산업용 음성 DB 메타데이터 표준화[1)]

주영희, 홍기형
성신여자대학교 미디어정보학부

Standardization of XML based Meta-data for Industrial Speech Databases

Young-Hee Joo and Ki-Hyung Hong
School of Media and Information, Sungshin W. University

yhjoo@media.sungshin.ac.kr, khhong@sungshin.ac.kr

Abstract

본고에서는 산업용 음성 DB를 위한 XML 기반 메타데이터의 표준화에 대한 현재 상황과 표준화 활동에 대하여 소개한다. 산업용 음성 DB는 구축에 많은 시간과 비용을 요구하며, 양질의 음성 처리 시스템 (인식/합성/인증)의 개발을 위해서는 가능한 많은 양의 음성 데이터가 필요하다. 산업용 음성 DB 메타데이터 표준화는 서로 다른 기관에서 구축한 음성 DB의 공유와 재사용을 원활히 하기 위하여, 2004년 9월부터 요구사항 분석을 시작하여, 2005년 3월 초안이 완성되었다. 본 표준안은 음성 DB 메타데이터의 구조를 XML 기반으로 정의한 것이며, 음성 파일 이름, 화자 식별자, 음소 기호와 같은 구조 외의 표준화 대상에 대해서는 다루지 않는다. 이미 ETRI와 SiTEC [5]에서 XML 기반의 메타데이터 구조와 내용 표준안을 제안한 바 있으나. [5]에서 제안한 구조는 평면 구조를 취하고 있어내용의 중복성 등의 단점이 있어, 이를 보완하여 음성 DB 데이터 모델을 객체지향 방식으로 설계하였다.

Ⅰ. 서론

산업용 음성 데이터베이스(이하 산업용 음성 DB)는 자동차, 로봇, 가전제품 등 다양한 산업 제품을 위한 음성 사용자 인터페이스(음성인식/합성)의 개발에 있어서 필수적이며 가장 첫 단계에서 구축되어야 한다.

다양한 산업 제품의 사용 환경을 반영한 산업용 음성 DB의 효과적인 구축 및 활용은 음성 사용자 인터페이스의 효율적인 개발과 성능향상에 있어 매우 중요하며, 다양한 산업 제품의 사용자가 필요로 하는 음성언어정보가 충실히 표현된 DB, 개발자가 기술개발에 용이하게 활용할 수 있는 구조화된 DB 구성이 요구된다. 그러나 지금까지 국내에서는 산업용 음성 DB 표준화에 대한 필요성은 충분히 인지하고 있었으나 구체적 노력이 미미하였다. 최근 대량의 산업용 음성 DB 구축이 진행되고 있으나, 각기 다른 기관에서 구축한 DB의 호환성에 문제가 발생하고 있다. 이에 따라 산업용 음성 DB의 호환성을 높이고, 활용성을 증대시키기 위해 메타데이터 표기방법 및 구조의 표준화가 절실히 필요한 시점이다.

본고는 산업용 음성 DB 메타데이터 표준안으로, 산업용 음성 DB의 구축 환경 정보, 구축에 참여한 화자(speaker) 정보, 화자가 발성한 단어 또는 문장 목록, 음성의 디지타이징 방법 및 저장 음성 파일의 개별 정보, 그리고 다양한 전사 정보의 표준 표시 방법을 제시한다. 세계적 공개 표준인 XML(eXtended Markup Language) 기반으로 개발하여, 향후 확장성을 보장하고, 표준화된 메타데이터 관리 및 구축 도구의 개발을 통하여 표준의 활성화를 꾀하고자 한다.

국외의 경우 프랑스의 ELRA(The European Language Resources Association) [1]와 미국의 LDC(The Linguistic Data Consortium) [2]를 주축으로 음성 DB 구축 및 표준화를 꾸준히 수행해 오고 있다. 표준화 대상으로는 수집시스템, 수집환경, 음성DB, 전사방법, 검증방법 등 다양한 분야에서 수행하고 있다. 그러나 국제 표준화 활동은 영어권의 음성 특성을 반영하고 있어, 우리말 음성의 고유한 특성을 반영해야 할 국내 표준화에 그대로 채택하여 사용하기에 적합하지 않은 부분이 있다 [3].

국내의 경우 음성 DB 메타데이터 표준안은 2003년도에 ETRI(한국전자통신연구원) 음성/언어정보 연구센터에서 개발한 공통 음성 DB 표준안 [4][5]이 있으나, 음

1) 본 고는 한국기술표준협회와 한국 SIT산업협회의 "표준 산업용 음성 DB 메타데이터 및 구축 도구 개발" 사업의 일환으로 수행되었으며, SiTEC의 협조와 지원으로 이루어 졌다.

성 DB 표준안은 정보통신망 기반 음성 기술 개발을 위한 음성 DB 메타데이터 규격으로 개발되었다. 음성 DB 정보를 크게 기본정보, 음성정보, 전사정보, 화자정보, 환경정보, 파일정보, 기타정보로 나누었다. 공통 음성 DB 표준안은 국내에서 처음으로 시도한 음성 DB 메터데이터 규격이라는 점에서 의미가 있으나, 다음과 같은 단점이 있다.

첫째, 음성 DB 메타데이터를 단순히 2단계 구조로 정의하였다. 이러한 평면적인 구조는 음성 DB를 구성하는 객체 사이의 집합적 포함관계나 연관관계를 정확하게 표현하기에 부족하다. [5]에서 기술한 모델은 하나의 음성 파일에 대한 것인지, 음성 DB, 즉 다수의 음성 파일에 대한 것인지가 불명확하다.

둘째, 화자, 음성 녹음채널, 음성 파일 등 DB를 구성하는 객체의 식별자와 이를 참조하는 객체 참조 개념이 결여되어 있어, 동일한 정보의 중복 및 특정 음성 파일의 화자의 식별이 어렵다. 예를 들어, 발화자가 100명이고 발화 문장이 100개가 존재할 경우 메타데이터에 총 10,000개의 음성 파일에 대한 정보가 존재하게 된다. 이 경우 발화자 한 명당 발화 문장 100개씩을 메타데이터 문서에 중복하여 입력해야 한다.

본고에서 제안하는 산업용 음성 DB 메타데이터는 국제기관인 LDC, ELRA의 용어 및 분류정보를 참조하고 [5]의 모델을 보다 객체 지향 구조로 보완하였다.

Ⅱ. 산업용 음성 DB 모델

자동차, 완구, 가전 기기, 산업용 기기 등 다양한 소음 및 주변 환경에서 높은 성능의 음성 기술(인식/합성/인증/코딩/강화) 개발 및 시험을 위해서는, 해당 산업용 음성 기술이 사용되는 동일한 환경에서 대량의 음성 데이터베이스의 구축이 필수적이다. 즉, 자동차에서 사용하려고 하는 음성 인식 기술의 개발을 위해서는 자동차의 주행 환경에서 사용자가 내릴 수 있는 음성 명령을 녹음하여, 기술의 개발에 사용하여야 자동차 환경에서 성능이 높은 인식 기술을 개발할 수 있다. 또한 음성은 화자(speaker)의 특성에 따라 동일한 환경이라고 하더라도 많은 차이가 있다. 사투리의 구사여부, 개인의 구강 구조의 특성, 성별, 나이 등에 따라 동일한 환경, 동일한 단어를 발성하더라도 음성 신호에서 많은 차이를 보이므로, 화자 독립형의 음성 기술의 개발에서는 가능한 많은 수의 화자로부터 음성을 수집하여야 한다. 따라서 완구용 음성 기술을 위한 음성 DB, 자동차용 음성 기술을 위한 음성 DB 등과 같이 산업용 음성 기술 개발을 위해서는 해당 기기의 사용 환경에서 많은 수의 화자로부터 음성을 수집하는 것이 필수적이다.

산업용 음성 DB는 산업 응용에서 적합한 음성 기술(인식/합성/인증/코딩/강화) 개발 및 시험을 위하여, 다수의 화자로부터 채취한 음성 파일의 집합으로 정의한다.

산업용 음성 DB 메타데이터는 산업용 음성 DB의 기본정보(구축 기관정보 등), 음성파일의 디지타이징 방법, 화자 정보 등과 같이 산업용 음성 DB에 대한 정보를 말한다.

본 메타데이터 규격의 대상이 되는 산업용 음성 DB는 그림 1과 같은 구성을 전제로 한다.

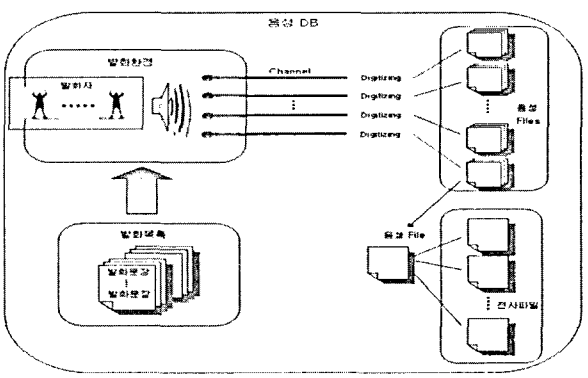

그림 1 산업용 음성 DB 메타데이터 구성

1) 음성 DB 내의 음성 파일은 다양한 발화환경에서 발화한 것을 수집한 것으로 가정하였다. 발화환경은 방음실, 백화점, 자동차 안, 가정과 같은 화자의 주변 환경을 의미한다.

2) 하나의 음성 DB는 화자의 음성을 여러 가지 방법으로 수집하기 위해서 다수의 녹음 채널이 있을 수 있다. 채널이란 녹음 시 사용하는 마이크, 디지타이징 형식, 통신망(전화망, IP network 등), 발화환경 등을 서로 다르게 할 수 있다.

3) 하나의 음성파일은, 화자가 발화목록의 한 문장을 발화한 음성을 특정한 채널로 녹음하고 디지타이징하여 저장한 것이다.

4) 하나의 발화목록은 여러 개의 발화문장으로 구성될 수 있다.

5) 하나의 음성 파일은 여러 종류의 전사정보를 가질 수 있다.

6) 화자 한 명은 다수의 발화 목록을 발성할 수 있다.

7) 발화시점을 다르게 하여 화자는 같은 발화목록을 여러 번 반복 발성하여 녹음할 수 있다.

Ⅲ. 산업용 음성 DB 메타데이터 모델

음성 DB의 구성요소는 화자, 발화문장, 채널, 음성파일, 전사정보이다. 음성 DB 메타데이터 구성요소들의 관계는 그림 2와 같다. 화자 한 명이 여러 개의 문장을

반복 발화할 수 있으므로 발화자와 발화는 1:n의 관계이다. 발화문장 하나는 여러 화자가 반복하여 발화하므로, 발화와 발화문장의 관계는 n:1이 된다. 그림 2에서 '발화'는 화자가 발화하는 행위 자체를 하나의 개체로 바라보기 위한 약한 개체(weak entity)이다. 화자 한 명이 하나의 발화문장을 발화할 때, 다수의 채널로 녹음할 수 있으므로 채널별로 음성파일이 생성되어 발화와 음성파일은 1:n의 관계이다. 이렇게 생성된 한 개의 음성파일에는 전사정보가 전사의 종류에 따라 여러 개 존재할 수 있다. 따라서, 하나의 음성파일과 전사정보는 1:n의 관계를 가진다.

하나의 음성파일과 관련 있는 화자, 채널, 발화문장, 전사정보와의 구체적인 관계는 그림 3과 같다. 음성파일에는 식별자(*File_Id*)가 있으며, 누가 발화하였는지 나타내기 위해 발화자의 식별자(*Speaker_Id*)를 참조한다. 어느 채널을 통해 녹음된 파일인지 나타내기 위해 채널의 식별자(*Channel_Id*)를 참조한다. 어떤 문장에 대한 발화인지를 나타내기 위해 발화문장 식별자(*Utterance_Id*)를 참조하며, 해당 음성파일과 관련한 전사정보를 모두 참조(Set of *Transcription_Id*)하고 있다. 그림 3에서 Set of *Transcription_Id*는 XML 스키마로 구현하면서 File의 하위 엘리먼트인 Transcription을 정의하는 구조로 표현하였다.

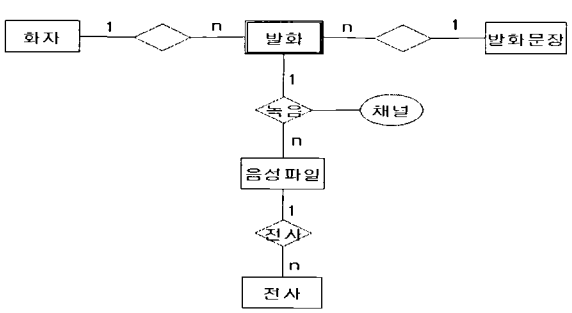

그림 2 음성DB 메타데이터 구성요소 및 관계도 (ER 모델)

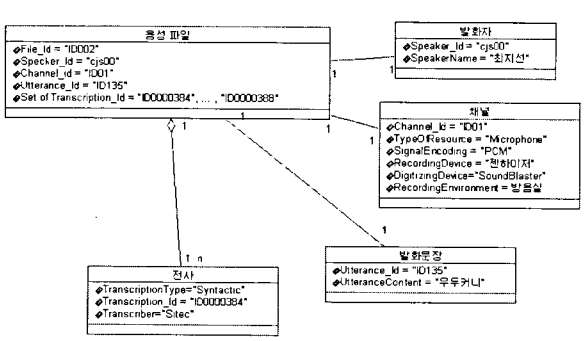

그림 3 음성 파일 하나와 다른 정보와의 관계도

Ⅳ. XML 기반 음성 DB 메타데이터

1. 메타데이터 설계

Ⅲ장에서 살펴본 바와 같이 산업용 음성 DB는 수집 채널, 화자 정보, 발화목록, 음성파일, 전사 정보로 구성된다. 제안하는 메타데이터 규격은 기존의 SiTEC과 ETRI에서 추진한 표준화 작업[4][5]과 비교하여 XML[6]의 특징을 최대한 살릴 수 있도록 정의하였다. 화자, 채널, 발화목록, 파일, 전사 정보에 각각의 식별자를 두고 구성요소 사이의 관계를 식별자의 참조로 표시할 수 있도록 하였다. 또한 연관되는 엘리먼트들의 그룹핑과 계층구조를 도입하여, 메타데이터의 구조를 체계화하여, 향후 음성 DB의 검색이나 재사용이 용이하도록 설계하였다.

XML 스키마로 정의한 메타데이터 구조는 그림 4와 같다. 본 메타데이터 정의에서 사용한 엘리먼트와 속성의 이름은 ELRA와 LDC에서 부분적으로 참고하였다[1][2]. 그림 4의 ①번, SpeechDBMetadata는 메타데이터인 XML 문서의 루트 엘리먼트이다. ②번 General은 해당 음성 DB의 기본정보를 위한 엘리먼트이다. ③번 Channels는 음성 DB 수집에서 사용한 다수의 채널에 대한 정보를 위한 엘리먼트이다. Channels 엘리먼트의 하위 엘리먼트로 Channel 엘리먼트를 다수 가질 수 있으며, 각 Channel 엘리먼트는 특정 수집채널에 대한 정보를 기술한다. ④번 Speakers는 화자 정보를 나타내는 엘리먼트이다. Speakers는 다수의 Speaker 엘리먼트를 가질 수 있고, 하나의 Speaker 엘리먼트는 특정 화자의 정보에 해당한다. ⑤번 UtteranceLists는 DB 수집에서 사용한 발화목록 정보이다. 하나의 DB 수집에서 다수의 발화목록이 있을 수 있으므로, 하위 엘리먼트로 다수의 UtteranceList 엘리먼트를 가진다. ⑥번 Files는 DB를 구성하는 각 음성파일에 대한 정보를 담고 있는 File 엘리먼트를 DB에 존재하는 음성파일의 개수만큼 하위 엘리먼트로 가진다. ⑦번 Trans- cription은 상위 엘리먼트인 File이 기술하고 있는 음성파일에 존재하는 전사정보를 위한 엘리먼트이다. 하나의 음성파일은 하나 이상의 전사정보가 있을 수 있으므로, 하나의 File 엘리먼트는 다수의 Transcription 엘리먼트를 가질 수 있다.

2. 메타데이터 구현 예

그림 5는 제안하는 산업용 음성 DB 메타데이터 규격을 특정 음성 DB에 적용한 예이다.

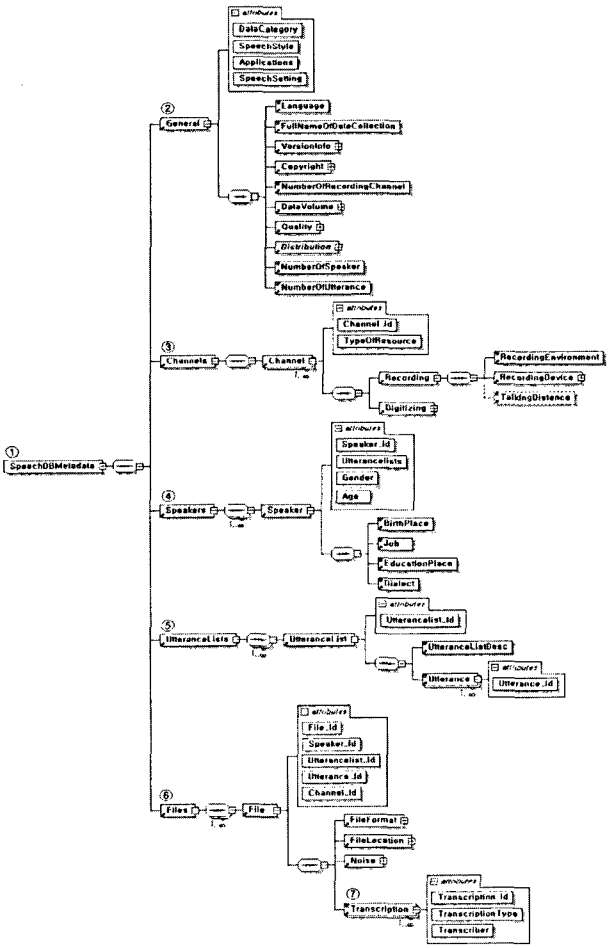

그림 4 산업용 음성 DB 메타데이터 스키마 구조

V. 결론 및 향후과제

본고에서 제안한 산업용 음성 DB를 위한 메타데이터 표준안의 보다 상세한 내용은 [7]에서 기술하였으며, ETRI와 SiTEC에서 기존에 제안한 규격 [5]를 객체 지향 구조로 보완하였다. 화자, 채널, 발화목록, 파일, 전사 정보에 각각의 식별자를 두고 구성요소 사이의 관계를 식별자의 참조로 표시할 수 있도록 하였다. 또한 연관되는 엘리먼트의 그룹핑과 계층구조를 도입하여, 메타데이터의 구조를 체계화하여, 향후 음성 DB의 검색이나 재사용이 용이하도록 설계하였다.

향후 표준화 계획은 본 표준안을 표준협회를 통하여 국가 표준안으로 제안할 예정이다. 또한, 대용량의 산업용 음성 DB의 메타데이터를 본 표준안에 따라 쉽게 구축할 수 있도록 하는 산업용 음성 DB 메타데이터 구축도구의 개발이 진행 중이다.

```
<SpeechDBMetadata xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
 <General SpeechStyle="Elicited" Applications="VoiceControl" DataCategory="VCVSequence">
   <Language>KOR</Language>
   <FullNameOfDataCollection>멀티모달 음성인터페이스를 위한 DB</FullNameOfDataCollection>
   ....
 </General>
 <Channels>
  <Channel TypeOfResource="Microphone" Channel_Id="ID01" DataFormat="FloatingPoint">
        <Recording>
                <RecordingEnvironment>방음실</RecordingEnvironment>
                <RecordingDevice>...</RecordingDevice>
                <TalkingDistance/>
        </Recording>
        <Digitizing>
                <DigitizingDevice>
                        ...
                </DigitizingDevice>
                <SignalEncoding type="PCM"/>
                <ByteOrder order="LitteleEndian"/>
                <BitsPerSample>8Bit</BitsPerSample>
                <SamplingRate>150kHZ</SamplingRate>
        </Digitizing>
   <Channel/>
   <Channel> .... </Channel>
    ...
</Channels>
<Speakers>
 <Speaker Speaker_Id="csy00" Age="20" Utterancelists="ID03 ID08" Gender="Female" >
        <SpeakerName>천신영</SpeakerName>
        <BirthPlace>서울</BirthPlace>
        <CurrentlyResidence>서울</CurrentlyResidence>
        <CurrentlyResidenceDuration>20</CurrentlyResidenceDuration>
   ...
 </Speaker>
 <Speaker/>
        ...
</Speakers>
<UtteranceLists>
  <UtteranceList Utterancelist_Id="ID03">
        <UtteranceListDesc>1번부터 105번까지 발화목록</UtteranceListDesc>
        <Utterance  Utterance_Id="ID000001">영</Utterance>
        <Utterance  Utterance_Id="ID000002">공</Utterance>
        ....
        <Utterance/>
 </UtteranceList>
 <UtteranceList/>
  ...
</UtteranceLists>
<Files>
  <File  File_Id="ID001" Utterancelist_Id="ID03" Channel_Id="ID01" Utterance_Id="ID000001"
   Speaker_Id ="cjs00">
        <FileFormat format="Wave">
                <HeaderSize>0Byte</HeaderSize>
                <FileSize>1024Byte</FileSize>
        </FileFormat>
        <FileLocation>
                <SpeechFileName>dig01.wav</SpeechFileName>
                <DirectoryPath>D:\data\digit\female\f2cjs0</DirectoryPath>
        </FileLocation>
        <Noise NumberOfNoise="1">
                <NoiseSection>
                        <NoiseType>음악</NoiseType>
                        <NoiseStart Time>00:01:34</NoiseStart Time>
                        <NoiseEndTime>00:01:50</NoiseEndTime>
                </NoiseSection>
        </Noise>
        <Transcription TranscriptionType="Syntactic" Transcription_Id="ID0001" Transcriber="Sitec">
        영</Transcription>
   </File>
   <File/>
    ...
</Files>
</SpeechDBMetadata>
```

그림 5 XML 기반 음성 DB 메타데이터 예제

참고문헌

[1] ELRA http://www.elra.info/services/speech_1.4.rtf, 2004.

[2] LDC http://www.ldc.upenn.edu/Catalog, 2004.

[3] 홍기형, 이욱재, "국제 음성기술 표준화 동향과 대응", *음성통신 및 신호처리 학술대회 논문집*, 20권, 1호, pp.185-188, 2003.

[4] 김상훈, 이용주, "음성 DB 표준화", *음성통신 및 신호처리 학술대회 논문집*, 20권, 1호, pp.181-184, 2003.

[5] 김상훈, 이영직, 한민수, "음성 DB 부가 정보 기술 방안 표준화를 위한 제안", *대한음성학회*, 말소리, 제47호, pp.110-119, 2003.

[6] XMLSchema http://www.w3.org/TR/xmlschema, 2004.

[7] 주영희, 홍기형, "산업용 음성 DB를 위한 XML 기반 메타데이터", *대한음성학회*, 말소리, 제 55호, pp.77-91, 2005.