

# 음성 DB의 메타데이터 표준화<sup>1)</sup>

김상훈

ETRI 음성/언어정보연구센터

## Meta-data Standardization of Speech Database

Sanghun Kim

ksh@etri.re.kr

### Abstract

In this paper, we introduce a new description method of annotation information of speech database. As one of structured description methods, XML based description which has been standardized by W3C will be applied to represent metadata of speech database. It will be continuously revised through the speech technology standard forum during this year.

### I. 표준화 필요성

최근 각종 휴대형 정보단말기, CTI 및 텔레메틱스 등 사용자 편의성을 증대하기 위한 음성인터페이스(VUI: Voice User Interface)가 매우 중요한 요소기술로 부각되고 있다. 특히, 각종 정보기기의 소형화와 이동성에 따라 언제, 어디서나 정보를 획득할 수 있는 정보검색 서비스가 향후 2-3년 내에 일반화될 것으로 예상되고 있고(가트너그룹 조사에 의하면 2004년까지 e-commerce 고객 중 40%가 무선을 통해 접속할 것으로 예측), 이러한 정보기기의 자연스러운 입력방식으로 음성인터페이스의 적용이 필수적인 요건으로 될 것이다. 최근 국제표준화 기관인 ETSI, ITU-T, IETF에서는 분산음성인식기술(DSR) 국제표준화를 추진하고 있

으며, 마이크로소프트/CISCO/Intel이 주도하고 있는 SALT (Speech Application Language Tag) 포럼에서는 멀티모달 환경에서의 음성인터페이스에 대한 표준화 작업을 수행하고 있는 바, 음성기술은 가까운 시일 내에 일반인들에게 매우 친숙한 기술이 될 것으로 예상된다.

음성처리기술을 개발하기 위해서는 음성/텍스트 DB가 필수적인 기반요소로 요구된다. 이미 미국, 유럽 등의 선진국에서는 산/학/연 공동으로 연구기관을 설립하여 공통 음성 DB를 구축 보급하고 있다[1]. 미국에서는 DARPA를 통하여 수집한 DB를 기반으로 1990년대 초에 LDC(Linguistic Data Consortium)를 구성하였고[2], 유럽에서는 1995년에 비영리기관으로 ELRA(European Language Resource Association)를 설립하였다. 국내에서도 최근 100여개의 음성 전문업체가 설립되어 다양한 분야에 음성정보기술을 응용하고 있다. 그 동안 대다수의 음성유관기관과 음성처리업체들은 개별적으로 소규모의 DB를 구축하여 연구와 자체 엔진개발에 활용하여 왔으나, DB 구축에 많은 시간과 비용이 소요되어 외국업체에 비해 음성정보처리기술개발 경쟁력이 점차 떨어지고 있는 상황이다. 이에 따라 정부도 음성 DB의 필요성을 인식하고 ETRI, 원광대 SiTEC을 통해 정부출연 대량의 공통 음성/텍스트 DB 구축이 진행되고 있다[3][4][5].

공통음성 DB는 사용자가 필요로 하는 음성언어정보가 충실히 표현된 DB, 사용자들이 기술개발에 용이하게 활용할 수 있는 구조화된 DB 구성이 요구된다. 그러나

1) 본 연구는 정보통신부 출연 “언어정보처리기술개발” 사업으로 수행되었습니다.

지금까지 국내에서는 DB 표준화에 대한 필요성은 충분히 인지하고 있었으나 구체적인 노력이 미미하였고, 최근 대량의 공통 음성 DB 구축이 진행되고 있는 바 DB의 호환성에 문제가 발생하고 있는 실정이다. 이러한 가운데 최근 음성기술 표준화를 위한 포럼이 구성되고 용어표준, 공통명령어 표준, DB 표준 등 음성처리기술에 필요한 표준을 차근차근 진행하고 있는 것은 매우 다행스러운 일이다. 본 논문에서는 공통 음성 DB의 호환을 용이하게 하고 활용성을 증대시키기 위한 노력의 일환으로 ETRI에서 수행하고 있는 음성 DB 메타데이터 표준화 방안을 소개하고, 이러한 작업이 현재 포럼을 통해 추진되고 있음을 알리고자 한다.

## II. 표준화 동향

국의동향으로, 유럽에서는 프랑스의 ELRA(European Language Resource Association)를 주축으로 DB 구축, DB 표준화, 평가, 배포 목적으로 SpeechDat 프로젝트를 수행한 바 있다. 표준화 대상으로는 수집시스템, 수집환경, 음성언어정보, 음성 DB, 전사방법, 검증방법, 평가방법 등 다양한 분야에 걸쳐 표준화를 추진하였다. 본 프로젝트에는 벨기에 L&H, 독일의 Siemens, 이탈리아의 CSELT 등 모두 17개국 기관이 참여한 바 있다[6].

COCOSDA(International Coordinating Committee on Speech Databases and Assessment)는 “음성입출력 평가법 및 음성 데이터베이스”에 관한 워크샵으로 주로 음성인식 및 합성시스템의 성능평가 방법 및 음성 데이터베이스를 다루고 있다. EAGLE(Expert Advisory Group on Language Engineering Standards)는 EU 지원으로 구성된 전문가 그룹으로 언어공학적 응용을 위한 각종 스펙의 가이드라인을 제시하고 음성 및 언어과학에 대한 연구 및 응용에 대한 광범위한 컨설팅과 기존에 구축된 음성 언어자원의 평가 및 효율적인 구축을 위한 방법론을 제시해 주고 있다. 한편 International Standards Organization(ISO)에서는 언어정보를 표현하기 위한 데이터 모델 표준화를 위해 ISO/TC 37/SC 4 라는 위원회를 구성하였다. 이 위원회에서는 언어자원(텍스트 코퍼스, 사전류, 음성)의 생성, 코딩, 처리, 관리 방법을 체계화하고, 용어표준(ISO/TC 37에서 이미 다루었음) 보다는 언어자원의 데이터 모델링, 마크업, 데이터 교환, 평가부분에 초점을 맞추어 언어자원의 효율화를 높이는 방향으로 표준화 작업을 진행하고 있다. 또 위원회는 ISLE(International Standards for Language Engineering)에서 수행한 표준화안과 EAGLES에서

작업한 표준화안을 ISO에 통합하고 있으며, EAGLES/ISLE에 포함하지 않은 동양권 언어자원에 대한 표준화도 확장하여 진행하고 있다.

국내동향으로는, 2001년 KAIST 전문용어공학연구센터에서는 과학기술부와 한국과학기술평가원의 기술용역사업에 의한 “대용량 음성(음향)/언어/영상 DB 구축 및 표준화”과제에 대한 사업발표회를 개최하였다. 이 발표회에서 음성/언어제품 개발 및 제품 평가를 위한 평가용 DB와 개발용 DB에 대한 설명회와 공개토론회를 갖고 시범 CD-ROM 및 자료집을 배포한 바 있다. 한국정보보호진흥원과 한국전자통신연구원은 생체인식용 생체특징(지문, 홍채, 얼굴, 음성)의 데이터 교환을 지원하기 위한 생체데이터 포맷 표준화 작업을 진행중에 있고 표준 데이터 구축도 병행하여 추진하고 있다. 최근에는 산업자원부 지원 표준화 포럼에서 음성정보 처리 관련 용어 정리를 추진하였고, 현재 보완 작업 중이다. DB 표준화와 관련, 원광대 SiTEC에서는 음소분할기준, 운율정보표기 체계화 등 주로 학술적인 목적으로 표준화를 시도하고 있으며[3], ETRI 음성/언어정보연구센터에서는 SiTEC과 공동으로 음성 DB 단체 표준화를 추진하고 있다.

## III. 음성 DB 표준화 제안

ETRI에서는 W3C(World Wide Web)에서 추진하고 있는 XML(Extensible Markup Language)을 사용하여 공통 음성 DB의 음성언어정보 표기방안을 표준화 할 계획이다. XML은 데이터/문서 교환을 용이하게 만들기 위한 W3C에서 추진하고 있는 차세대 마크업 표준으로 데이터 표현과 데이터 접근에도 매우 강력한 기능을 가지고 있다.

한편, 음성 DB를 XML로 표기하기 위해서는 메타데이터 표기를 위한 의미태그 선정, 메타데이터 표현 구조 그리고 편집/검색하기 위한 도구가 필요하다. 이들 중 현재 음성 DB 메타데이터 표현용 의미태그 선정작업을 음성정보처리기술 포럼을 통해 추진하고 있으며, 1차 선정된 초안은 <표 1>와 같다. 1차 안으로는 기본정보 등 7개의 대분류와 42개의 속성으로 의미태그를 분류하였다. 이는 ETRI에서 그 동안 DB 구축 경험을 바탕으로 공통 음성 DB에 반드시 필요하다고 판단되는 음성언어정보를 태그로 할당하였으며, 현실적으로 가능한 정보를 우선 표기하고 추후 필요로 하는 정보에 따라 태그를 정의하여 추가/삭제할 계획이다. <그림 1>은 메타데이터 표기용 의미태그를 이용하여 XML 스크립트를 작성한 예를 보여주고 있다.

XML을 사용할 경우, DB의 카타고리별 분류, 소팅

등 사용자가 원하는 정보를 용이하게 얻을 수 있는 부가적인 장점이 있다.

<표 1> 음성언어정보 메타데이터 XML 태그 정의(안)

대분류	속성	
	영문표기	의미
기본 정보 (DB_Info)	Language	언어
	Version	버전
	ApplicationCategory	응용분야
	NumberOfSpeaker	발성화자수
	NymberOfUtterance	발화수
	DataCatagogy	DB종류
	RecordingDate	녹음날짜
	FilingDate	수정날짜
	RevisionHistory	수정기록
	Distributer	수행기관
음성 정보 (Wave_Info)	SamplingRate	주파수
	NumberOfByte	바이트수
	ByteOrder	바이트순서
	EncodingLaw	인코딩방법
	NumberOfBit	비트수
	NumberOfChannel	채널수
	SignalToNoiseRatio	SNR
전사 정보 (Annotation_Info)	SpeechSection	음성구간
	NoiseSection	잡음구간
	RecordingTrans	녹음문장
	SpokenTrans	철자전사
	PhoneticTrans	발음전사
화자 정보 (Speaker_Info)	SpeakerName	화자명
	Sex	성별
	Age	나이
	Region	지역
	Dialect	방언
환경 정보 (Environment_Info)	RecordingEnviron	발성환경
	NoiseEnviron	잡음환경
	RecordingDevice	녹취장비
	DigitizingDevice	A/D장비
	CommunicationDevice	통신종류
파일 정보 (File_Info)	FileCategory	파일종류
	FileName	음성파일명
	DirectoryPath	파일위치
	HeaderSize	헤더크기
	FileLength	파일길이
	FileFormat	파일포맷
	NumberOfRepeat	반복차수
	TimeInterval	녹음주기
	Distance	발성거리
기타 정보 (Miscellaneous_Info)	QualityStatus	품질상태

#### IV. 향후 추진 내용

음성 DB 표준화와 관련 향후 ETRI에서 추진할 연구 내용은 다음과 같다. 첫번째, 각 기관간에 상이하게 사용하고 있는 DB 표현방식(파일명명, 디렉토리 구조, 파일포맷, 전사방법 등)을 통일하고자 하며, 이는 SiTEC과 공동작업을 통해 추진하고자 한다. 두번째, 음성 DB의 부가정보인 음성언어 메타데이터를 XML로 표기하기 위한 국외 표준화 현황을 조사하고, 이로 부터 포럼에서 자체적으로 국내표준을 새롭게 작성해야 하는지 아니면 국제표준을 따라야 하는지 등 음성 DB의 XML 표기 표준화 방향을 결정하고자 한다. 국내표준을 작성할 경우, 국내 공통음성 DB의 음성언어 부가정보 XML 표기 first draft를 2003년 11월까지 관련 전문가의 의견 수렴 후 작성하고 2003년 12월 공청회를 거쳐 표준화 가능한 부분을 우선 표준화하고자 한다.

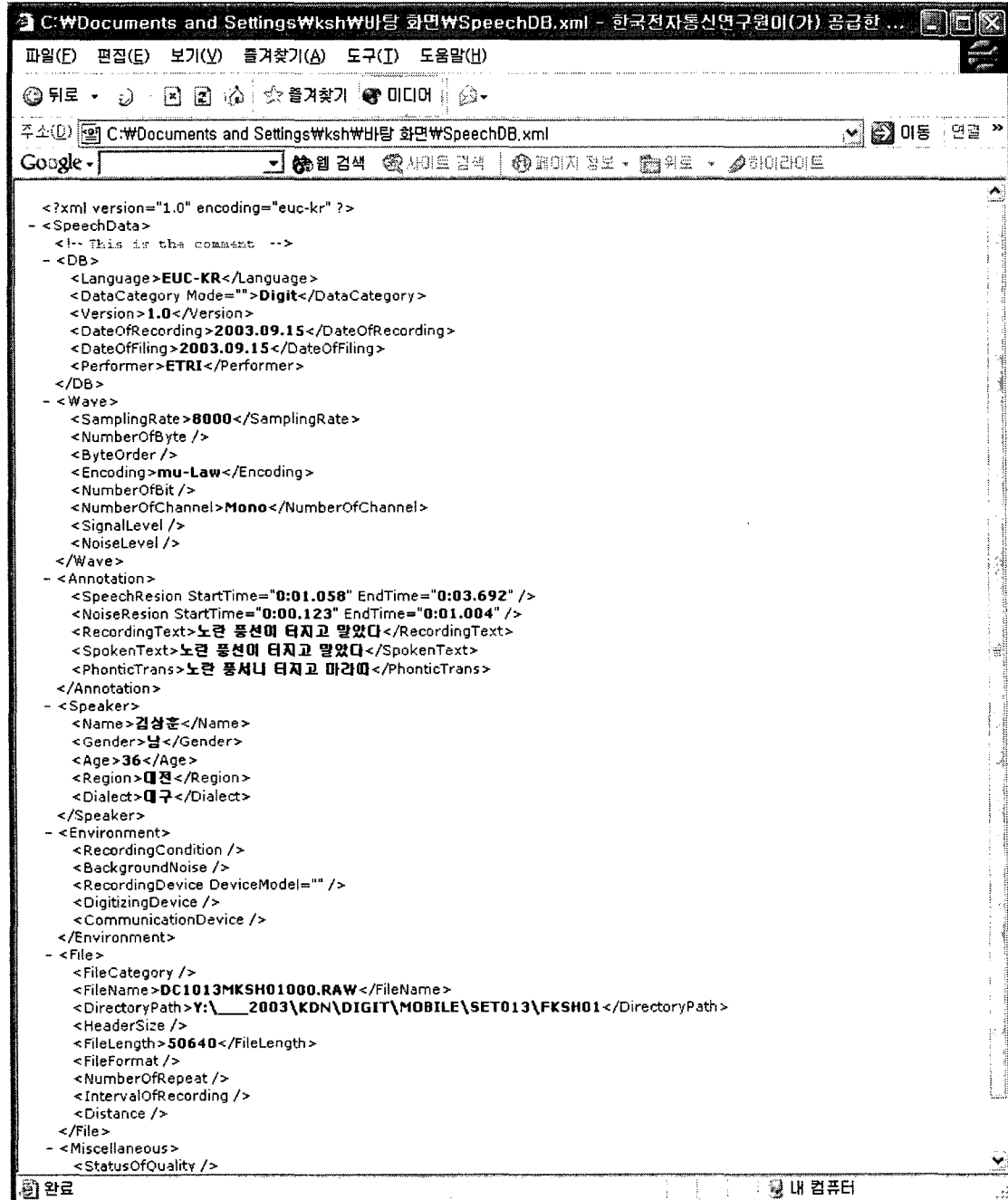
#### V. 결론

XML 기반 음성언어정보의 메타데이터 표준화는 음성 DB의 사용을 매우 용이하게 해주며, 각 기관간의 DB 공유도 활성화할 수 있어 국내 음성자원의 효용성을 극대화할 수 있는 계기가 될 것으로 보인다. 그러나 ISO 음성언어자원 표준화 위원회에서도 지적한 바 있듯이, 다양한 언어현상(linguistic phenomena)을 표기하는데 여러 가지 이론적 접근 방법이 존재하기 때문에 하나로 통일하는 것이 매우 어렵다는 것과 여러 프로젝트를 통해 수년동안 구축한 방대한 언어자원과 소프트웨어가 새로운 표준이 채택됨에 따라 무용지물이 될 수 있다는 불안감으로 인해 표준화가 어렵다는 것이다. 또한 현재 진행하고 있는 표준화 작업에 대해 음성언어 전문가들에게 충분히 인지되지 못하고 있는 실정이다. 따라서 이러한 문제를 조금이나마 해소하기 위해서는 관련 전문가들의 적극적인 참여를 통해 충분한 검토와 표준화 확산 노력이 필요하다.

#### 참고문헌

- [1] ETRI 음성/언어정보연구센터 <http://voice.etri.re.kr>.
- [2] LDC home page <http://www ldc.upenn.edu>.
- [3] 김봉완, 이용주, "음성정보기술산업지원센터의 음성 코퍼스 구축 현황 및 계획," 한국음향학회 하계학술대회논문집, 제21권, 제1(s)호, pp.49-52, 2002.
- [4] 김상훈, 박문환, 김현숙, "통신망환경 한국어 공통 음성 DB 구축", 대한음성학회 춘계학술발표대회, pp.23-26, 2003.

- [5] 김상훈, 오승신, 정호영, 전형배, 김정세, "공통음성 DB 구축," 한국음향학회 하계 학술대회 논문집, 제 21권, 제1(s)호, pp.21-24, 2002
- [6] SpeechDat Technical Report "Specification of speech database interchange format".



<그림 1> 음성언어정보 메타데이터 XML 실현 예