 논문에서 구현한 시스템의 성능을 확인하기 위해 테스트용 웹 페이지를 생성한 후 웹 페이지를 통해 XML-SRGS와 사용자의 음성파일을 입력 받은 다음 그 인식 결과를 텍스트로 반환 하였다. 본 실험을 위해 음성인식을 위한 인식기는 HTK[11]를 이용하여 구현한 인식기를 이용하였다.

본 논문에서 구현한 웹 서비스 소비자와 음성인식 웹 서비스 제공자는 <표2>와 같이 Microsoft의 .NET 환경에서 구현하였다.

<표 2> 구현환경

| 구분 | 웹 서비스 소비자 | 웹 서비스 제공자 |
|---|---|---|
| 운영체제 | Microsoft Windows XP | |
| 개발 플랫폼 | Microsoft Framework2.0 / Microsoft WSE 1.0 | |
| 개발 도구 | Microsoft Visual Studio 2005 | |
| 개발 언어 | C# , ASP.NET | |

EBNF Converter의 입력인 (그림 4)로 기술된 XML-SRGS가 변환된 EBNF 파일(그림 8)을 확인 한 결과 EBNF Converter의 기능이 정상적으로 동작되는 것을 알 수 있었다.


(그림 10) EBNF Converter 결과 : phoneGram

또한 phoneGram파일이 제대로 변환되었는지 확인하기 위해 이 파일을 이용하여 HTK가 인식 실험을 하도록 하였다. 음성녹음을 위해 사람이 없는 일반 강의실에서 오디오 샘플크기 16bit , 채널1(mono)오디오 샘플속도 16KHz, 오디오 형식 PCM 방식으로 녹음 환경을 구축하였다. 발성자는 성인 여자 3명, 성인 남자 2명을 대상으로 인식 단어를 각각 5번씩 발음 하였으며, 인식 결과는 100회 중 95번 정확히 인식하였다.

## 5. 결론

본 논문에서는 W3C에서 제안한 Speech Interface Framework 구조에서 음성인식을 원격 시스템에서 수행하고 Voice Browser가 인식 결과를 사용할 수 있는 웹 서비스 환경을 제공하는 새로운 SIF를 제안하였다.

또한 Voice Browser에서 사용하는 표준 XML-SRGS로 기술된 grammar를 음성인식 시스템에서 사용하는 EBNF로 기술된 grammar 변환해주는 EBNF Converter를 구현하였다.

여러 언어를 인식하는 여러 음성 엔진이 웹 서비스 제공자의 ASR 시스템이 설치된다면 Voice Brower는 언제든지 다른 언어도 인식이 가능할 것이다.

추후에 EBNF 뿐만 아니라 여러 음성엔진을 위한 gram-mar로 자동으로 변환할 수 있는 기능을 converter에 추가하고 나아가 SALT나 VoiceXML을 해석하여 음성인식 시스템과 Voice Browser 간의 소통이 가능하게 해주는 인터프리터를 설계 및 구현하고자한다.

## 참고문헌

[1] Mecanovic, D, and Shi, H. "Voice User Interface Design for a Telephone Application Using VoiceXML", Lecture Notes in Computer Science, No. 3399, Web Technologies Research and Development, pp. 1058-1061.

[2] Susan J.Boyce, "Natural Spoken Dialogue Systems for Telephony Application", Communications of the ACM, Vol.43, No.9, Sep. 2000.

[3] 오만수, "음성 인터페이스의 기술현황과 표준화 동향", 군산대 교육대학원, 석사 논문 2007.

[4] 홍기형, 정민화, "유비쿼터스 환경의 정보서비스를 위한 음성 기술 표준화 동향" , 정보과학회지 제24권 제1호, pp11-18., 2006. 01.

[5] Introduction and Overview of W3C Speech Interface Framework , http://www.w3.org/TR/voice-intro/ , W3C Working Draft 4 December 2000.

[6] Hunt, A. and McGlashan. S. , Eds. , W3C Speech Recognition Grammar Specification Version 1.0, March 16, 2004(see http://www.w3.org/TR/speech-grammar/).

[7] Kinfee Tadesse Mengistu, Andreas Wendenuth, "Telephone-Based Spoken Dialog System Using HTK-based Speech Recognizer and VoiceXML", Fortschritte der Akustik, 2007.

[8] 장준식, 윤재석, "Voice 브라우저의 설계 및 구현" , 한국해양정보통신학회 춘계종합학술대회지 제8권 제1호, 2004.

[9] IISPLLab, "IISPL Generator Specification", "http://www.wins.or.kr/IISPL Generator, 2008. 02.

[10] XSLT Style Sheet , http://www.w3.org/TR/speech-grammar/grammar-transformer.xsl

[11] 홍인숙 , "HTK를 이용한 음성 인식 시스템 구현", LAB Project, http://home.wins.or.kr/ , 2008. 02.