음성DB구축을 위한 국제간 활동현황
(COCOSDA'94에서 발표된 내용을 중심으로)

창원대학교 제어계측공학과
조 철 우

International Cooperative Works for Preparing Speech Database

Dept. of Control and Instrumentation Enginering, Changwon National University
Jo, Cheol-Woo

요 약

최근 음성처리기술의 성교화, 고도화를 위해서 대량의 다양한 데이타베이스가 필요하게 되었고 또 자동통역전화 등 국제적 연결을 위한 응용분야가 개발됨에 따라 자국의 언어가 아닌 다른 나라의 음성에 관한 데이타베이스가 필요하게 되었다. 이에 따라 자연히 필요한 데이타베이스의 규격이나 종류등의 상호 공동 관심사를 논의할 필요가 있게 되었고 이의 논의를 위한 워크샵등의 모임이 형성되게 되었다. 본 고에서는 이러한 모임중의 대표격인 COCOSDA의 활동에 관하여 언급하고 우리나라에서 음성데이타베이스분야에 관련하여 관심을 기울여야할 부분에 대하여 열거해 보았다.

1. 서 론

컴퓨터가 문명의 이기로 인간생활에 중요한 역할을 맡게 됨에 따라 보다 편리한 입출력 수단으로서의 음성의 역할은 점점 증대되고 있다 이를 위한 음성인식, 합성, 분석에 관한 연구도 세계각국에서 진행되어왔으며 컴퓨터 하드웨어기술의 발달로 시스템적인 측면의 문제점들도 하나 하나씩 해결되어 가고 있는 상황에서 이제 알두 기술은 실용화단계에 까지 이르고 있다.
그러나 이러한 음성처리 관련 기술의 발달에 있어서 가장 걸림돌이 되고 있는 것은 바로 어떻게 다양하고도 원하는 특성을 갖는데이타베이스를 확보하는가 하는 문제이다. 다양한 데이타베이스의 확보는 이미 제작된 시스템의 평가와 성능향상에 매우 중요하기 때문이다. 이러한 문제는 이미 우리보다 수십년 앞서서 음성연구를 해오고 있는 음성연구의 선진국이라 할 수 있는 미국, 일본을 비롯하여 유럽 각국의 경우에도 예외는 아니어서 기존의 데이타베이스를 공유한다든지 데이타베이스에 관한 정보를 상호교환하는 체계를 만들어 가고 있고, 나아가서는 다른 나라의 언어에 대한 데이타베이스를 확보하기 위하여 국제적인 교류활동을 행해오고 있다.
이러한 활동의 대표적인 사례가 바로 Eurospeech및 ICSLP 등의 음성관련 국제학술대회와 함께 개최되고 있는 COCOSDA이다. 본 고에서는 94년 일본 요코하마에서 열린 ICSLP'94의 위성 세션으로 개최되었던 COCOSDA의 내용을 중심으로 국제적인 음성데이타베이스 교류동향을 기술하고자 한다. 또한 아시아권의 동향에 관하여도 언급한다.

2. COCOSDA

COCOSDA 는 'Committee for the International Cooperation and Standardization of Speech Database and Speech Input/Output Assessment Methods'의 약자로 문자그대로 음성 데이타베이스, 음성입출력 평가기술의 국제협력과 표준화를 위한 위원회이다.

2.1 COCOSDA관련 일지

관련 모임 일지

.1982 Gaithersburg, USA organized by Dave Palett
.1989 September
Noordwijkerhout, the Netherlands
(ESCA ETRW organized by Louis Pols
1990 11월
일본 고베 ICSLP'90의 satelite event
.1991 9월
이태리 Chaivari Eurospeech '91의 satelite event
***(cocosda의 공격형성)
.1992 7월
Elsnel/ESCA/Salt Workshop on 'Integrating Speech and Natural Language' 더블린
.1992 10월
ICSLP'92의 satelite event 카나다 Banff
.1993 9월
Eurospeech '93의 satelite event, 독일 Berlin
.1994 9월
ICSLP'94의 satelite event, 일본 Yokohama

COCOSDA모임

1994년 요코하마의 모임을 중심으로

이 모임은 다음과 같은 과정으로 진행된다. 우선 기조발표가 사전에 정해진 발표자에 의해 행해진 다음 세부 분과별로 나뉘어 개별 토의를 행하는 방식으로 진행된다.
1994년의 모임에서는 첫 날 예비모임을 가진 뒤 분야별로 나누어 지정된 발표자료를 발표하고 문제점을 토의하는 방식으로 진행되었다. 모임은 크게 데이타베이스에 관한 내용과 인식, 합성시스템의 평가에 관한 내용으로 나누어 진행되었다. 이중 데이타베이스에 관한 내용은 전원이 공

통으로 참가하여 진행하고 인식과 합성에 관한 내용은 각 분야별로 나누어 진행하였다.

1994.9.22 Plenary session
9 23 Individual Working Group meetings
(Corpora and Labelling : Millar* & Itahashi
Recognition : Pallett* & Furui
Synthesis : Pols*
* working group governers)
Cross-group discussions on labelling

연급된 내용들

다음은 이 모임에서 논의된 주제들을 정리한 것이다. 앞으로 국내의 음성 데이타베이스 개발과 보급에 방향을 제시해 줄 것으로 보고 개요를 싣는다

Corpora and Labelling

TED and NEWS
POLYPHONE
PROPHERTY RIGHTS
SPEAKER RECOGNITION DATA CORPUS DESIGN
PROSODIC TRANSCRIPTION
ORIENTAL SPOKEN LANGUAGE DATABASE DESIGN

Physical Quality & Transcription
Multi-Speaker & Multi-Sensor Description
Infra Structure for Labelling Schemes
Labelling Oriental Language Data
Compression Standards

LDC(Linguistic Data Consortium)관련 활동

자동음성인식, 자연언어처리등 대규모 언어데이타가 필요한 분야의 연구,개발을 위한 공통 데이타베이스를 확보하기 위한 컨소시엄이다. 미국 펜실바니아 대학에 본부를 두고 있으며 회원 단체가 원장한 비용을 분담하여 데이타베이스를 개발,사용할 수 있게하고있다.

LDC소장 데이타베이스 목록

- TIMIT and NTIMIT speech corpora
- Resource management speech corpus (RM1, RM2)
- Air travel information system(ATIS0) speech corpus
- Association for Computational Linguistics - Data Collection Initiative text corpus(ACL-DCI)
- TI connected Digits Speech Corpus(TIDIGITS)
- TI 46-word Isolated Word Speech Corpus(TI-46)
- Road Rally conversational speech corpora(including 'Stonehenge' and 'Waterloo' corpora)
- Tipster Information Retrieval Test Collection
- Switchboard speech corpus ('Credit Card' excerpts and portions of the complere switchboard collection)
- Machine readable spoken english speech corpus (MARSEC)
- Edinburgh Map Task speech corpus
- Message understanding conference(MUC) text corpus of FBI terrorist reports
- Continuous Speech Recognition - Wall Street Journal speech corpus(WSJ-CSR)
- Pen Treebank parsed/tagged text corpus
- Multi-site ATIS speech corpus (ATIS2)
- Air Traffic Control(ATC) speech corpus
- Hansard English/French paralel text corpus
- European Corpus Initiative multi-language text corpus (ECI)
- Int'l labor organization/Int'l Trade Union multi-language text corpus (ILO/ITU)
- Machine-readable dictionaries/lexical database (CUMLEX, CELEX)

TED(Translanguage English Database/Terrible English Database)
- Eurospeech'93에서 발표자들로부터 수집
Eurospeech'95에서 시험예정
사원자에 한해서 CDROM형태로 배포하고 사험하게함

NEWS
- 대규모 신문낭독음성 데이타베이스
분장을 구하기 쉽다.
다른 언어에 대해 같은 의미의 내용을 얻을 수 있다.
현재 Corpora. BREF
WSJ(Wall Street Journal)
British English WSJ

Copy Right of Linguistic Resources in Multimedia Society

Texas Instruments Tsukuba R&D center, Ikuo KUDO

음성 데이타베이스와 관련한 저작권 문제에 관하여 논의하였는데 다음과 같은 주제들로 요약된다

1. 수집에 많은 노력이 든다.
2. 배포의 방법? LDC의 경우는 CDROM
3. 데이타베이스의 배포범위? Academic or Commercial?
일본의 경우 ATR, ASJ데이타베이스는 academic use only
4. 왜 음성데이타베이스가 공개되지 않는가?
5. We need wisdom
데이타베이스의 보급에 따라 새로운 데이타베이스를 창출할 수 있으며 결국 그 혜택을 입게된다.

Speaker Recognition Corpus Design문제

Australian National Univ., Bruce Millar

Basic Parameters

Identification and Verification
Text-dependent and Text Independent

Major domain for design decision

1. Speaker Selection and description
2. Material selection and description
3. Environment selection and description
4. Speaking task selection and description
5. Signal transforms and description

유럽의 음성데이타베이스 교류

ESCA(European Speech Communication Association)를 중심으로 형성된 공동연구활동에서 각종 데이타베이스가 제작되고 보급되고 있다

Recent developments in Europe
on Spoken/Written Language Resources

LIMSI-CNRS(France), J. Mariani

RELATOR(European repository of Linguistic resources)

-Within the CEU language research & engineering(LRE) program
1994년 1월 시작
Univ. of PISA(Coordinator, Italy), Center for Cognitive Science(Edinburgh, UK), Univ. of Stuttgart(Germany), DFKI(Saarbrucken, Germany), LIMSI-CNRS(France), ICP(France), Institute de Droit(France), INESC(Portugal), Univ. of Copenhagen(Denmark)

Goal: Provide a European infrastructure for the distribution of Spoken and Written language resources.

EuroCocosda
목적:
- to establish the infrastructure for a concrete European contribution to Cocosda
- to provide support for the European component of Polyphone
- to coordinate the European component of an NL-Speech Cocosda initiative, NEWS, a multi-language newspaper text and speech database
- to produce a non-native English corpus, TED, containing spontaneous and read speech and complementary text materials.
- to foster links with other European work, for example: EAGLES, SQALE & RELATOR within LRE and projects in ESPRIT III
#SQALE: Speech Recognition system quality Assessment
#RELATOR: Repository of Speech & Language Resources
#EuroCocosda: European interface to Cocosda

PARTNERS:
UCL, UK
LIMSI-CNRS, France
University of Amsterdam, Netherlands
CSELT, Italy
University of Munich, Germany

3. 아시아권의 음성DB관련 활동

아시아권 관련모임의 구성현황

(1) 일본의 데이타베이스 관련 활동
ATR, ETL, ASJ, JEIDA등의 기관과 대학에서 개별적으로 데이타베이스 작성작업을 하고 있으며 국가 프로젝트로 작성작업이 지원되고 있다
(2) 중국의 데이타베이스 관련 활동
1992년 시작된 국가 프로젝트의 일부로 Chinese Speech Corpus가 작성되어 있다.

Oriental COCODA의 발족
1994년 일본 요코하마에서 개최된 ICSLP'94의 COCOSDA에서 Shuichi Itahashi교수에 의해 동양권의 음성처리에 필요한 문제점을 논의하기 위한 목적으로 동양권의 별도 모임을 만들것을 주창하여 가칭 Oriental COCOSDA가 구성되었다. 이때 논의된 것은 다음과 같다.

Itahashi교수의 발기문

1. 지역적인 문제들을 지역의 노력에 의해 해결해야 한다.
2. 서구 국가들의 동양권 언어에 관한 관심이 증대되고 있다.
3. 동양권 언어들은 서구 언어과는 다른 독특한 성격이 있다.
a) 다양한 성질을 띠며 다른 언어권에 속한다.
b) 중국, 한국, 일본등과 같이 다른 표기시스템이 있다
c) 다양한 로마자 표기법이 있다
d) 문자표기법과 음성학적, 음운론적 표기의 상관관계가 확실하지 않은 부분이 많다.
4. 동양권은 지역적으로 인접해 있다.

COCOSDA에서 동양권의 COCOSDA설치에 관한 발표가 있은 직후 한, 중, 일 3국의 관련 참석자가 모여 이의 설치에 동의하고 추후 연락을 통해 실무적인 문제들을 논의하기로 하였고 이후의 연락은 일본의 Itahashi교수가 맡기로 하였다

이 모임에서 Oriental COCOSDA구성과 관련하여 거론된 관련 단체명은 다음과 같다. [5]

1) 중국
Chinese COCOSDA
Prof. Jialu Zhang and ? members
2) 한국
KCCSLP: Korean Coodinating Committee for Spoken Language Processing
Prof. Sougil Ann and ? members
3) Speech Database Committee, Acoustical Society of Japan
Prof S. Itahashi and 29 members
4) Speech Input/Output Systems Committe, JEIDA
Prof. S. Itahashi and 19 members

JEIDA: Japan Electronic Industry Development Association

5) LRSI: Linguistic Resources Sharing Initiative
Dr. T.Yokoi and 24 members

6) Database Workshop of RWCP(Real World Computing Partnership) is interested in creating linguistic corpora and image database
Prof. S.Hahashi and 13 members

7) ATR Interpreting Telecommunications Research Laboratories
Dr. Yamazaki and members

9) Grant-in-Aid for Scientific Research on Priority Areas Project "Spoken Dialog"
Prof. H Fujisaki and 14 members

Oriental COCOSDA와 관련한 문제점들

이 모임은 아직 초기 단계로 발기후의 구체적인 활동은 보고되지 않고 있으나 초창기부터 일본의 주도로 세워졌고 기타 국가들의 참여가 극히 미미하므로 앞으로 적극적인 참여가 요구되며 우리의 말과 글의 특성이 국제 사회에서 앞으로 논의될 데이타베이스의 규격이나 음소표기법, 로마자화 표기법등에 정확히 반영되지한 필요성이 있다. 이와 관련하여 주목할만한 점은 이번 Oriental COCOSDA의 출발에 내비하여 중국이 사전에 자국의 활동상황을 소개하는 준비를 하여 발표를 할 기회를 가졌던 것과는 대조적으로 우리 나라에서는 그러한 소개의 기회조차 갖지 못했다는 것이다. 이는 음성데이타베이스와 아의 국제적 교류 및 동향에 대한 국내의 인식부족, 활동의 미비함을 원인으로 꼽지 않을 수 없다. 이 문제는 최근까지 문제가 되어왔던 컴퓨터 코드의 표준화에 못지 않은 중요성을 갖고 있다고 보인다. 또한 이러한 모임에 정기적으로 한 개 이상의 단체에서 대표를 파견하여 우리의 입장을 알릴 필요가 있다. 이러한 활동에 있어서는 음성분야의 주요관련 연구소, 학회등에서의 관련된 문제들에 관한 활발한 논의와 함께 적극적인 참여를 위한 노력이 필요하다. 이러한 문제는 개인의 차원이 아닌 국가적 차원에서 보아야하며 이점에서 우리나라 음성연구를 주도하고 있는 연구소나 학회 수준에서의 참여가 지속적으로이어져야 할 것이 바람직하다고 보인다. 또 이들 모임에 참가하기 위해서 사전에 국내의 활동에 대한 현재까지의 상황을 파악하고 의견을 모아 정리할 필요가 있다고 생각된다.
다음은 이러한 목표를 위해 앞으로 국내의 학술단체, 학교, 연구소등에서 취해야할 방향을 나름대로 정리해 본 것이다.

1. 음성데이타베이스 및 레이블링등과 관련한 국내 논의의 활성화
2. 기존 국내 음성데이타베이스의 정리, 공유화
3. 국제 음성데이타베이스 관련 모임에의 적극적 참여
4. 관련 연구소, 학술단체의 음성데이타베이스에 대한 인식제고
5. 우리말 음성의 풍부한 로마자 표기법 설정
   - 레이블링, 데이타베이스의 제작등에 활용
6. 우리말 데이타베이스의 표준형식 설정
7. 음성인식, 합성 시스템에 관한 평가방법 마련 및 평가용 데이타베이스의 구축

등이 앞으로 국내에서 시급히 해결되어야할 문제들로 생각된다.

4. 결론

이상에서는 음성데이타베이스교류를 위한 국제적 활동에 관하여 간단히 알아보았다. 국제화, 세계화 시대를 맞아 이제는 우리 말에 관한 연구도 더이상 국내의 수준에만 안주할 수 없게 되어 있다. 이 대표적인 사례의 하나로 이미 한글 코드의 표준화안이 국제적인 논의의 대상이 되어 왔고 이제 우리 말의 표기법과 음성 데이타베이스까지도 예외가 될 수 없는 상황에 이르렀다. 그러나 아직도 음성데이타베이스와 관련 분야에 관한 국내의 사정은 열악한 상태이다.
이 문제를 해결하는 방법은 국내의 관련 분야의 학계, 연구소, 기업들에서 활발한 논의을 진행시켜나가는 수 밖에는 없을 것이다.

5. 참고문헌

1) Report on the COCOSDA workshop. Pacific Convention Plaza. Yokohama, Japan. 1994
2) Report on the COCOSDA Workshop. Haus am Kollnischen Park. Berlin, Germany. 1993
3) Frequently Asked Questions, comp speech, usenet
4) Materials on COCOSDA workshop, Yokohama, Japan. 1994
5) Email on Oriental COCOSDA from Hahashi Shuichi