Spoken Dialogue System Framework Based on User-Generated Content

Keiichi Tokuda, Akinobu Lee,
Daisuke Yamamoto (Nagoya Institute of Technology),
Junichi Yamagishi (University of Edinburgh)

uDialogue Project

User-Generated Dialogue Systems: uDialogue

Period:

– Oct. 2011 - March 2017

Participants:

– Speech Processing Lab, NITech (Tokuda Group)
– IT Center, NITech (Yamamoto Group)
– CSTR, University of Edinburgh (Renals Group)

Background (1/2)

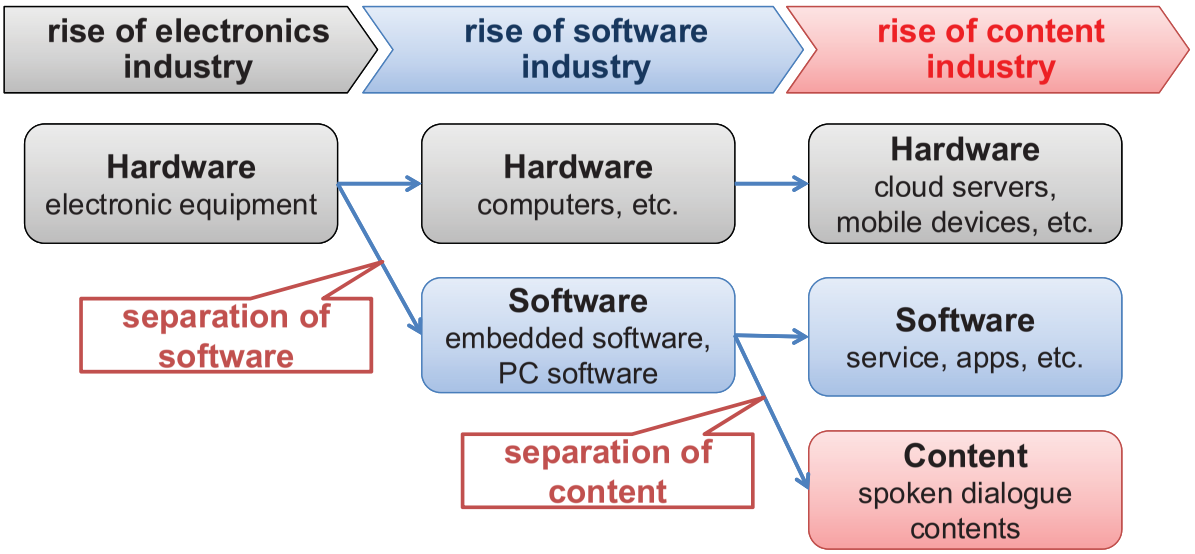
- **Speech Interface**
  - speech is the most basic form of communication

- **During past several years practical use advances**
  - ex) Google voice search, voice translation
  - ex) speech reservation systems (British Airways (UK), Amtrak (US), etc.)

General users want to use spoken dialogue systems?

Background (2/2)

Industrial structure in telecommunications



- **Separation of content**
  - Users can easily participate content creation
  - Generation of attractive spoken dialogue contents
    - voice, facial expression, gesture, timing, humor, etc.

Examples of User Generated Content

Printed matter
books, news papers, dictionaries

WIKIPEDIA
separate contents from dictionaries

Movie and Music
CD, video, DVD, broadcasting

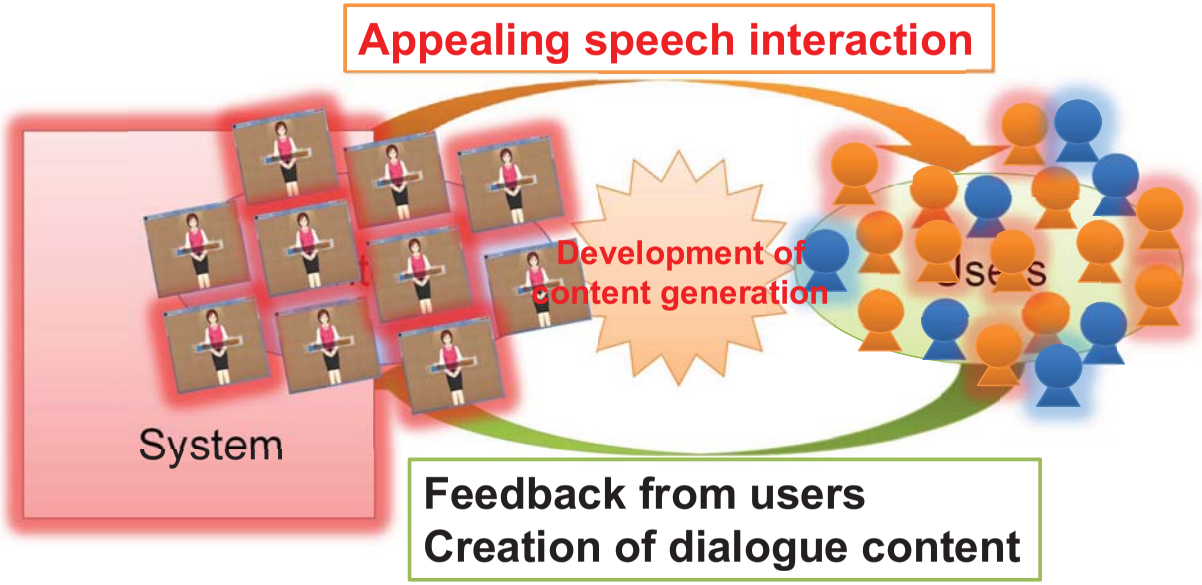
YouTube USTREAM
separate contents from media

Spoken dialogue systems

JST CREST uDialogue Project
separate spoken dialogue content from the systems
→ **attractive interaction**

Concept of the project

- **Investigation of mechanisms for generating attractive spoken dialogue content**



Strategy for accelerating content generation

- **Open source (software)**
  - not only source code but also file format etc.
  - easy to extend
- **Digital signage (hardware)**
  - wide-spreading
  - time and place can be specified
  - university campus would be a test bed
- **Internet (network)**
  - participation of more users
  - incentive to content creation
  - discovery of excellent creators

Digital signage in the campus

NITech speech processing lab.

Keiichi Tokuda
Director
Speech synthesis
Speech processing

Akinobu Lee
Speech recognition
Spoken dialog system
Voice interaction

Yoshihiko Nankaku
Speech recognition
Statistical modeling
Image recognition

Keiichiro Oura
Speech synthesis
Singing voice synthesis

Kei Hashimoto
Speech recognition
Acoustic modeling

+ 2 Post-docs
+ 4 Ph.D. students

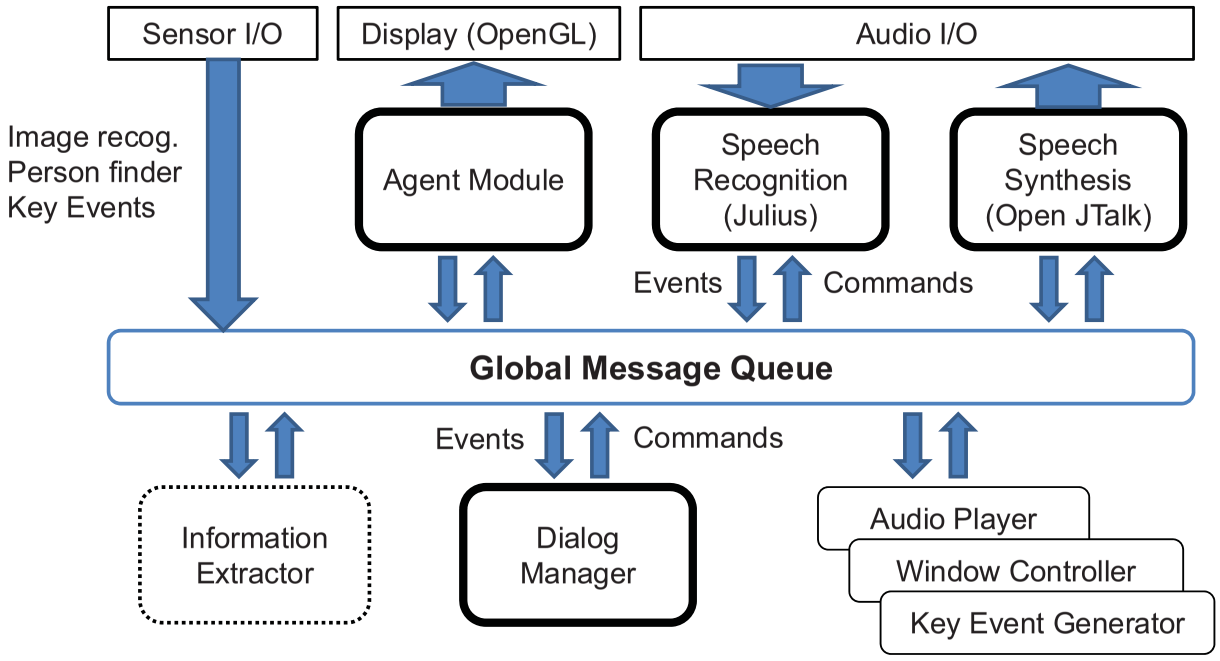
MMDAgent

~ Toolkit for building voice interaction systems ~

- **Modern speech technologies**
  - High-quality, low-latency speech recognition and synthesis
- **Life-like interaction with 3D objects**
  - Rich model expression (bones, morphs, IKs, physics…)
  - Detailed motion management (online motion composition)
- **Open source, open format**
  - All components can be made by freely-available tools
    - Acoustic model, language model, 3D object, motion, dialog…
  - All-in-one package, runs on Win/Mac/Linux + OpenGL

… to promote exploration of speech interaction to everyone

System overview



Speech recognition module

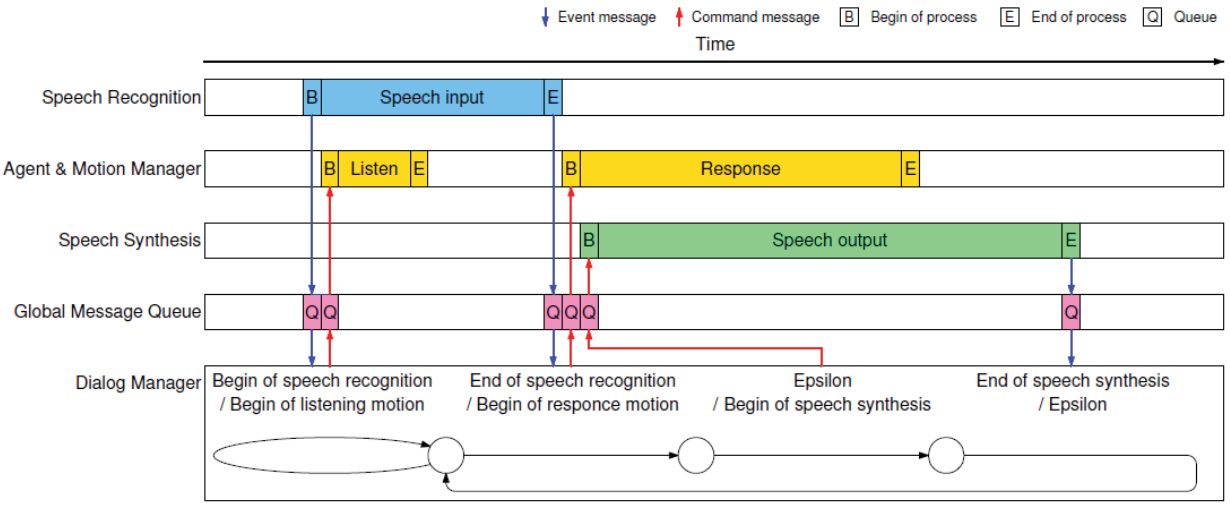
- **"Julius" LV speech recognition engine**
  - Open source, open format
  - Small footprint, rapid and low-latency recognition
  - http://julius.sourceforge.jp/
- **Built as a plugin for MMDAgent**
  - All functions are available in the original version
    - GMM-based input rejection, multi-model recognition etc…
  - Support task-dependent additional dictionary
    - Can change keywords for each task at run time
- **Default system for Japanese**
  - 60k-word general Japanese 3-gram trained from Web
  - SI triphone HMM

Speech synthesis module

- **Diverse voice expression for dialog system**
  - Various speech styles should be controlled
- **HMM-based TTS**
  - Statistical parametric method
    - Small footprint (voice parameters are stored as statistics)
    - Less distortion than concatenative synthesis
    - Easy to control the output speech by parameter conversion
- **HMM Training tool**
  - HTS: HMM-based speech synthesis system
    http://hts.sp.nitech.ac.jp/
- **Synthesis plugin for MMDAgent**
  - Open JTalk: a Japanese TTS system
    http://open-jtalk.sourceforge.net/

Dialog management with events

- **Messages in the global queue**
  - Events from the modules: internal status changes
  - Command to the modules: actions to be issued
- **Allows asynchronous processing**



Demonstration

0. System Introduction
1. “Hello world”
2. Expressive speech synthesis
3. Barge-in
4. Inverse kinematics and physics simulation
5. On-line motion composition

Information Technology Center Group

(Yamamoto Group)

Members

Ichi Takumi
Professor

Computer system network | Natural disaster science | Measurement engineering | Intelligent informatics

Group leader

Daisuke Yamamoto
Associate Prof.

Web service | Multi-media | GIS

Takahiro Uchiya
Associate Prof.

Computer system network | Intelligent informatics

Ryota Nishimura
Assistant Prof.

Spoken dialogue

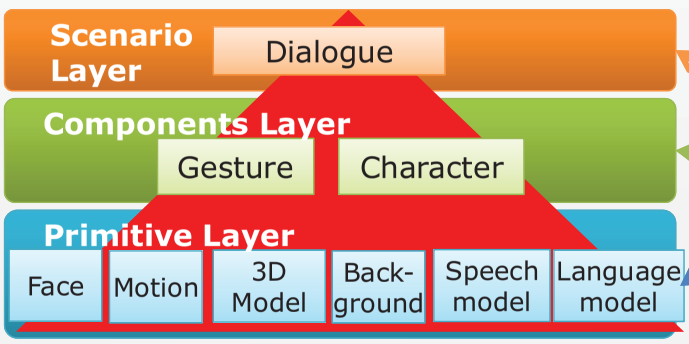
We are belong to
Nagoya Institute of Technology

Task: Environment for user-based spoken dialog content generation

- **Editing contents on the Web easily**
  - **Divide, Share** and **Reuse** the layers of dialog content
  - Automatic contents generation based on **campus DB**.
- **Mobile Services for spoken dialog**

Campus DB

Syllabus | Calendar | Individual data | Bulletin board



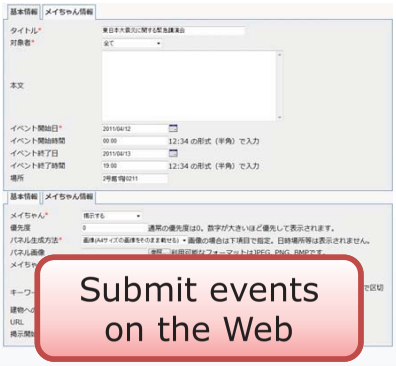
Share

Web Editing

Generate contents

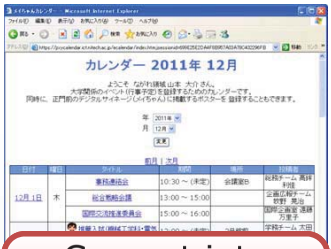
Students, Staffs

Event Calendar with Mei-chan



Submit events on the Web

Title
Body
Date
Keyword



Convert into spoken dialog content

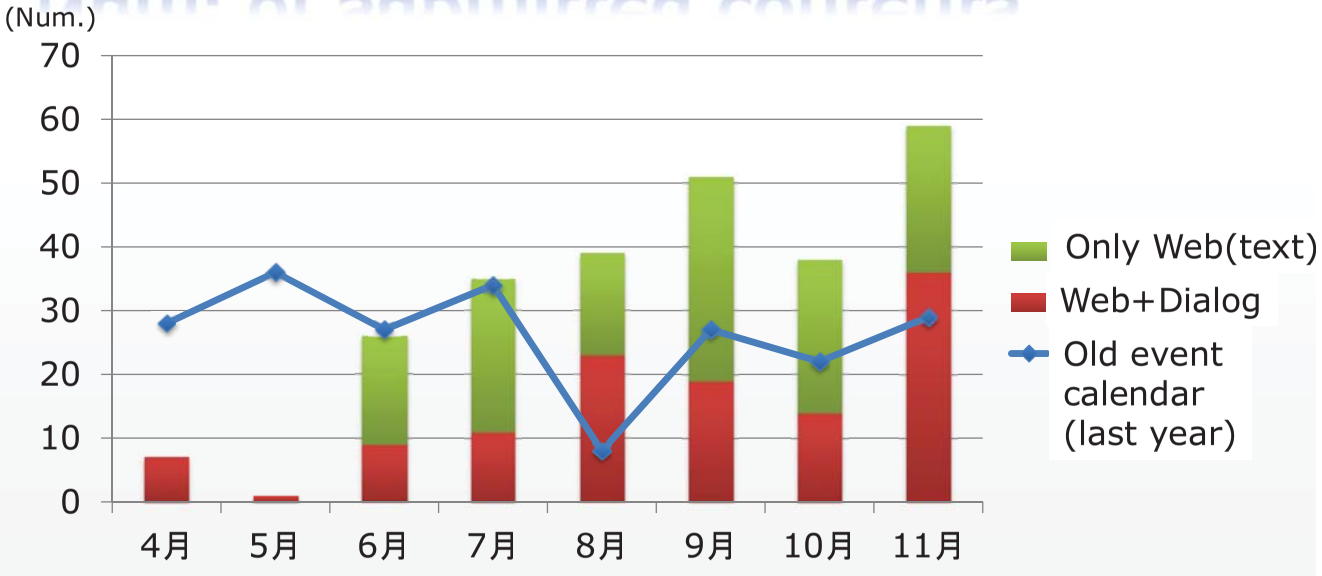
Image
Scripts
Speech

Content on the digital signage

- **Event Calendar**
  - Existing calendar sharing service for NITech
  - Students and staffs post and share events a lot every day.
- **Event Calendar with Mei-chan**
  - Convert events into spoken dialog content for Mei-chan
  - Available since June 2011 at NITech.

demo

Num. of submitted contents



Num. of contents is **2x** bigger than last years.
Including students' original contents.

Mobile Mei-chan

Internet

skype

MMDAgent

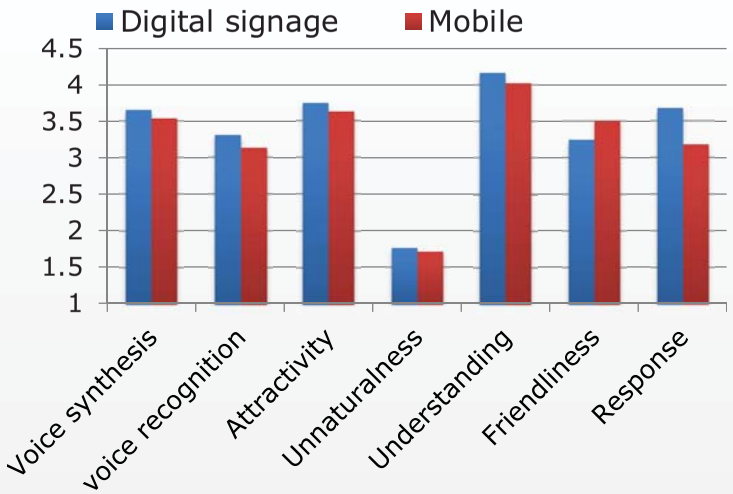
- **A Mobile Spoken Dialog Service**
  - Based on Video phone function of Skype and MMDAgent
  - Any users can communicate with Mei-chan by using their smart phones
    - iPhone, Android, mobile PCs

Open experiment (March 2012)

- **Guidance system based on Mobile Mei-chan**
  - For the 74th National Convention of IPSJ
  - 2836 participants
  - 4 iPhones, 10 servers

demo

- **Questionnaires**
  - 121 answers



Although Response is not better, Friendliness is better than signage.

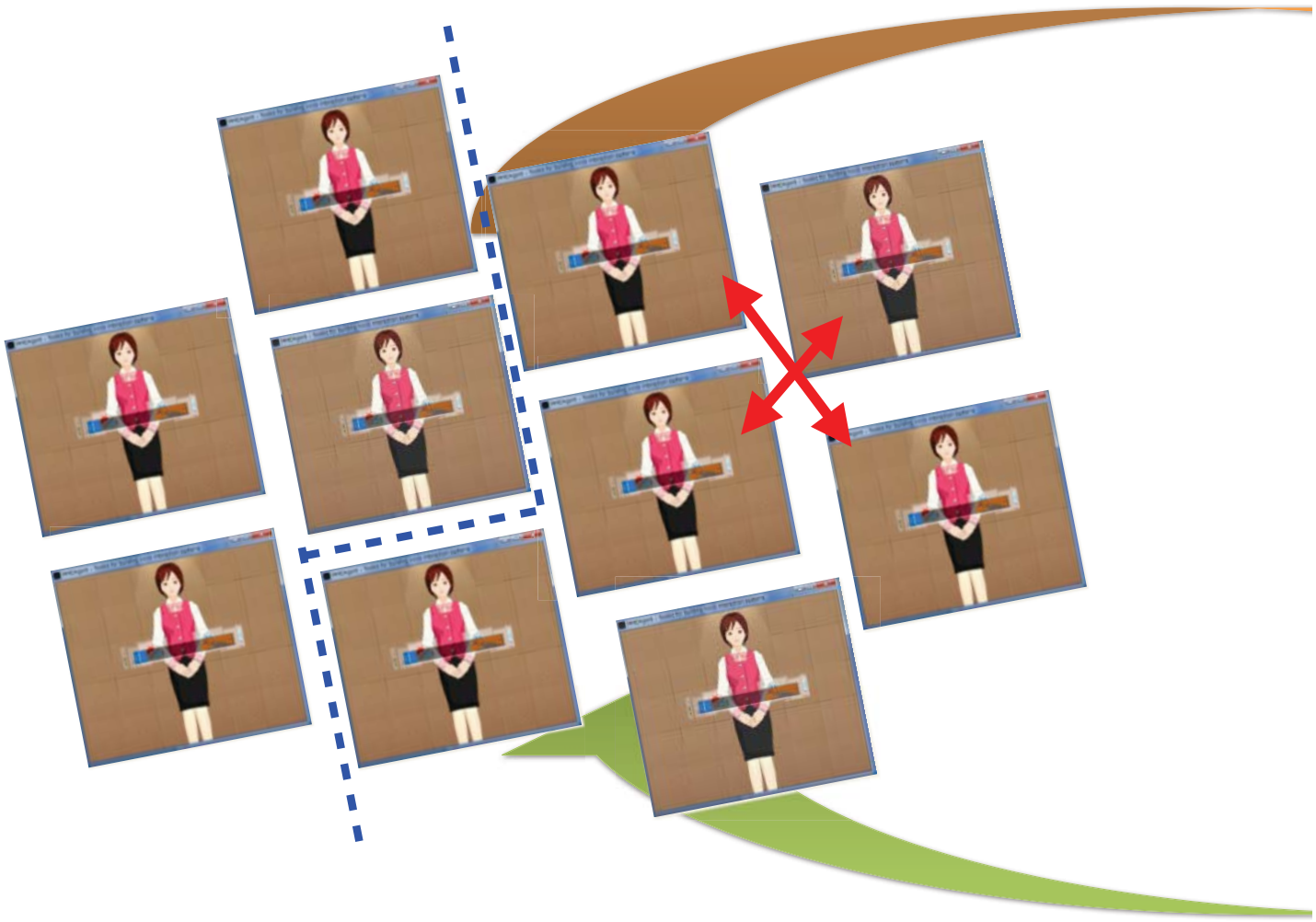




+ 3PhD students

JST CREST uDialogue Project

Crowd Source Content = Knowledge of Spoken Dialogue System



Summary

- **User generated dialogue systems**
  - investigation of the mechanism for generating attractive spoken dialogue content
  - develop basic speech technologies for enhancing the attractiveness
  - extending to statistical approach to spoken dialogue modeling