Effects of vowel types on perception of speaker characteristics of unknown speakers

ATR Human Information Science Laboratories
Tatsuya Kitamura and Parham Mokhtari

This research was supported by the Ministry of Internal Affairs and Communications on their Strategic Information and Communications R&D Programme.
Introduction

Å Speech signal conveys

- Linguistic info.
- Identity
- Emotional state
- Health condition
- etc.

Auditory face (Belin et al. 2004)
Speaker individualities

Humans have individual voice characteristics as they have individual face.

Acoustic characteristics of speech sounds vary due to the differences of:

- Phoneme
- Pitch frequency
- Speaking style
- Emotional state
- Health condition
- Communication channel
- etc.

Intra-speaker variation

But, humans can robustly identify speakers of familiar voices. Why? How?
Aim

To investigate human abilities to identify speakers despite intra-speaker variations.

Hypotheses

- Humans perceive speaker individualities that persist across intra-speaker variations.
- Humans have intra-speaker variation models in their mind.

Three psychoacoustic experiments focused on intra-speaker variations due to vowel differences.
Experiments 1 & 2

- Speaker identification tests.
- To confirm whether there are speaker individualities common to sustained vowels.
- To investigate effects of dynamic features on identifying the speakers.
- To show effects of pitch frequency (F0) on identifying the speakers of vowels and sentences.
Experiments 1 & 2
Speech data & participants

Å Experiment 1
   • 4 sustained Japanese vowels uttered by 4 male native Japanese speakers.
     • /a/, /e/, /i/, and /o/.
     • Approx. 0.6 sec.

Å Experiment 2
   • 3 Japanese sentences uttered by 4 male speakers
     • /arajimu gen3itsu wo subete3ibu no ho:e ne3imagetanoda/
     • /ijiju:kaN bakari niw:yo:ku o jiwzi jita/
     • /terebi ge:mu ja pasokoN de ge:mu o jite asobu/

Å Sampling rate (Fs): 16 kHz
Å 2 tokens for each vowel and sentence.
Å 9 listeners (2 males and 7 females)
Experiments 1 & 2

Stimuli

Å Experiment 1
  ï V1: Speech waves with normalized amplitude.
  ï V2: Speech waves with normalized amplitude and pitch frequency.

Å Experiment 2
  ï S1: Speech waves with normalized amplitude.
  ï S2: Speech waves with normalized amplitude and pitch frequency.

Å Pitch frequencies of V2 and S2 were tuned to a mean value by using the STRAIGHT analysis-synthesis system (Kawahara et al. 1999).
Experiments 1 & 2
Procedure

Å ABX test

- Participants were asked to select which of the first two speakers produced the third stimulus.

Å Example

- Stimulus V1
- Stimulus V2
- Stimulus S1
- Stimulus S2

Fig. ABX sequence for Exp. 1
Experiments 1 & 2
Results

Stimulus V1

amp. normalized vowel

percent correct [%]

0 20 40 60 80 100 120

/a/ /i/ /o/

** statistically significant difference (p < 0.01)

Stimulus S1

amp. normalized sentence

percent correct [%]

0 20 40 60 80 100 120

sentence1 sentence2 sentence3

F0 normalized vowel

percent correct [%]

0 20 40 60 80 100 120

/a/ /i/ /o/

Stimulus V2

F0 normalized sentence

percent correct [%]

0 20 40 60 80 100 120

sentence1 sentence2 sentence3

Stimulus S2

percent correct [%]

0 20 40 60 80 100 120

sentence1 sentence2 sentence3
Experiments 1 & 2

Discussions

Å Possible explanation for the differences of results of Exp. 1 & 2

1. The participants could obtain dynamic features of each speaker from the different sentences, which were absent in the sustained vowel stimuli.

2. They needed speech sounds with dynamic variations in order to obtain invariant static features as cues to speaker identification.

3. They identified the speakers using speaker characteristics obtained from phonemes common to the sentences.
Experiment 3

To investigate possible speaker characteristics common to sustained vowels.

Focused on

- Higher frequency region of speech spectra.
- Glottal source pattern.
Experiment 3
Method

Å Speech data
   ï 3 sustained Japanese vowels
   ï 6 male speakers.
   ï /a/, /e/, and /o/.
   ï Approx. 1 sec.
   ï 3 tokens for each vowel
   ï Fs=16 kHz

Å ABX test procedure
Å Reference vowel is /a/.
Å 12 female participants

Å Stimuli: 4 types
   original speech wave
   normalize amplitude
   normalize F0
   down sampling (Fs=5 kHz)
   fix F0 constant
   randomize frame sequence
   stimulus 1
   stimulus 2
   stimulus 3
   stimulus 4
Experiment 3

Results

**amp. normalized**

**F0 normalized**

**down sampled**

**Frame randomized**

statistically significant difference ($p < 0.01$)
Conclusions

1. There are speaker individualities common to the sustained vowels.

2. The possible cues common to the sustained vowels:
   1. the mean of the pitch frequency,
   2. higher frequency region of speech spectra,
   3. glottal source patterns.

3. The mean of pitch frequency is not a significant cue for identification of the unknown speaker of sentences.
Speech spectra

Speaker A

Speech spectra

Speaker B
Speech production model proposed by Honda et al. (2004)

A Source-filter model of speech production
(Fant 1970)

Glottal source → Vocal tract filter → Radiation → Speech

A Honda’s model

Glottal source → Hypopharynx filter → Vocal tract proper filter → Radiation → Speech
Experiment 1
Participants

Å 9 listeners (2 males and 7 females).
Å They have never met with the speakers nor listened to the speakers' voice.
Å No hearing impairments.