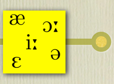
Are language learners myna birds?

-- A note of warning from a serious speech engineer --



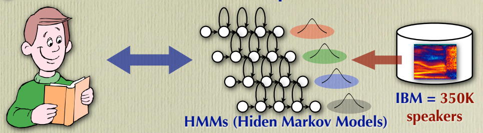
Nobuaki Minematsu
Graduate School of Engineering
The University of Tokyo

Outline of this presentation

- Something weird about ASR..., what's weird?
  - Is the current speech technology pedagogically-sound enough?
  - It seems to assume that a learner is a kind of myna bird.
- What in a teacher's voices should a learner imitate?
  - What in a father's voices should a child imitate acoustically?
  - A holistic and speaker-invariant pattern embedded in an utterance.
  - Context-sensitive perception and interpretation of stimuli
- Proposal of a new speech technology
  - Holistic and speaker-invariant representation of an utterance
  - Experimental verification of the validity of the proposed technology
- How to use the new technology for CALL
  - **!!! Visit our courseware demonstration on tomorrow !!!**

Something weird about ASR..., what's weird?

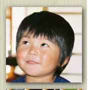
- The voices of a student are compared to those of natives

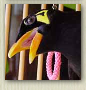


  - The demand from **A**utomatic **S**peech **R**ecognition (ASR) technology
    - It needs a huge number of training speakers to cover speaker differences
    - Difference in age and gender as well as that in microphone and channel.
  - The voices of a student are compared to distributions of natives.
    - Some normalization techniques are included to cancel these differences.
    - But the initial pronunciation scores are calculated as differences between the voices of a student and the averaged voices among native speakers.

What about boys and girls?

- No child needs so many speakers to understand speech.
  - A major part of the speech it hears is from its mother and father.
  - After it begins to talk, a large part of the speech it hears is its own.
  - Hearing a speaker-balanced corpus is completely impossible!!
- Two very fundamental facts
  - Language acquisition is based on the vocal imitation.
  - But no child imitates the voices of their parents.
    - Hearing a very good boy, no one can guess its parents.
  - Myna birds imitate the voices of their owners.
    - Hearing a very adept myna bird, one can guess its owner.
  - Are learners myna birds to the averaged voices?
- A simple and fundamental question
  - Which aspect of a father's voices does a child imitate?



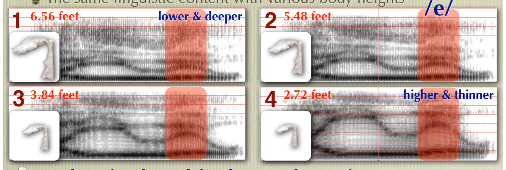


What in a father's voices is imitated?

- No child imitates the voices of its father.
- A bad hypothesis
  - "Children decompose an utterance into a phoneme sequence and then, each phoneme is converted into its sound by their mouths."
  - "They have very little phonemic awareness and cannot convert an utterance into a phonemic sequence."
- Then, what in a father's voices is imitated by children?
  - "It is the holistic sound pattern of the word, called word Gestalt."
    - Kato'03, Lieberman'80, Kato'03, Shaywitz'05, and Hayakawa'06
    - But they don't show the acoustic definition of the word Gestalt.
- One important and maybe true thing about the Gestalt
  - The Gestalt has to be speaker-invariant.
    - If it is not, children have to try to imitate the voices of their fathers.

Acoustic variability of speech

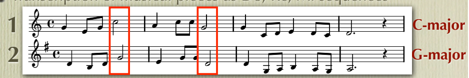
- Speech varies due to body size differences.
  - The same linguistic content with various body heights



- Another simple and fundamental question
  - Does the perception of category <x> in different segments require that some absolutely common features have to exist there?
  - Our answer is **NO!!!**. What's yours?
    - The answer from **IBM** is expected to be **YES!!!**.

Absolute and relative sense of sounds

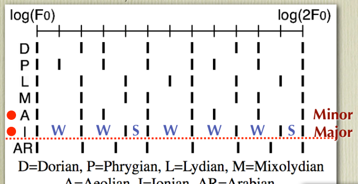
- Perception of two physically different tones as identical
  - Transcription of musical pieces as Do, Re, Mi sequences



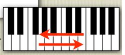
    - **1** : So-Mi-So-Do, **2** : Re-Si-Re-So   Absolute Pitch (AP)
    - **1** : So-Mi-So-Do, **2** : So-Mi-So-Do   Relative Pitch (RP) with verbalization
    - **1** : La-La-La-La, **2** : La-La-La-La   RP without verbalization
  - Two methods of naming tones : pitch names and syllable names
    - P names are assigned to physically-absolute properties of tones
      - Officially, they are CDEFGAB but Do, Re, Mi are often used (fixed Do).
    - S names are assigned to functions of tones, which are relatively defined.
      - Do(=Tonic), Re, Mi, Fa(=Subdominant), So(=Dominant),,, (movable Do)
      - S name perception follows the perception of the arrangement of tones.

RP requires the key-invariant tonal system

- Various scale structures (tonal arrangement)
  - 1 octave = log(F_0) --- log($2F_0$) with 12 semitone intervals
  - 8 tones are arranged so that they have 5 whole and 2 semi intervals.
  - With a different sound system, music can take on a different color.



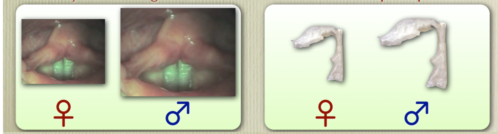
D=Dorian, P=Phrygian, L=Lydian, M=Mixolydian
A=Aeolian, I=Ionian, AR=Arabian



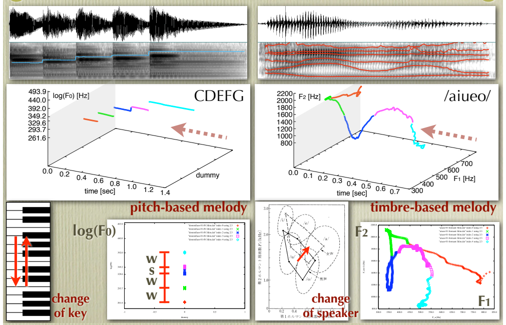
  - RP people cannot identify an isolated sound at all.

Absolute or relative, that is the question.

- Why is a father's voice lower in pitch than a mother's?
  - Because his vocal chords are heavier and longer, a physical reason.
  - With **relative pitch**, we perceive the equivalence bet. the two.
  - Only with strong **absolute pitch**, the invariant perception is hard.
- Why is a father's voice deeper in timbre than a mother's?
  - Because his vocal tract is longer, another very physical reason.
  - Then, why don't we assume **"relative timbre"** perception?
  - Only with strong **absolute timbre**, is the invariant perception hard?

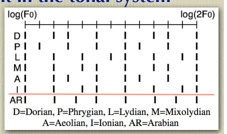


Speech = dynamic pattern of timbre



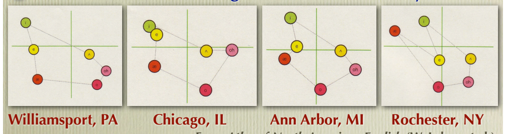
Sound system of music and language

- Variation of sound arrangement in the tonal system
  - Classical church music
    - Dorian, Phrygian, Lydian, ...
  - Major and Minor
    - Ionian = Major, Aeolian = Minor
  - Arabic scale



D=Dorian, P=Phrygian, L=Lydian, M=Mixolydian
A=Aeolian, I=Ionian, AR=Arabian

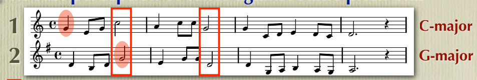
- Variation of sound arrangement in the timbre system



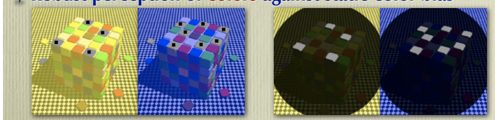
From *Atlas of North American English* (W. Labov et al.)

Context-sensitive interpretation of stimuli

- Robust perception of tones against static pitch bias



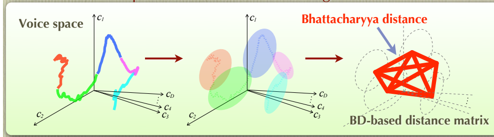
  - □ Two absolutely different tones are perceived as identical.
  - ● Two absolutely identical tones are perceived as different.
- Robust perception of colors against static color bias



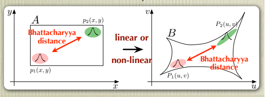
  - Two absolutely different colors are perceived as identical.
  - Two absolutely identical colors are perceived as different.

A novel and new speech technology

- A holistic and speaker-invariant sound pattern
  - A full set of speech (timbre) contrasts = a geometrical structure

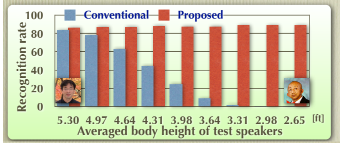


- Robustly-invariant features between two spaces
  - Every event is characterized as distribution not as point.
  - $-\log \oint \sqrt{p_1(x,y)p_2(x,y)}dxdy$
    $\equiv -\log \oint \sqrt{P_1(u,v)P_2(u,v)}dudv$
  - Contrasts are invariant.



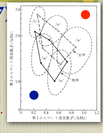
Use of the new technology for robust ASR

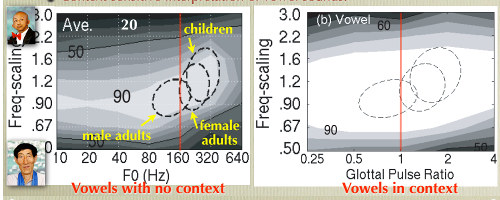
- Recognition of isolated words of /V1-V2-V3-V4-V5/
  - 120 words of /aiueo/, /aeoui/, /ioaeu/, etc
  - Training and testing
    - 4M + 4F x 120 words x 5 times = 4,800 for training
    - 4M + 4F x 120 words x 5 times = another set of 4,800 for testing
  - Comparisons (FFT-cepstrums are used and #distributions = 20 > 5)
    - Conventional : statistical modeling of very variable speech substances
    - Proposed : statistical modeling of very invariable speech contrasts



Ability to identify isolated sounds is needed?

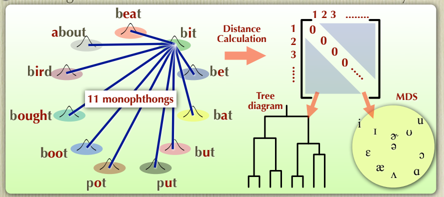
- Vowel sounds of giants and fairies(Hayashi'07)
  - Can humans identify vowels of giants and fairies?
  - Identification of **isolated** vowel sounds is difficult.
  - Identification of vowel sounds **in context** is possible.
    - **Meaningless** sequences of morae are used in experiments.
    - Context-sensitive interpretation of vowel sounds.

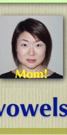




How to use this new technology for CALL

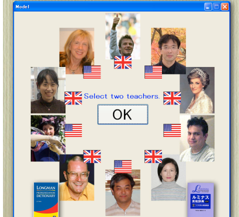
- A vowel training system for everybody!!
  - Learning not of individual vowels but of an entire vowel system



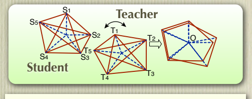


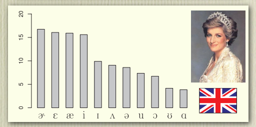
Which vowels to correct first in your case?

- A window for "favorite teacher selection"
  - A user interface impossible with the conventional technology
- Which vowels to correct first to become like him/her?
  - The system can show the shortest cut to the model pronunciation.



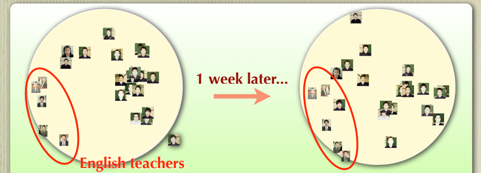
Select two teachers.





Classification of learners

- Changes of students in a class before and after training



- **!!! Visit our courseware demonstration on tomorrow !!!**

Pronunciation Clinic