Speech Resources for Scientific Research and its Application

Kimiko YAMAKAWA Shuichi ITAHASHI (NII)

Objective?

NII has established Speech Resources Consortium (NII-SRC) so as to promote dissemination and distribution of speech resources.

NII-SRC conducts collection, distribution, investigation and research on speech resources (including speech data and software tools) necessary for developing science, education and industry related to speech.

What are we doing?

We contribute to the development of various research including speech recognition and synthesis by collecting and distributing speech corpora or speech databases which are difficult to develop individually.

Another scientific contribution by supplying valuable material for phonetics and sociolinguistics by preserving dialects and minority languages.

1. What is “Speech Corpus/Corpora”

What is corpus/corpora?

A corpus means a systematic collection of data for research with some additional information to be used for research.

(Ex.) Speech corpus, text corpus, multimedia corpus, image corpus, etc.

Variety and use of speech corpora

【Use】Analysis, synthesis, recognition of speech; analysis of discourse and dialects; preservation of languages, etc.

【Variety】Isolated words, continuous speech, read speech, dialogues, dialects, multilingual speech; speech by non-native speakers, infants, aged people; speech in noisy or reverberant environments.

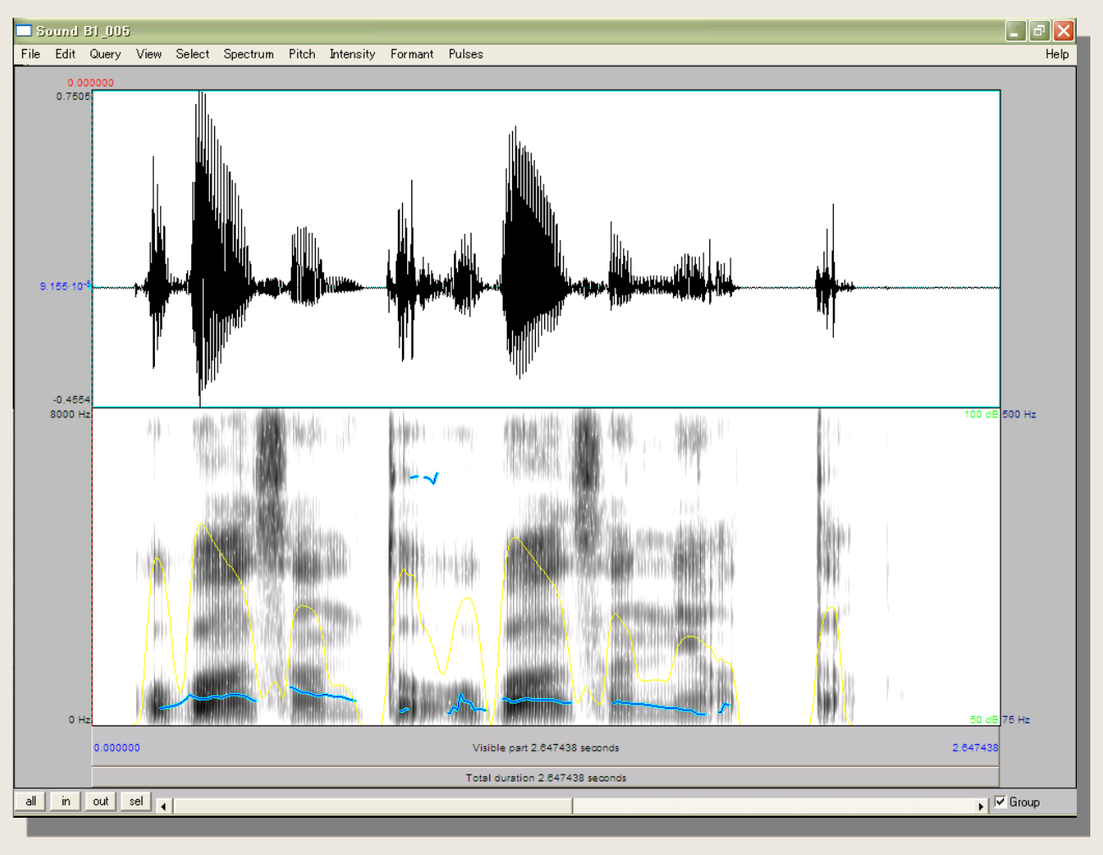
Recording media of speech corpora

The major recording media are used though it varies according to the use or data size. DVD-R is the most common currently. On-line distribution will be available soon.

【Recording media for speech corpora】

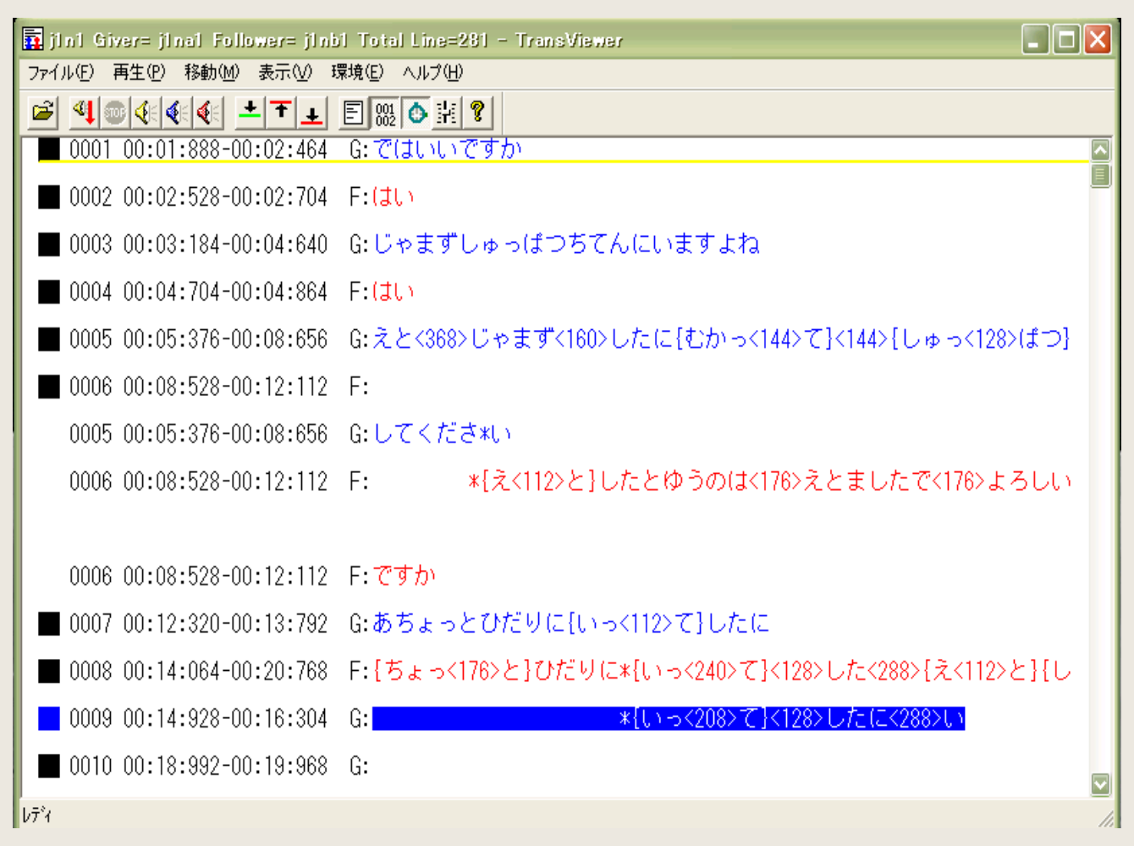
CD-R, DVD-R, HDD, DAT, LD, etc.

Contents of speech corpora



Analysis data

Video



```
j1n1 Giver= j1na1 Follower= j1nb1 Total Line=281 - TransViewer
ファイル(F) 再生(P) 移動(M) 表示(V) 環境(E) ヘルプ(H)
■ 0001 00:01:888-00:02:464 G:ではいいですか
■ 0002 00:02:528-00:02:704 F:はい
■ 0003 00:03:184-00:04:640 G:じゃまずしゅっぱつちてんにいますよね
■ 0004 00:04:704-00:04:864 F:はい
■ 0005 00:05:376-00:08:656 G:えと<368>じゃまず<160>したに[むかっ<144>て]<144>[しゅっ<128>ぱつ]
■ 0006 00:08:528-00:12:112 F:
  0005 00:05:376-00:08:656 G:してくださ*い
  0006 00:08:528-00:12:112 F:        *[え<112>と]したとゆうのは<176>えとましたで<176>よろしい

  0006 00:08:528-00:12:112 F:ですか
■ 0007 00:12:320-00:13:792 G:あちょっとひだりに[いっ<112>て]したに
■ 0008 00:14:064-00:20:768 F:[ちょっ<176>と]ひだりに*[いっ<240>て]<128>した<288>[え<112>と][し
■ 0009 00:14:928-00:16:304 G:                    *[いっ<208>て]<128>したに<288>い
■ 0010 00:18:992-00:19:968 G:
レディ
```

Transcription data

2. What is SRC?

Why speech corpora, now?

- Development of speech processing technology
- Quantitative research in linguistics-related areas
- Importance of preserving languages and dialects

Massive speech data of various kinds necessary

Problems of speech corpora

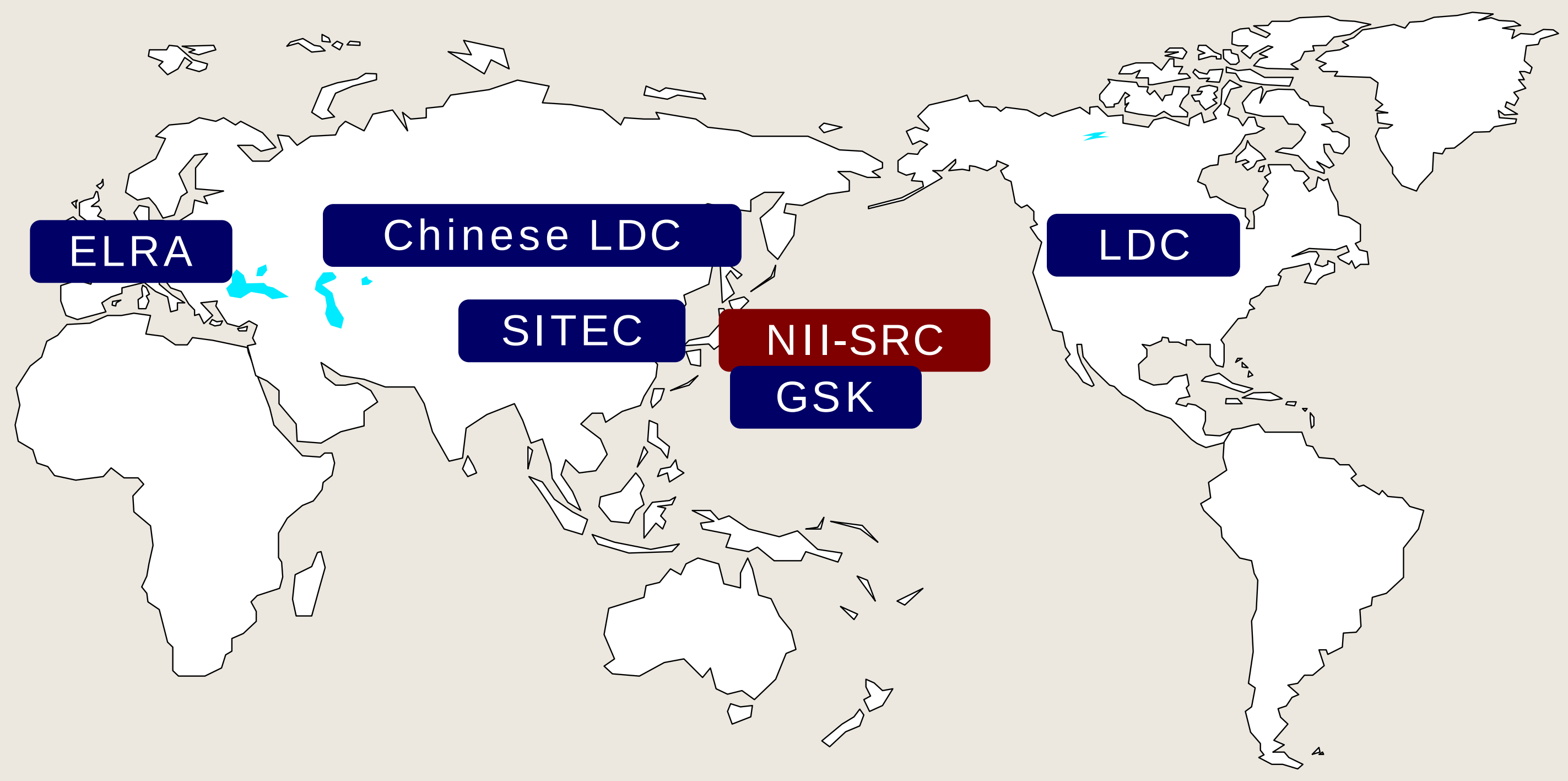
- Most corpora are developed for a project.
- It requires cost, time and labor to create.
- Expensive.
- Not open to the public.

A common framework is required for creation, collection, accumulation, distribution and sharing

Speech Resources Consortium (NII-SRC)

Launched by NII in 2006 in order to collect, manage, and distribute various speech corpora.

Currently, 31 corpora are available from NII-SRC.



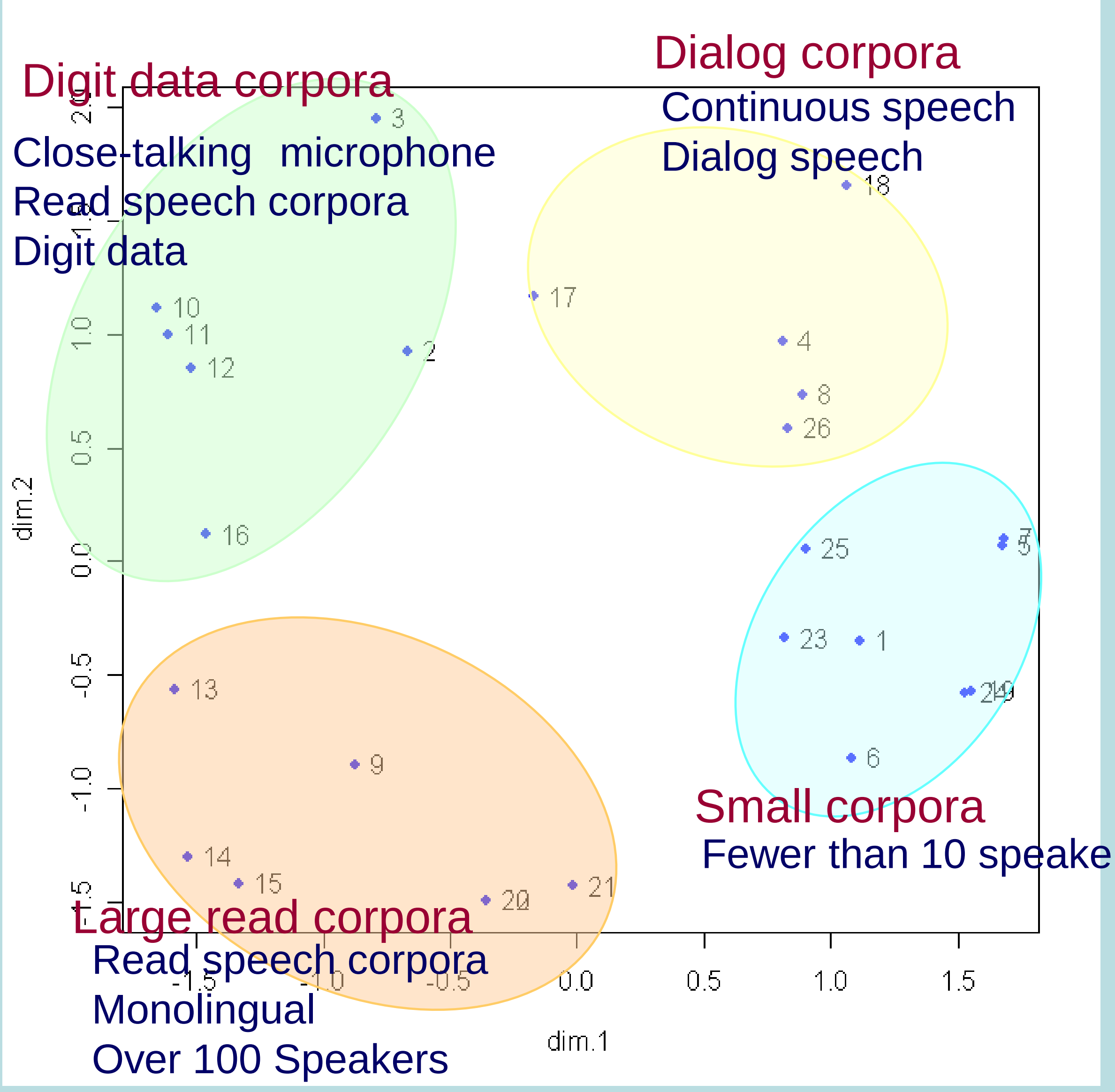
Speech-related Organizations in the World.

3. Categorization of speech corpora

Corpus attributes (8 attributes and 58 items)

Attribute	Item	
Input device	7 items	Type of input device (ex. Desk-top microphone)
Input environment	5 items	Recording environment (ex. Soundproof room)
Number of speakers	10 items	Number of speakers
Speaking style	4 items	Style of speech (ex. Continuous speech)
Speech mode	5 items	Speech mode (ex. dialog, read speech)
Data mode	9 items	Other information (ex. Sampling frequency)
Language	4 items	Type of language (ex. Monolingual)
Purpose	14 items	Keyword for use or development (ex. Recognition)

4. Corpus similarity visualization



Speech Resources Consortium, National Institute of Informatics
URI： http://research.nii.ac.jp/src/eng E-mail： src@nii.ac.jp