CUHK – PKU – UST
Joint Research Centre for Language and Human Complexity
語言與人類複雜系統聯合研究中心

Spoken Language Technologies and Applications

Tan Lee 李丹

DSP and Speech Technology Laboratory
Department of Electronic Engineering
http://dsp.ee.cuhk.edu.hk

April 2014

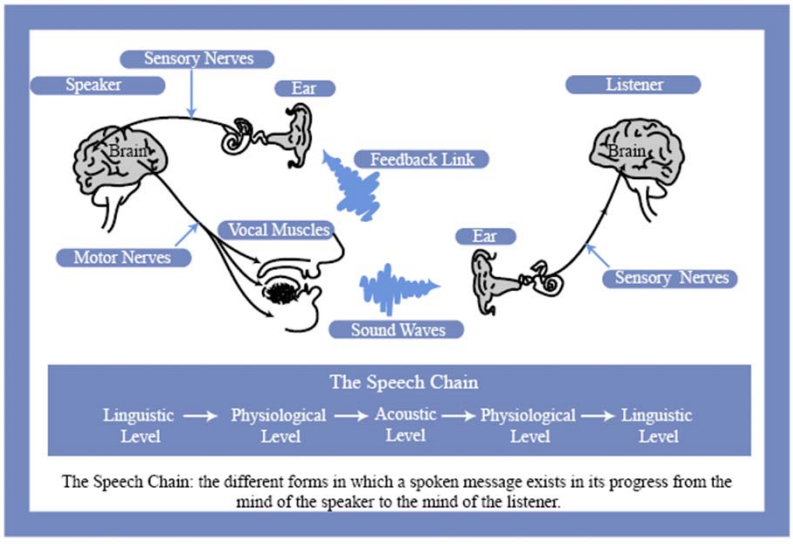
Outline

- Fundamentals of computer speech processing
- Spoken language technologies
- State of the art and limitations
- Applications in speech, hearing and language

Speech communication

- Speaker → listener
- Face-to-face
- Telephone/Internet
- Spoken language
- Not only linguistic information

The Speech Chain



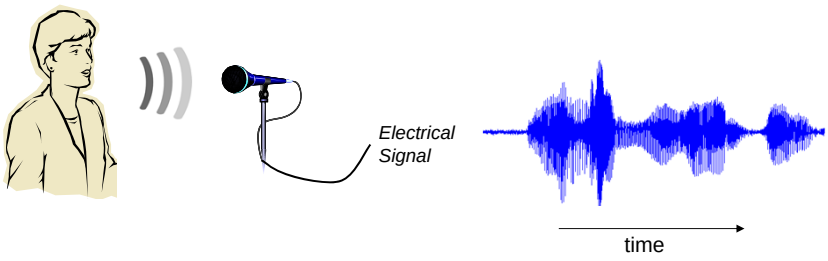
The Speech Chain: the different forms in which a spoken message exists in its progress from the mind of the speaker to the mind of the listener.

(Denes & Pinson, 1963)

Engineering perspectives

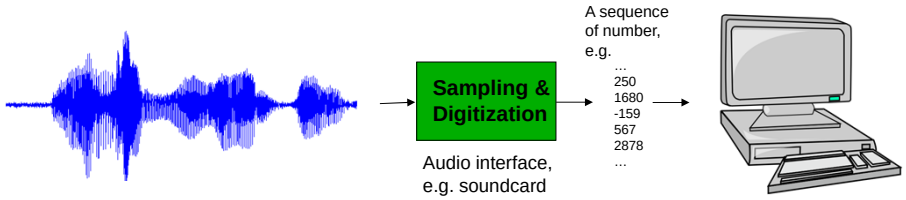
- Speech is a kind of signal transmitted acoustically and electrically
- Variability vs. invariability
- Can we capture, analyze, modify, and regenerate speech signals ?
- Computational models of speech communication
- What for ? What are the applications ?
- Technologies have limitations

Capturing speech



Captured signals come with ***background noise, competing talkers, echo/reverberation, microphone distortion, …***

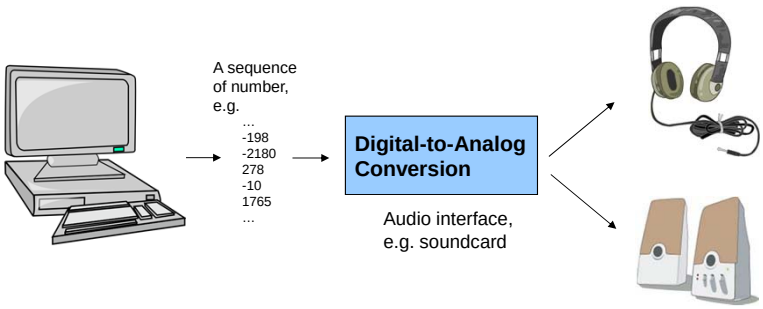
Digitizing speech



Analyzing speech

- Extracting (computing) signal features
- Spectrogram, pitch, duration, intensity
- Establishing linguistic features
- Phonetic unit, stress, intonation, break
- Detection, recognition/identification, classification, verification
- Statistical modeling with large corpus

(Re-)Generation of speech



Speech technologies

- Speech recognition
- Speaker recognition
- Spoken language recognition
- Speech synthesis
- Speech compression
- Speech enhancement

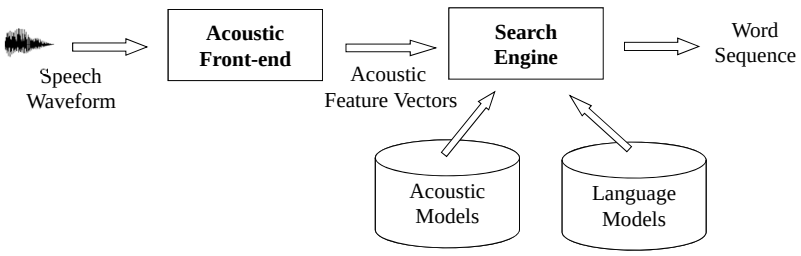
Automatic speech recognition (ASR)

- A process to map speech signals to words
- "recognize a word" ≠ "understanding the speech"

Complete system: LVCSR

Basic principle:

To find the most probable word sequence based on *signal similarity* under *language constraints*



Applications of ASR

- Voice command & control
- Spoken dialogue system
- Automated call-centre service
- Voice dictation: speech-to-text
- Automatic transcription of audio recordings
- …

State of the art and limitations

- Excellent performance achievable in matched conditions
- Server-based systems better than stand-alone: searching in a larger space
- Robustness to environment change needs to be improved
- Inability of handling new words, new domains, new speaking styles, …

Text-to-speech (TTS)

Basic steps:

- Text normalization: from input text to speech symbols
- Prosody specification
- Speech generation

Speech generation

- Concatenation of pre-recorded segments
- Speech modification by signal processing
- Change of spectral features
- Change of intensity, pitch, duration
- Change of "voice"

Applications of speech generation

- Public information system
- Spoken dialogue system
- eBook, eNewspapers
- Voice morphing
- Expressive speech
- …

State of the art and limitations

- More mature and usable than ASR
- Excellent speech intelligibility and quality
- Very good naturalness and intonation
- Text normalization never perfect
- Trainable system: learn to produce “new” voice

State of the art and limitations

- More mature and usable than ASR
- Excellent speech intelligibility and quality
- Very good naturalness and intonation
- Text normalization never perfect
- Trainable system: learn to produce “new” voice

Demonstration

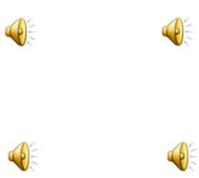
Speech Recognition Breakthrough for the Spoken, Translated Word

Speech enhancement

- To reduce background noise
- To suppress voice of competing speaker
- To compensate environmental reverberation
- To improve perceptual quality of speech
- To improve intelligibility of speech

State of the art and limitations

- Stationary noise can be effectively suppressed
- Multi-talker babble is challenging
- Competing speaker is a nightmare



Other related technologies

- Speaker authentication: “voice print”
- Spoken language identification
- Spoken term detection
- Language proficiency assessment
- Singing voice synthesis
- …

Technology trends

- Robust
- Adaptive
- Multi-lingual
- Multi-modal: with vision, gesture, haptics
- Expressive
- Customized
- Personalized

DSP & Speech Technology Lab.

Department of Electronic Engineering, CUHK
Room 329, Ho Sin Hang Engineering Building

P.C. Ching 程伯中
Chair Professor
Vice-President, CUHK

William S.-Y. Wang 王士元
Research Professor
CLHC Director

Tan Lee 李丹
Associate Professor
Associate Dean of Engineering

Frank K. Soong 宋國平
Adjunct Professor
Speech Group Manager
Microsoft Research Asia

Speech Research @ EE, CUHK

- Since 1987
- Early research focused on Cantonese ASR & TTS
- Lot of work on lexical tones and prosody
- Well known as resources centre for Cantonese speech processing
  - Hundreds of hours of speech recording, covering microphone and telephone speech, continuous speech and isolated words, Cantonese-English code-mixing speech
  - Cantonese pronunciation lexicon
  - In-house developed ASR and TTS software
  - available for research (write to me if you are interested)

Features of our research

- Started by developing core component technologies and infrastructure
- Focused on Cantonese
- Emphasizing on the use of linguistic knowledge
- Not only for human-computer interaction, but also for human-human speech communication
- Inter-disciplinary studies and collaboration
- Providing engineering tools for research in language and speech

Selected thesis titles (DSP Lab, EE, CUHK)

Automatic Recognition of Isolated Cantonese Syllables Using Neural Networks	PhD 1996
Inverse Solution of Speech Production Based on Perturbation Theory and Its Applications to Articulatory Speech Synthesis	PhD 1998
Large Vocabulary Continuous Speech Recognition for Cantonese	MPhil 2000
Cantonese Text-to-Speech Synthesis Using Sub-syllable Units	MPhil 2001
Information Fusion for Monolingual and Cross-Language Spoken Document Retrieval	PhD 2002
On the Robustness of Static and Dynamic Spectral Information for Speech Recognition in Noise	PhD 2004
Use of Tone Information in Cantonese LVCSR Based on Generalized Character Posterior Probability Decoding	PhD 2005
Speaker Recognition Using Complementary Information from Vocal Source and Vocal Tract	PhD 2005
Short-time Independent Component Analysis for Blind Separation of Speech Sources	PhD 2007

Model-based Speech Separation and Enhancement with Single-microphone Input	PhD 2008
Exploitation of Effective Temporal Cues for Lexical Tone Recognition of Chinese	PhD 2009
Spoken Language Identification with Prosodic Features	PhD 2010
A Perceptual Study on Linearly Approximated F0 Contours in Cantonese Speech	PhD 2010
Development of a Cantonese-English Code-mixing Speech Recognition System	PhD 2011
Design and Evaluation of Tone-enhanced Strategy for Cochlear Implants in Noisy Environment	MPhil 2011
Speech Periodicity Enhancement Based on Transform-domain Signal Decomposition and Robust Pitch Estimation	PhD 2011
Model-based Single-microphone Speech Separation Using Conditional Random Fields	PhD 2014

Inter-disciplinary research
Speech and Hearing Disorders

- Speech processing technologies are useful in assessment and rehabilitation of speech and hearing disorders
- Hearing aids and cochlear implants
- Pitch-controlled alaryngeal speech production
- Assessment of disordered voice and speech

Speech enhancement for better hearing

- Hearing impairment is a common disorder
- Various degrees: mild, moderate, severe, profound
- Conductive vs. sensori-neural
- Hearing aids (HA): sound amplification
- Cochlear implants (CI)
  - Surgically implanted
  - Electrical stimuli delivered via implanted electrodes

Cochlear Implant

- A surgically implanted electronic device
- For severe-to-profound sensori-neural hearing loss
- To compensate the impaired acoustic-to-electrical conversion
- The device takes in acoustic signal and delivers electrical stimuli at cochlea

Figure from Harvey Dillon, ***Hearing Aids***

Cochlear implants

Problems of existing CI devices

- Hearing in noise is difficult and uncomfortable
- Pitch cues are not effectively represented
- Significant performance difference in tone identification abilities of normal-hearing and CI listeners

Our work

- Collaborative research with medical doctors and audiologists
- New signal processing method to make pitch cues more prominent
- New methods to suppress background noise
- Subjective listening tests
  - Tone recognition
  - Speech recognition
  - In quiet and in noise

"Listening" tests

Test interface

Voice disorders

- 7% of general population have voice problems
- abnormality of pitch, volume, resonance and/or quality, and/or a voice”
- irregular masses on vocal systems or inefficient use of laryngeal musculature
- Intensive and inappropriate use of voice
- School teachers are highly risky

Perceptual evaluation of voice

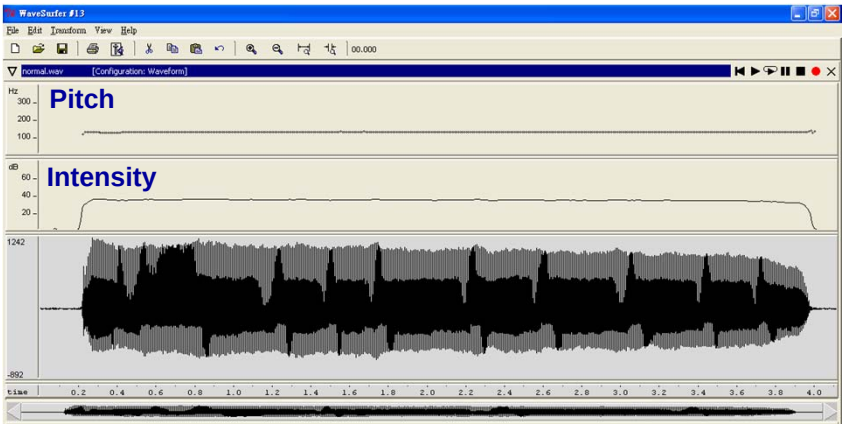
- Listening to patient's voice and making judgement
- Done by experienced voice clinicians
- Subjective, difficult to establish a standard
- Effective training tool can help inexperienced clinicians to learn

Training tool on voice assessment

Objective voice assessment

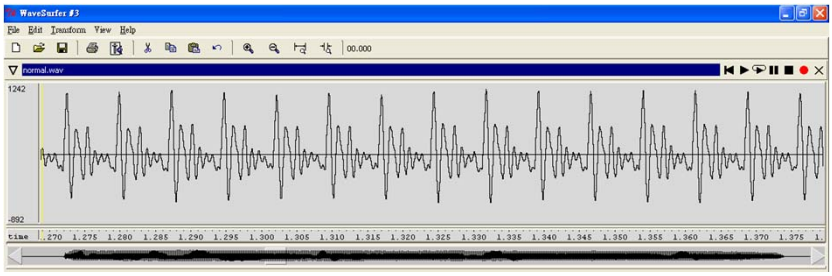
- Voice disorders lead to irregularities of speech signals
- Abnormal change of pitch and intensity
- Atypical waveforms and spectrum
- Speech processing methods are useful to classify and quantify voice disorders

Normal voice

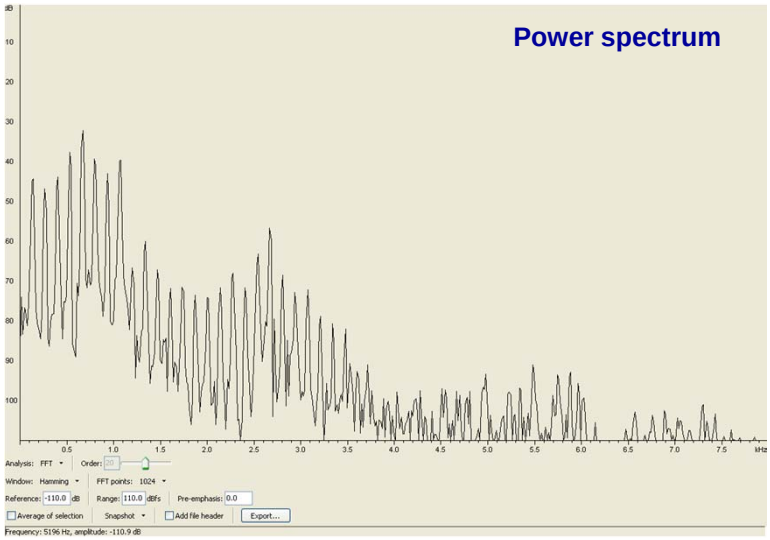


Normal voice

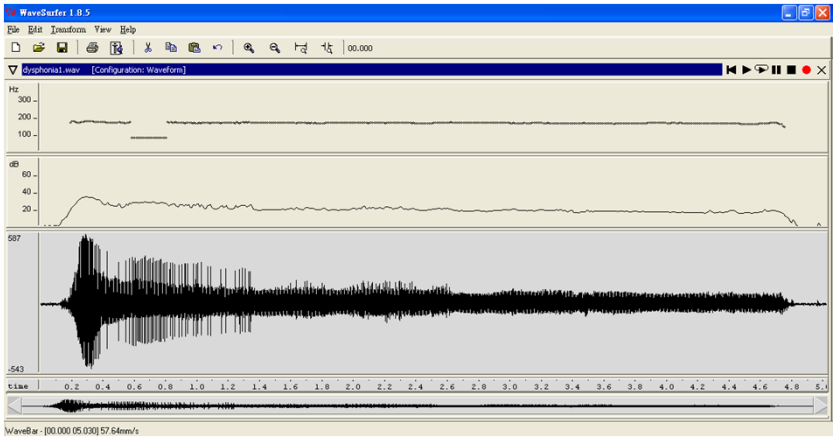
Periodic waveform cycles



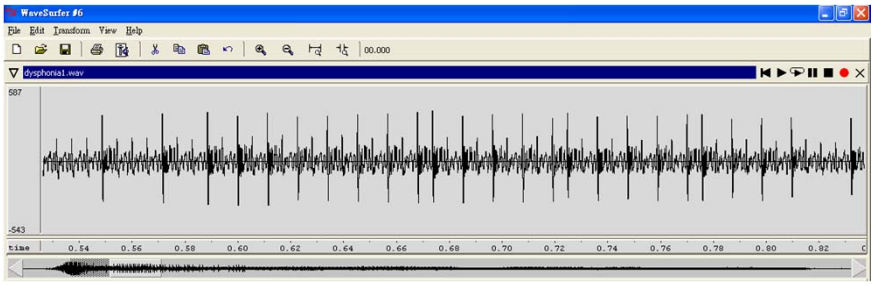
Normal voice



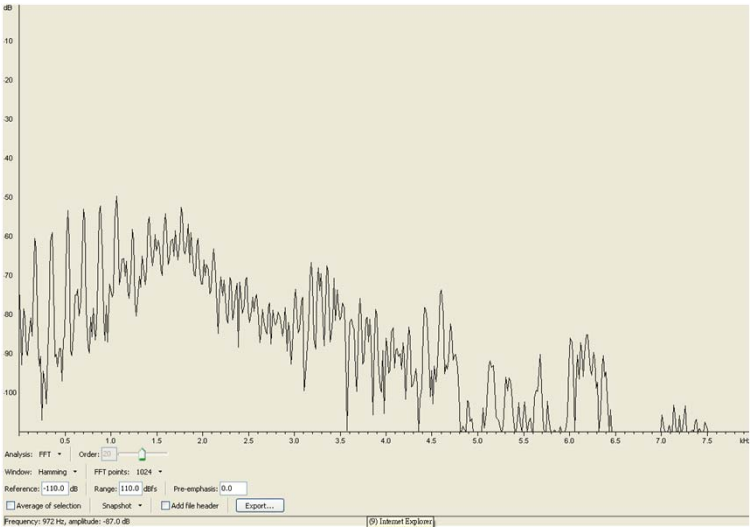
Pathological voice



Pathological voice



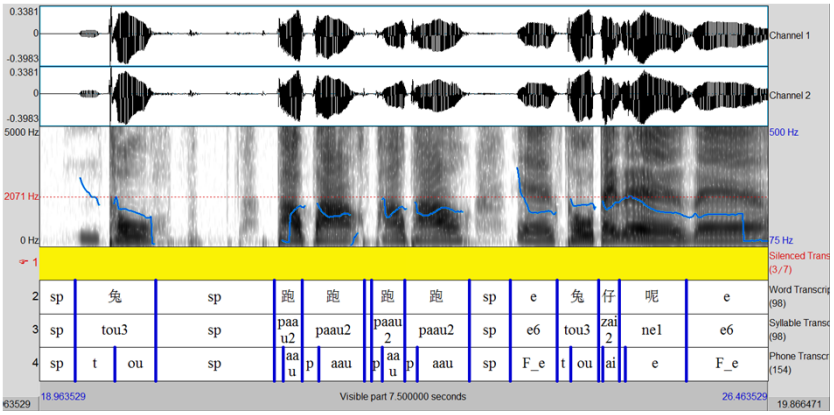
Pathological voice



Analysis of aphasia speech

- **Aphasia: impairment of language caused by stroke or other type of injury to the brain**
- **Aphasia patients have difficulties producing fluent and intelligible speech**
- **Cantonese AphasiaBank: audio and video recordings of 180 normal speakers of Cantonese and 180 individuals with Aphasia**
- **Current work aims at objective analysis and assessment of aphasia speech**

Analysis of aphasia speech



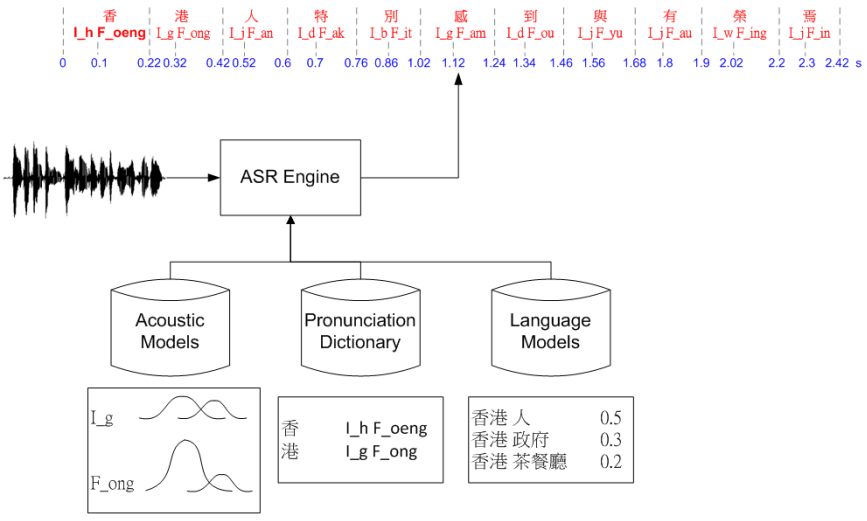
Inter-disciplinary research
Linguistics and Language

- Low-cost digital storage and rapidly growing transmission bandwidth
- Easy to find or create large corpus for linguistic and language studies
- Computer speech processing techniques can help linguistic research in various ways
- Automated annotation of speech recordings
- Statistical analysis of large corpus
- Automatic discovery of unknown patterns

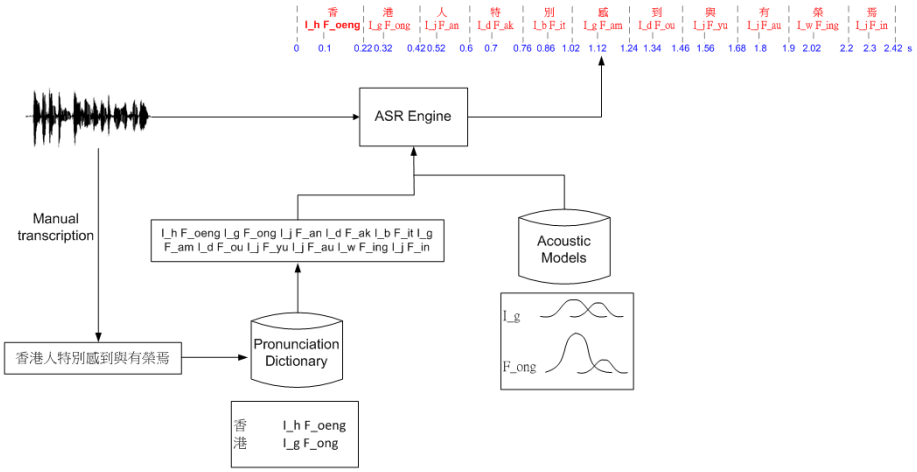
Automatic alignment of speech

- In linguistic research, we often need to segment and label speech
- To locate boundaries of words, syllables, vowels, consonants
- To locating breaks (silence periods)
- Large speech corpus: tedious, time-consuming, inconsistent
- ASR technology can be used for automatic time alignment

ASR system

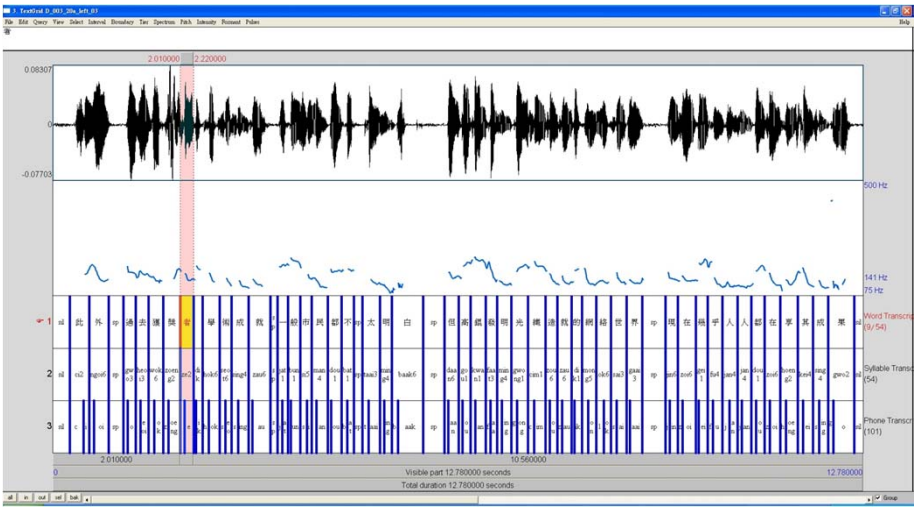


Forced alignment

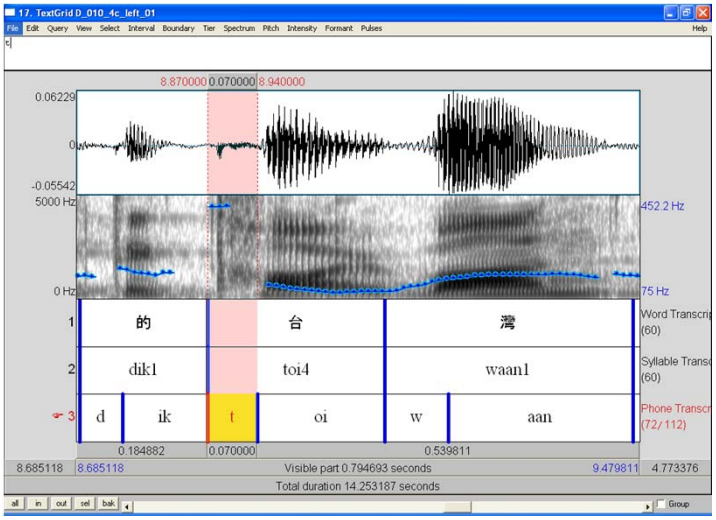


Procedures

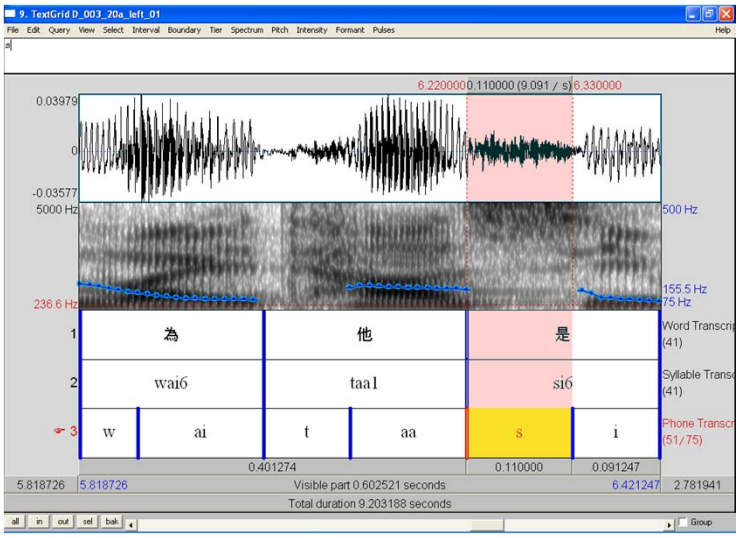
- Recordings transcribed manually
- Chinese characters and syllable sequences
- Converted into sequences of Initials (聲母) and Finals (韻母)
- Automatic alignment using acoustic models in our existing Cantonese ASR system
- Results represented in Praat format



Example 1


17. TextGrid D_010_4c_left_01
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
8.870000 0.070000 8.940000
0.06229
0
-0.05542
5000 Hz
0 Hz
452.2 Hz
75 Hz
1 的 台 灣
Word Transcri (60)
2 dik1 toi4 waan1
Syllable Trans (60)
3 d ik t oi w aan
Phone Transcr (72/112)
0.184882 0.070000 0.539811
8.685118 8.685118 Visible part 0.794693 seconds 9.479811 4.773376
Total duration 14.253187 seconds
all in out sel bak Group

Example 2



Mining a Year of Speech

- There are 100 million hours of broadcast collection in the world
- 11,415 years
- "Mining a Year of Speech": a joint project of Univ. of Oxford and University of Penn.
- `http://www.phon.ox.ac.uk/mining`
- 5234 hours of English speech processed (American and British)
- Automatic time alignment at phone and word level

Making use of automatic alignment

"Spoken corpora are not just repositories of words: they also embody specific phrases or constructions, as well as revealing - through their audio aspect - particularities of people's voices and habits of."

(Mining a Year of Speech, White Paper)

By efficiently analyzing a large amount of speech, we can answer many interesting questions with reproducible scientific evidences:

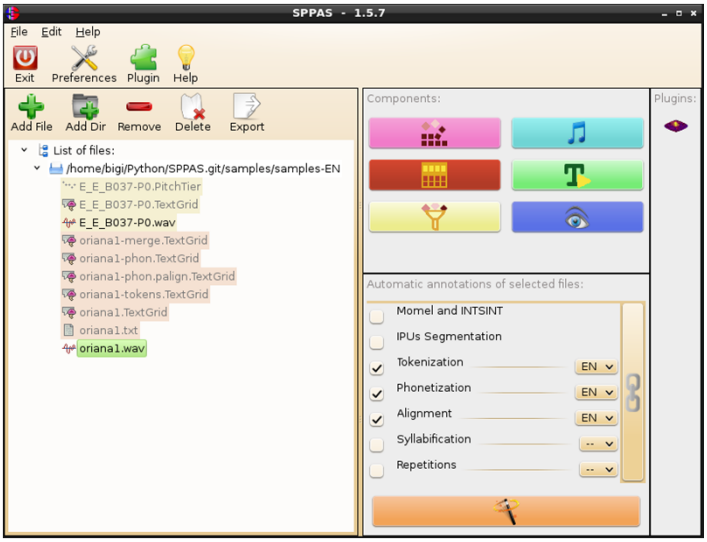
- ***"Do women speak fast than men ?"***
- ***"How is Indian English different from British English ?"***
- ***"How do people speak a long sentence ?"***

SPPAS

- **SPeech Phonetization Alignment and Syllabification**
- **A tool to automatically produce annotations from a speech recording and its transcripti**
- **Open source software by Laboratoire Parole et Langage, CNRS & Aix-Marseille Université,**
- **Currently support English, French, Spanish, Italian, Japanese, Taiwan Mandarin, Mainland Mandarin**
- **Cantonese will be added soon !**

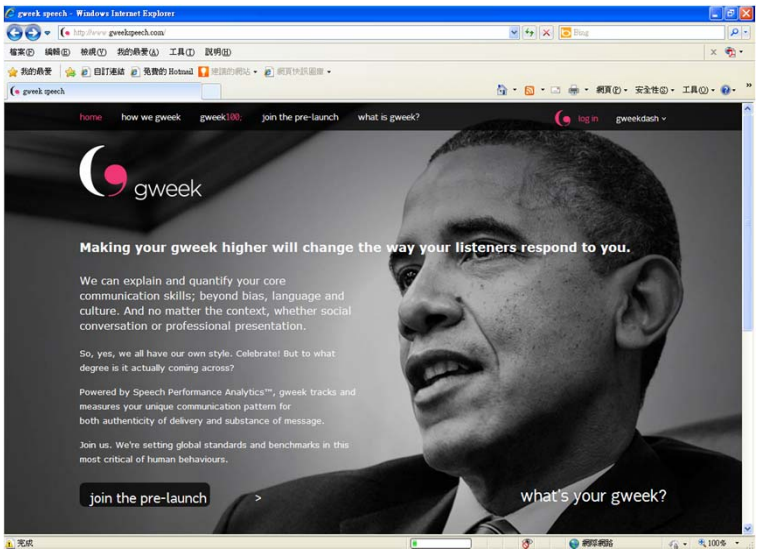
SPPAS

(downloadable from http://aune.lpl-aix.fr/~bigi/sppas/)



What is good speech ?

www.gweekspeech.com



To conclude ...

- Technological advancement is not just about Siri, YouTube, Dropbox, 小米盒子 ...
- Current technologies empower us with unprecedented abilities to explore the world
- Technologies are wide-angle camera, binoculars, and microscopes
- Learn to use them to study
  - The world of LANGUAGES, and
  - The LANGUAGES of the world

Thank you !

tanlee@ee.cuhk.edu.hk

http://dsp.ee.cuhk.edu.hk