

Speaker Identification (SID) in Low-Rate Coded Speech

Andrew Catellier and Stephen Voran

{acatellier;svoran}@its.bldrdoc.gov

Institute for Telecommunication Sciences

Telecommunications Theory Division

Effect of Transmission on Multimedia Quality of Service

17-19 June 2008 – Prague, Czech Republic



ITS



SID: Motivation

- New equipment for first responders
- Anecdotal complaints about system performance
 - Speaker emotional state, and speaker identity obscured
- Legitimate concerns!

SID: Motivation

- To help improve the platform, these problems must be measured
- Two experiments were designed to measure the problem
- We'll be talking about SID today

SID: Experiment Design

- Specifications:
 - Unfamiliar talkers
 - Clips with and without prosodic information (short and long clips)
 - Six simulated communication systems
 - Manageable experiment length

SID: Experiment Design

- Realization:
 - Tactical Speaker Identification Database
 - Used three males and three females
 - Three clip lengths: sentence, four digits, two digits
 - MELP, IMBE 7.2, 3.6 kbps (with and without impairments), MNRU
 - 360 total clips, experiment length around one hour

SID: Experiment Design

- CI clips produced by resampling at 8 kHz, filtering (160-3640 Hz bandpass), and then normalized to -26 dB below clipping
- Low-rate vocoders in C2, C3, C5 and C6 are similar to those used in Public Safety communication systems
- C5 and C6 have additional transmission impairments

Condition (C)	Description
C1	Null (no further processing)
C2	IMBE Codec, 7.2-kbps gross 4.4-kbps net
C3	MELP Codec, 1.2-kbps net
C4	MNRU, $Q = 6$ dB SNR
C5	IMBE Codec, 3.6-kbps gross 2.45-kbps net 7% BER, random
C6	C5+Packet Impairments+C5 Packet Impairments: create 60 ms packets, delete 10% of packets at random, insert the same number of empty packets at random and apply PLC to them

Table 1. Six conditions used in the experiment.

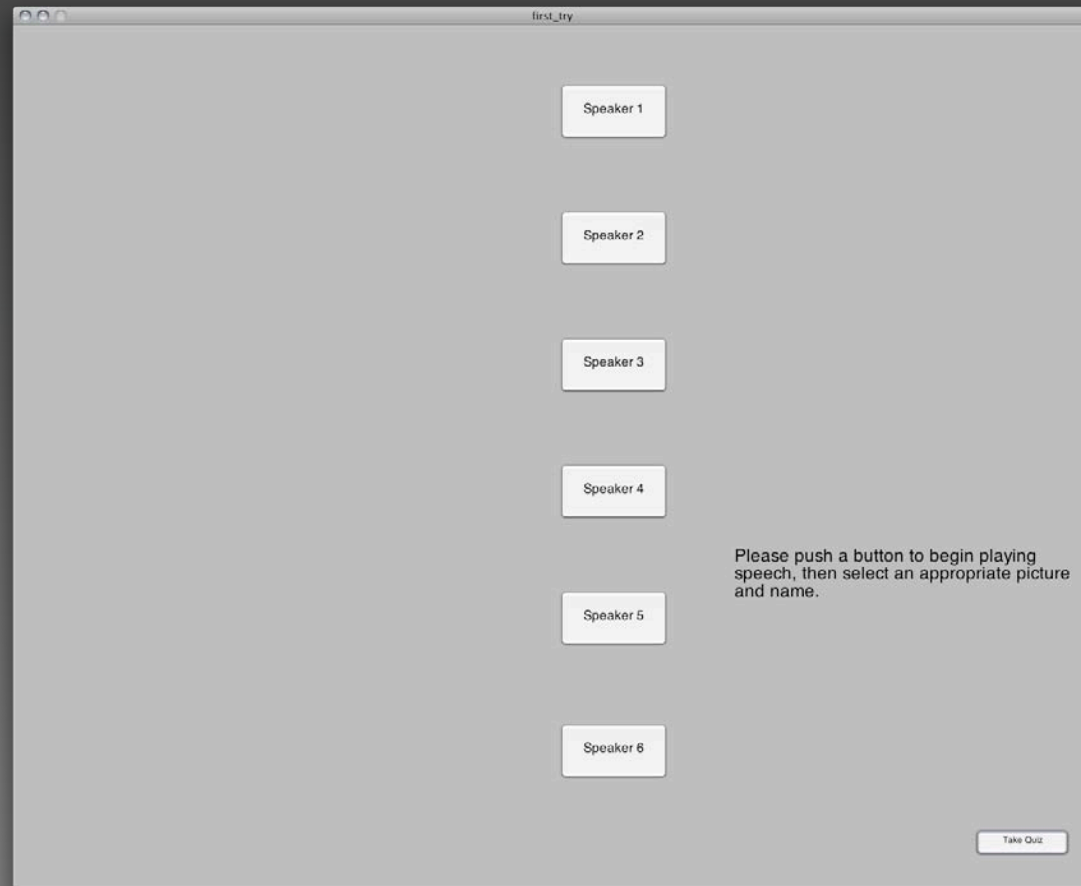
SID: Experiment Design

- Problems:
 - Training listeners to accurately recognize any given speaker
 - False confidence in training
 - Mid-test mistraining

SID: Experiment Design

- Training:
 - Used set of clips where speakers were giving directions (semi-spontaneous)
 - Allowed listeners to assign appropriate memory aids: a name and a face

SID: Experiment Design



SID: Experiment Design

The screenshot shows a web application window titled "first_try". On the left, there is a grid of 10 small portrait photos of different women. To the right of the photos is a scrollable list of names, including: Mary, Patricia, Livia, Barbara, Elizabeth, Jennifer, Maria, Susan, Margaret, Dorothy, Lisa, Nancy, Karen, Betty, Helen, Sandra, Donna, Dani, Ruth, Sharon, Michelle, Laura, Sarah, Kimberly, Deborah, Jessica, Sherry, Cynthia, Angela, Melissa, Brenda, Amy, Anna, Rebecca, Virginia, Kathleen, Pamela, Martha, Debra, Amanda, Stephanie, Carolyn, Christine, Marie, Janet, Catherine, Frances, Ann, Joyce, Diane, Alice, Julie, Heather, Tamara, Doris, Gloria, Evelyn, Jean, Cheryl, Michael, Katherine, Joan, Ashley, Judith, Rose, James, Kelly, Nicole, Judy, Christina, Kathy, Theresa, Beverly, Denise, Lamey, Irene, Jane, Lori, Rachel, Marilyn, Andrea, Kathryn, and Elizabeth. Below the list is a text input field with the placeholder "...or type a name." and a small button. To the right of the name list are six buttons labeled "Speaker 1" through "Speaker 6". Below these buttons, the text "now playing speaker 1" is displayed. Further down, a larger text prompt reads: "Please pick an appropriate picture and select or type an appropriate name for the current speaker." At the bottom right, there are two buttons: "Stop Audio" and "Take Quiz".

SID: Experiment Design

The screenshot shows a web-based interface for the SID experiment. At the top, the window title is "first_try". The main area is titled "Chose a name...". On the left, there is a grid of 10 small portrait photos of different women. In the center, a vertical list contains 40 names: Cici, Isabella, Rosario, Oia, Jantine, Milie, Lupe, Alisa, Lou, Maribel, Suzanne, Betta, Susana, Elso, Cecile, habelle, Lesley, Jocelyn, Paige, JONI, Rachelle, Lucia, Stephanie, Alta, Ester, Pella, Gracela, Imogenia, Joline, Krishna, Lacey, Gloria, Cathrina, Ken, Ursula, Luzie, Kirsten, Shana, Adeline, Maria, Jayne, Jaclyn, Grace, Sonnia, Carreola, Mania, Rosalind, Charly, Tomi, Beatriz, Manosi, Ciance, Jeanne, Shaina, Angeline, Paola, Lily, Hozzie, Shauna, Mike, Claudette, Catherine, Angela, Gabriela, Autumn, Katharine, Summer, Jodie, Staci, Lisa, Christi, James, Justine, Elma, Luella, Margot, Dominique, Jocelyn, Rene, Marina, Mango, Mavis, and Colia. Below the list is a text input field with the prompt "...or type a name." and a small button. To the right of the name list are six buttons labeled "Speaker 1" through "Speaker 6". On the far right, a large portrait of a woman is shown. Below it, the text "now playing speaker 1" is displayed. Further down, a larger instruction reads: "Please pick an appropriate picture and select or type an appropriate name for the current speaker." At the bottom right, there are two buttons: "Stop Audio" and "Take Quiz".

SID: Experiment Design

The screenshot shows a web-based interface for a speech identification experiment. The window title is "first_try". On the left, there is a grid of 10 small portrait photos of different women. In the center, a scrollable list titled "Chose a name..." contains 30 names: Cici, Justine, Rosario, Oia, Janine, Maline, Lupe, Alisa, Lou, Maribel, Suzanne, Betta, Susana, Elso, Cecile, babelle, Lesley, Jocelyn, Paige, JONI, Rachelle, Lucia, Stephanie, Alta, Ester, Pella, Gracela, Imogenia, Joline, Krishna, Lacey, Gloriana, Catharina, JESSICA, Ursula, Luzie, Kirsten, Shana, Adeline, Maria, Jayne, Jaclyn, Grace, Sondra, Carmeda, Mania, Rosalind, Charly, Tonia, Beatriz, Manosi, Ciance, Jeanne, Shaina, Angeline, Paola, Lily, Rozzie, Shauna, Mike, Claudette, Cathleen, Angela, Gaudilo, Autumn, Katharine, Summer, Jodie, Stasi, Lisa, Christi, James, Justine, Elma, Luella, Margot, Dominique, Jocelene, Rene, Martina, Margo, Mavis, and Colbie. Below the list is a text input field with the prompt "...or type a name." and a small button. On the right side, there are six buttons labeled "Speaker 1" through "Speaker 6". A large portrait of a woman is displayed, with the name "Keri" centered below it. Underneath the name, it says "now playing speaker 1". Below this, there is a text instruction: "Please pick an appropriate picture and select or type an appropriate name for the current speaker." At the bottom right, there are two buttons: "Stop Audio" and "Take Quiz".

SID: Experiment Design

first_try

Chose a name...

James
John
Robert
Michael
William
David
Richard
Charles
Joseph
Thomas
Christopher
Daniel
Paul
Mark
Donald
George
Kathleen
Steven
Edward
Dean
Ronald
Anthony
Kevin
Jason
Matthew
Gary
Timothy
Jane
Larry
Jeffrey
Frank
Scott
Eric
Stephan
Andrew
Raymond
Gregory
Joshua
Jerry
Dennis
Walter
Patrick
Peter
Harold
Douglas
Henry
Carl
Arthur
Ryan
Roger
Joe
Juan
Jacob
Albert
Jonathan
Justin
Levy
Gerald
Keith
Samuel
Willie
Ralph
Lawrence
Nicholas
Roy
Benjamin
Bruce
Brandon
Adam
Harry
Fred
Wayne
Bobby
Steve
Louis
Jeremy
Aaron
Harold
Howard
Eugene
Carlton
Russell
Burt

...or type a name.

Speaker 1

Speaker 2

Speaker 3

Speaker 4

Speaker 5

Speaker 6

now playing speaker 2.

Please pick an appropriate picture and select or type an appropriate name for the current speaker.

Stop Audio

Take Quiz



SID: Experiment Design

The screenshot shows a software window titled "first_try" with the following components:

- Chose a name...:** A list of 40 names including James, John, Robert, Michael, William, David, Richard, Charles, Joseph, Thomas, Christopher, Daniel, Paul, Mark, Donald, George, Kenneth, Douglas, Edward, Brian, Ronald, Anthony, Kevin, Jason, Matthew, Gary, Timothy, Jose, Larry, Jeffrey, Frank, Scott, Eric, Stephen, Andrew, Raymond, Gregory, Joshua, Jerry, Dennis, Walter, Patrick, Peter, Harold, Douglas, Henry, Carl, Arthur, Ryan, Roger, Jose, Juan, Justin, Albert, Jonathan, Justin, Larry, Gerald, Keith, Samuel, Willie, Hugh, Lawrence, Nicholas, Roy, Benjamin, Bruce, Brandon, Adam, Harry, Fred, Wayne, Bob, Steve, Louis, Jeremy, Aaron, Randy, Howard, Eugene, Carlos, Russell, and Barbara.
- Speaker Selection:** A vertical column of six buttons labeled "Speaker 1" through "Speaker 6".
- Speaker Image:** A large portrait of a man with short brown hair and a grey t-shirt, identified as Speaker 2.
- Instructions:** Text that reads "Please pick an appropriate picture and select or type an appropriate name for the current speaker."
- Controls:** "Stop Audio" and "Take Quiz" buttons at the bottom right.
- Input Field:** A text box at the bottom left with the prompt "...or type a name."

SID: Experiment Design

The screenshot shows a web-based interface for the SID experiment. The window title is "first_try". On the left, there is a grid of 12 small portrait photos of different men. To the right of the grid is a scrollable list of names. The names listed are: James, John, Robert, Michael, William, David, Richard, Charles, Joseph, Thomas, Christopher, Daniel, Paul, Mark, Donald, George, Kenneth, Douglas, Edward, Brian, Ronald, Anthony, Kevin, Jason, Matthew, Gary, Timothy, Jane, Larry, Jeffrey, Frank, Scott, Eric, Stephen, Andrew, Raymond, Gregory, Joshua, Jerry, Dennis, Walter, Patrick, Peter, Harold, Douglas, Henry, Carl, Arthur, Ryan, Roger, Joe, Juan, Justin, Albert, Jonathan, Justin, Larry, Gerald, Keith, Samuel, Willie, Hugh, Lawrence, Nicholas, Roy, Benjamin, Bruce, Brandon, Adam, Harry, Fred, Wayne, Billy, Steve, Louis, Jeremy, Aaron, Randy, Howard, Eugene, Carlos, Russell, and Roberto. Below the list is a text input field with the placeholder "...or type a name." and a small button. On the right side of the interface, there are six buttons labeled "Speaker 1" through "Speaker 6". A large portrait of a man is displayed, with the name "Dennis" written below it. Underneath the name, it says "now playing speaker 2". Below this is a text prompt: "Please pick an appropriate picture and select or type an appropriate name for the current speaker." At the bottom right, there are two buttons: "Stop Audio" and "Take Quiz".

SID: Experiment Design

The screenshot shows a web-based application window titled "first_try". On the left, there is a grid of 18 small portrait photographs of women. To the right of the grid is a scrollable list of names, including Mary, Patricia, Livia, Barbara, Elizabeth, Jennifer, Maria, Susan, Margaret, Dorothy, Lisa, Nancy, Karen, Betty, Helen, Sandra, Donna, Danni, Ruth, Sharon, Michelle, Laura, Sarah, Kimberly, Deborah, Jessica, Sherry, Cynthia, Angela, Melissa, Brenda, Amy, Anna, Rebecca, Virginia, Kathleen, Pamela, Maria, and Dora. Below the list is a text input field with the placeholder "...or type a name." and a small button. In the center, there are six buttons labeled "Speaker 1" through "Speaker 6" arranged vertically. To the right of these buttons, the text "now playing speaker 3." is displayed. Below this, a larger instruction reads: "Please pick an appropriate picture and select or type an appropriate name for the current speaker." At the bottom right, there are two buttons: "Stop Audio" and "Take Quiz".

SID: Experiment Design

first_try

Chose a name...

Mary
Patricia
Linda
Barbara
Elizabeth
Jennifer
Marie
Susan
Margaret
Dorothy
Lisa
Nancy
Karen
Betty
Heleen
Dorinda
Dorinda
Cami
Ruth
Sharon
Michelle
Lanae
Sarah
Kimberly
Deborah
Jessica
Shirley
Cynthia
Angela
Melissa
Brenda
Amy
Anna
Rebecca
Virginia
Katherine
Pamela
Martha
Della
Janet
Stephanie
Carolyn
Christine
Marie
Janet
Catherine
Frances
Ann
Joyce
Diane
Alice
Julie
Heather
Teresa
Dora
Gloria
Evelyn
Jean
Cheryl
Michael
Katherine
Joan
Ashley
Judith
Rose
Janice
Kathy
Nicole
Judy
Christina
Kathy
Theresa
Beverly
Denise
Laney
Irene
Jane
Lori
Rachel
Marylyn
Andrea
Kathryn
Laurie

Speaker 1

Speaker 2

Speaker 3

Speaker 4

Speaker 5

Speaker 6

now playing speaker 3.

Please pick an appropriate picture and select or type an appropriate name for the current speaker.

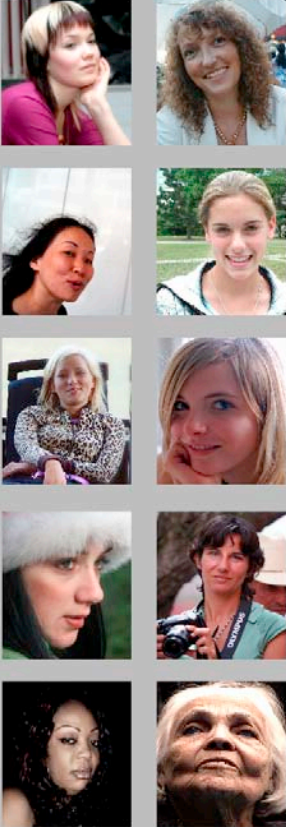
...or type a name.

Stop Audio Take Quiz

SID: Experiment Design

first_try

Chose a name...



- Mary
- Patricia
- Linda
- Barbara
- Florence
- Jennifer
- Maria
- Susan
- Margaret
- Dorothy
- Lois
- Nancy
- Egan
- Betty
- Helen
- Bandra
- Donna
- Dani
- Hull
- Sharon
- Michelle
- Laura
- Sarah
- Kimberly
- Deborah
- Jessica
- Bonny
- Cynthia
- Angela
- Melissa
- Brenda
- Amy
- Anna
- Rebecca
- Virginia
- Kathleen
- Parvula
- Maria
- Dora
- Amanda**
- Suzanne
- Carolyn
- Christine
- Maria
- Janet
- Catherine
- Frances
- Ann
- Joyce
- Diane
- Alice
- Julie
- Hester
- Teresa
- Dora
- Gloria
- Evelyn
- Jean
- Cheryl
- Mildred
- Katherine
- Jean
- Ashley
- Judith
- Rose
- Janice
- Kelly
- Nicole
- Judy
- Christina
- Kathy
- Theresa
- Beverly
- Denise
- Laney
- Terrie
- Jane
- Lori
- Rachel
- Maryn
- Andrea
- Kathryn
- Patricia

Speaker 1

Speaker 2

Speaker 3

Speaker 4

Speaker 5

Speaker 6

now playing speaker 3

Please pick an appropriate picture and select or type an appropriate name for the current speaker.

...or type a name.

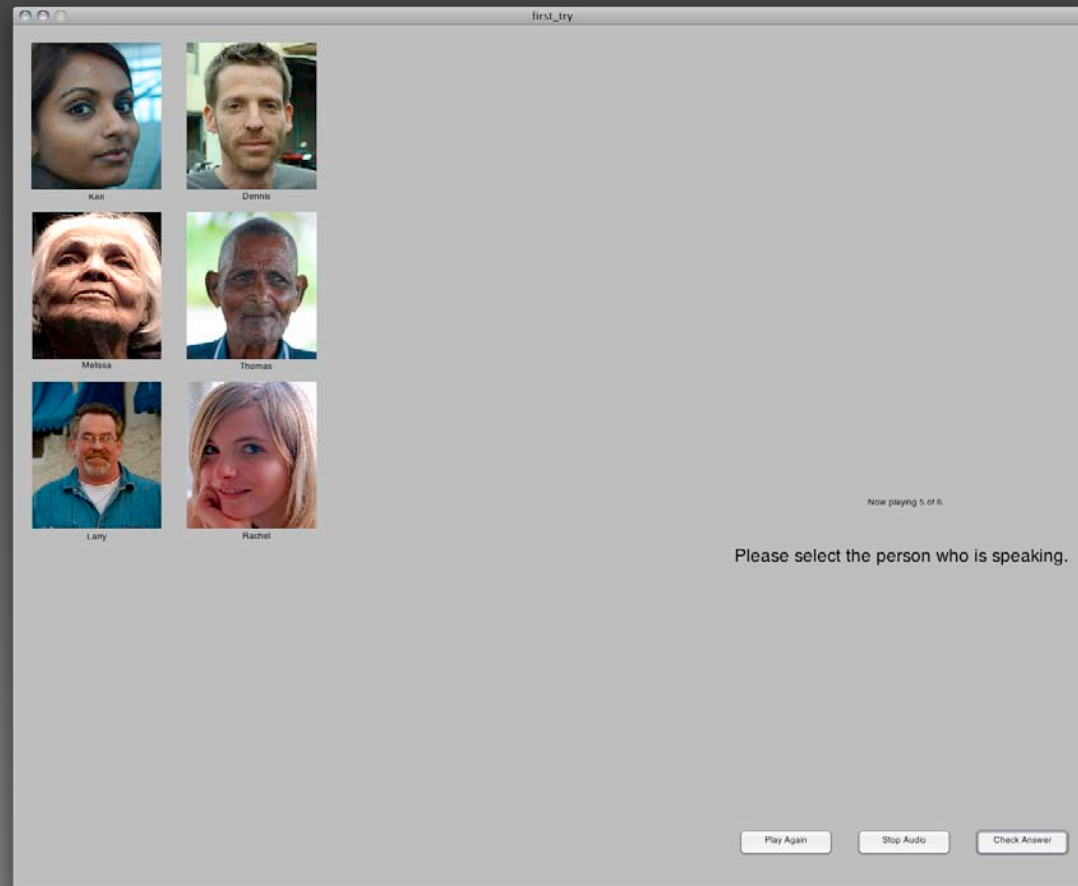
Stop Audio

Take Quiz

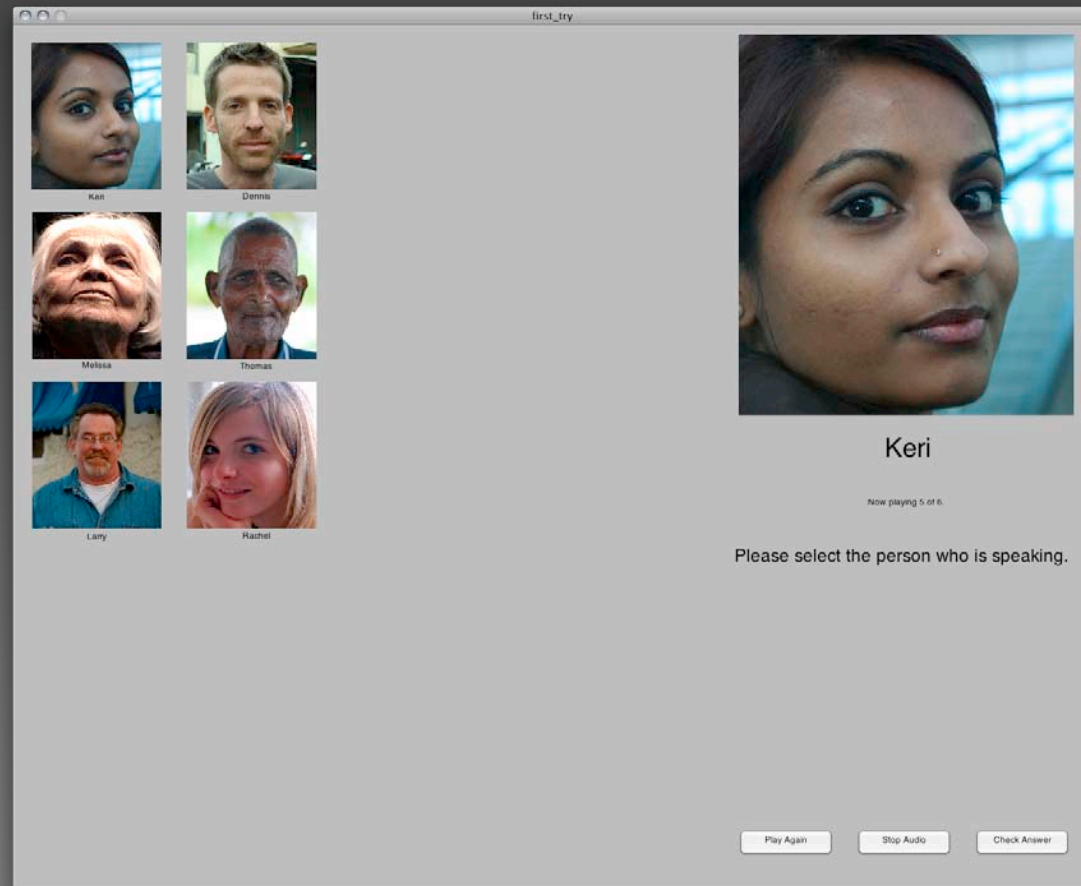
SID: Experiment Design

- Quiz:
 - Undistorted speech
 - Provided feedback to listener about state of training
 - Familiarized listener with test process

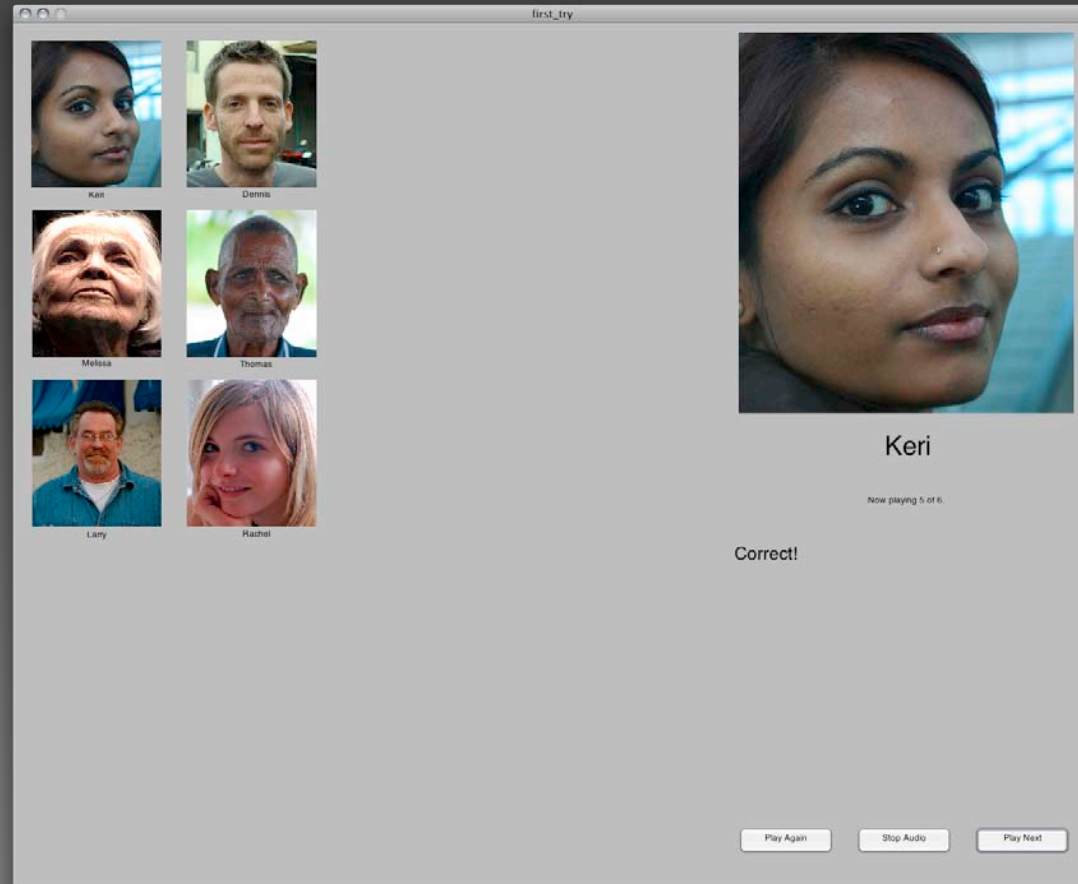
SID: Experiment Design



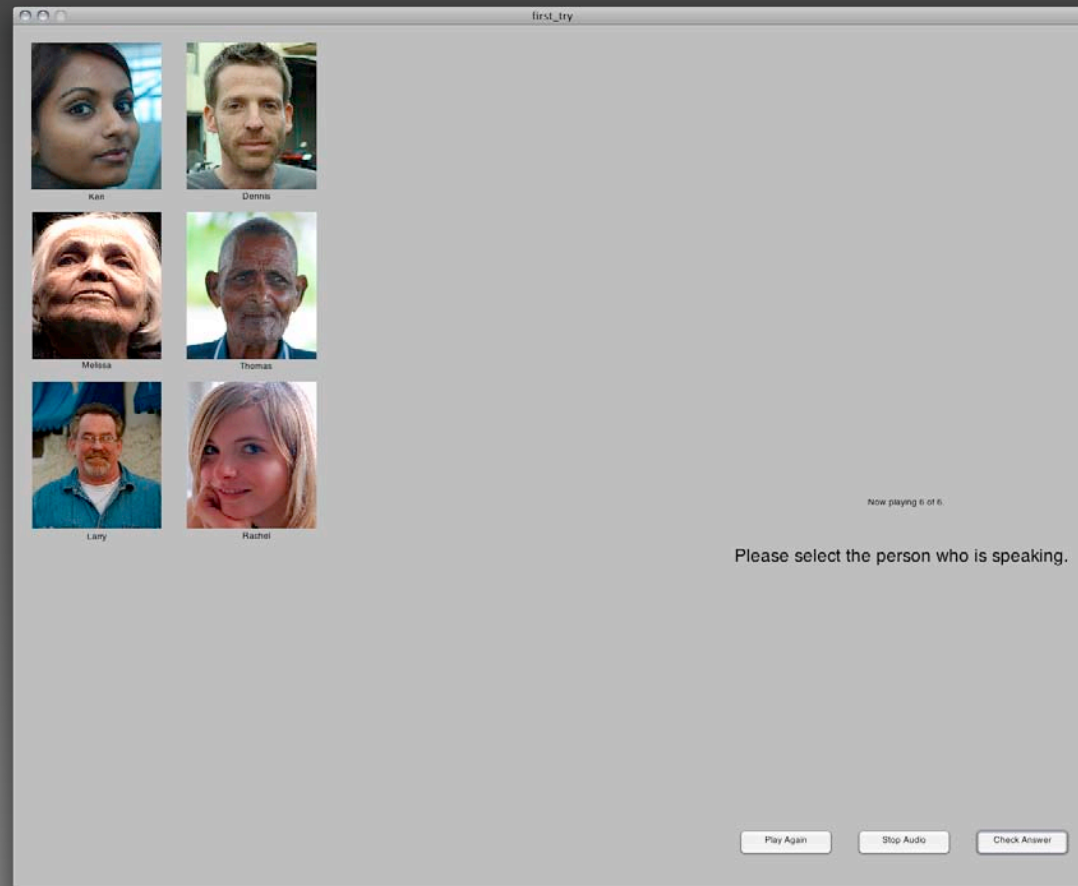
SID: Experiment Design



SID: Experiment Design



SID: Experiment Design



SID: Experiment Design

first_try

Kari Dennis

Melissa Thomas

Larry Rachel

Larry

Now playing 6 of 6.

Please select the person who is speaking.

Play Again Stop Audio Check Answer

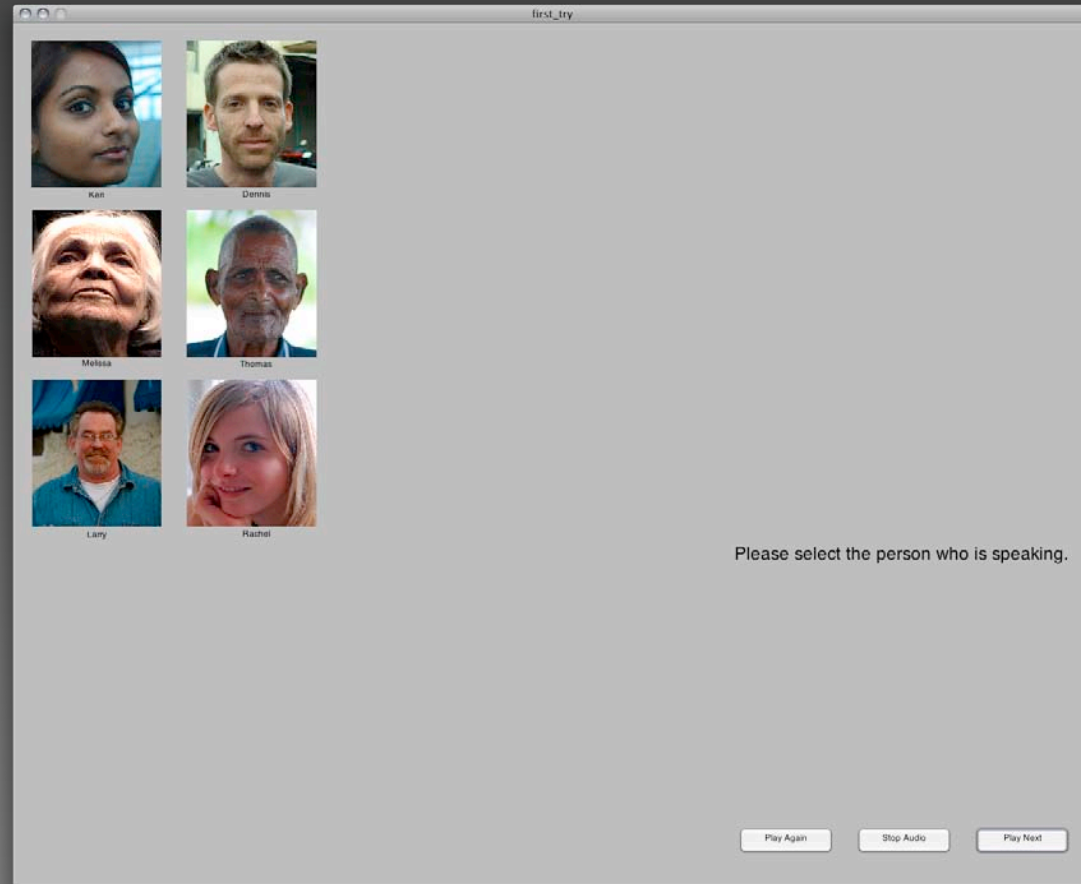
SID: Experiment Design

The screenshot shows a web browser window titled "first_try". On the left, there is a 3x2 grid of six small portrait photos of different people, each with a name label below it: Kari, Dennis, Melissa, Thomas, Larry, and Rachel. On the right, a larger portrait of a man is displayed, with the name "Dennis" centered below it. Underneath the name, it says "Now playing 6 of 6." Below that, a message reads "Incorrect. The correct answer is shown above." At the bottom right of the interface, there are three buttons: "Play Again", "Stop Audio", and "Play Next".

SID: Experiment Design

- Test sessions
 - Similar to the quiz session, but with no feedback
 - Three test sessions
 - Each clip length had its own session (sentences were first, then four digit clips, then two digit clips)
 - Same interface for all three clip lengths

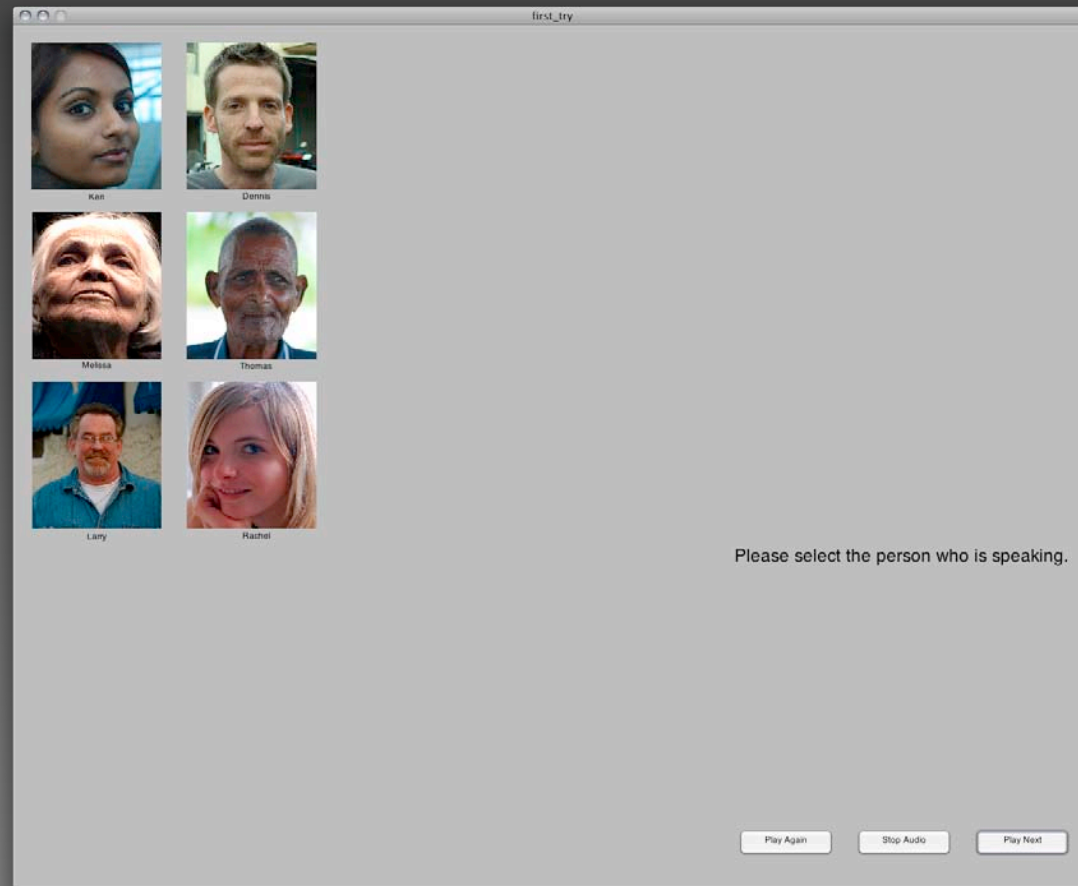
SID: Experiment Design



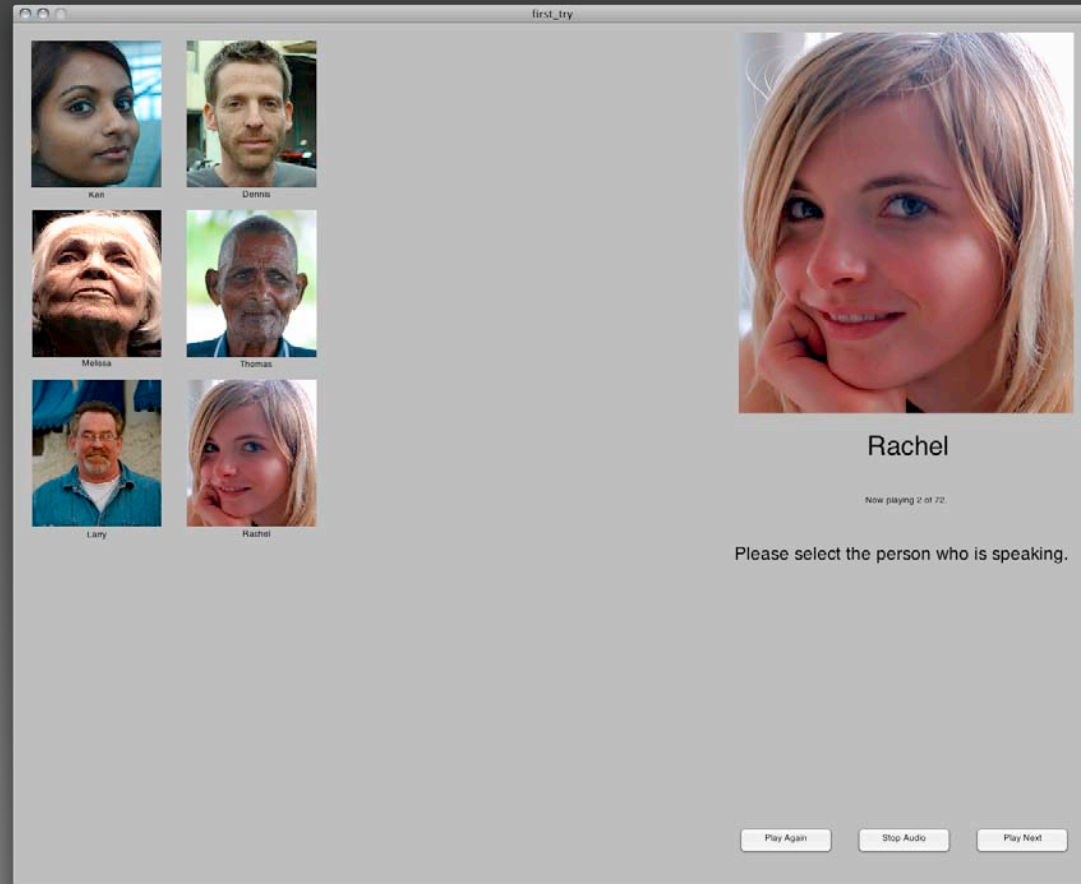
SID: Experiment Design

The screenshot shows a software window titled "first_try". On the left, there is a 3x2 grid of small portrait photos of six different people, each with their name written below: Kar, Dennis, Melissa, Thomas, Larry, and Rachel. On the right, a larger portrait of an elderly man, Thomas, is displayed. Below this portrait, the name "Thomas" is written in a large font. Underneath the name, it says "Now playing 1 of 72". Below that, the instruction "Please select the person who is speaking." is displayed. At the bottom right of the window, there are three buttons: "Play Again", "Stop Audio", and "Play Next".

SID: Experiment Design



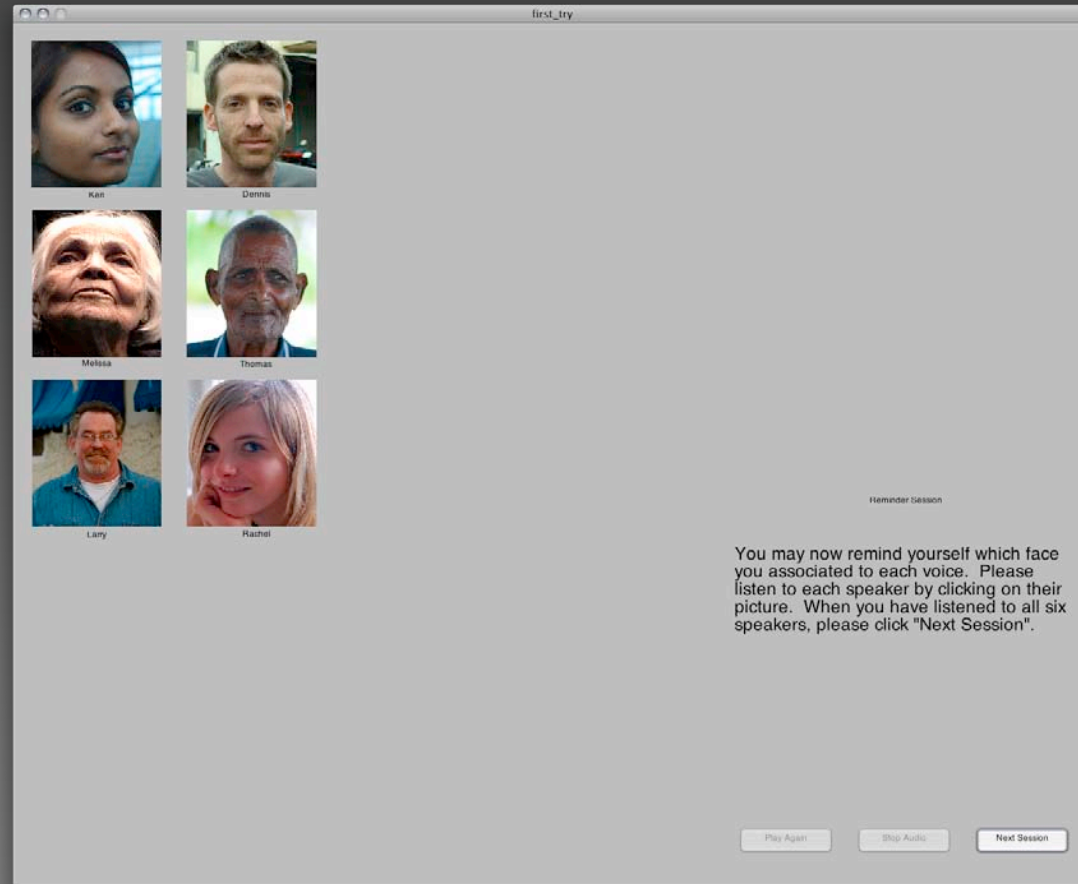
SID: Experiment Design



SID: Experiment Design

- Reminder sessions
 - Designed to keep the effects of mid-test mistraining from tainting results of next session
 - Clips representative of those to be heard in the next session were used
 - Listeners had to listen to each speaker at least once

SID: Experiment Design



SID: Experiment Design

first_try

Kari Dennis

Melissa Thomas

Larry Rachel

Larry

Reminder Session

You may now remind yourself which face you associated to each voice. Please listen to each speaker by clicking on their picture. When you have listened to all six speakers, please click "Next Session".

Play Again Stop Audio Next Session

SID: Experiment Design

first_try

Kari Dennis

Melissa Thomas

Larry Rachel

Melissa

Reminder Session

You may now remind yourself which face you associated to each voice. Please listen to each speaker by clicking on their picture. When you have listened to all six speakers, please click "Next Session".

Play Again Stop Audio Next Session

SID: Experiment Design

The screenshot shows a software interface for a speech identification experiment. At the top, the window title is "first_try". On the left, there is a grid of 16 small portrait photos of different women. To the right of this grid is a scrollable list of names. Below the name list is a text input field with the prompt "...or type a name." and a small button. In the center, there are six buttons labeled "Speaker 1" through "Speaker 6". On the right side, there is a large portrait of an elderly woman with white hair, labeled "Melissa". Below her name, it says "now playing speaker 3". At the bottom right, there are two buttons: "Stop Audio" and "Take Quiz".

Chose a name...

Mary
Patricia
Linda
Barbara
Elizabeth
Jennifer
Marie
Susan
Margaret
Dorothy
Lisa
Nancy
Karen
Betty
Heather
Dorinda
Dorinda
Cami
Hull
Sharon
Michelle
Lanae
Sarah
Kimberly
Deborah
Jessica
Shirley
Cynthia
Angela
Cynthia
Brenda
Amy
Anna
Rebecca
Virginia
Kathleen
Pamela
Martha
Debra
Amanda
Stephanie
Carolyn
Christine
Marie
Janet
Catherine
Frances
Ann
Joyce
Diane
Alice
Julie
Heather
Teresa
Dora
Gloria
Evelyn
Jean
Cheryl
Michael
Katherine
Joan
Ashley
Judith
Rose
Janice
Kathy
Nicole
Judy
Christina
Kathy
Theresa
Beverly
Denise
Laney
Irene
Jane
Lori
Rachel
Maryann
Andrea
Kathryn
Laurie

Speaker 1

Speaker 2

Speaker 3

Speaker 4

Speaker 5

Speaker 6

Melissa

now playing speaker 3

Please pick an appropriate picture and select or type an appropriate name for the current speaker.

Stop Audio

Take Quiz

...or type a name.

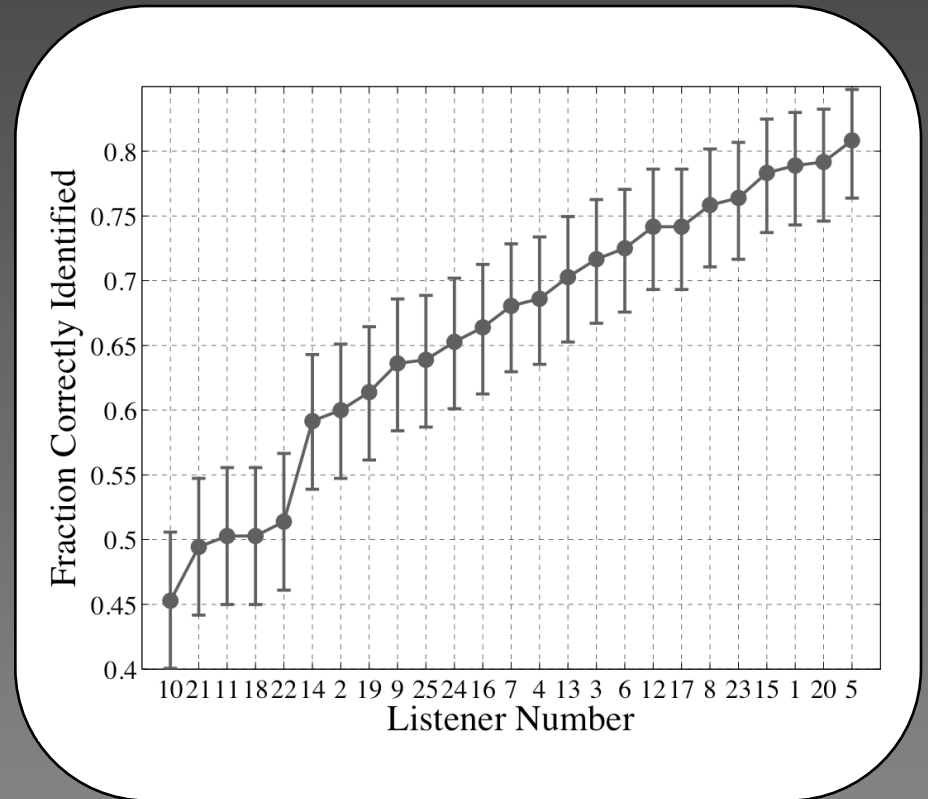
SID: Administrivia

- 25 listeners
 - 15 male
 - 10 female
 - Ages 37-64, Mean: 49
 - Scientists, mathematicians, IT professionals, desk workers
 - Native languages: English (22), Spanish (1), German (1), Russian (1)



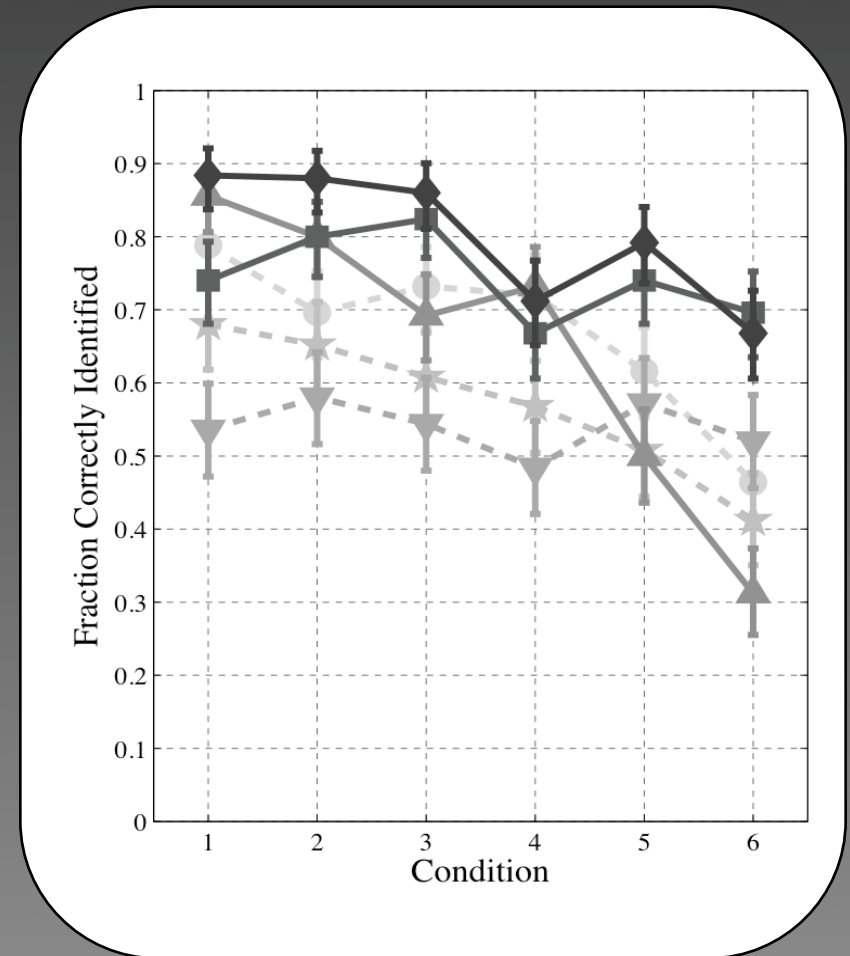
SID: Results

- Per Listener Results
 - Mean fraction of correct identifications: .662
 - 20 listeners fall between fractions .59 and .81
 - Two hearing aid users (14,16), one subject deaf in one ear (20)
 - Experiment administrator achieved a fraction correct of .98 (not included in analysis)



SID: Results

- Per Speaker Results
 - Dotted lines = males
 - Solid lines = females
 - One female very recognizable (also has Ecuadorian accent)
 - Males more often confused



SID: Results

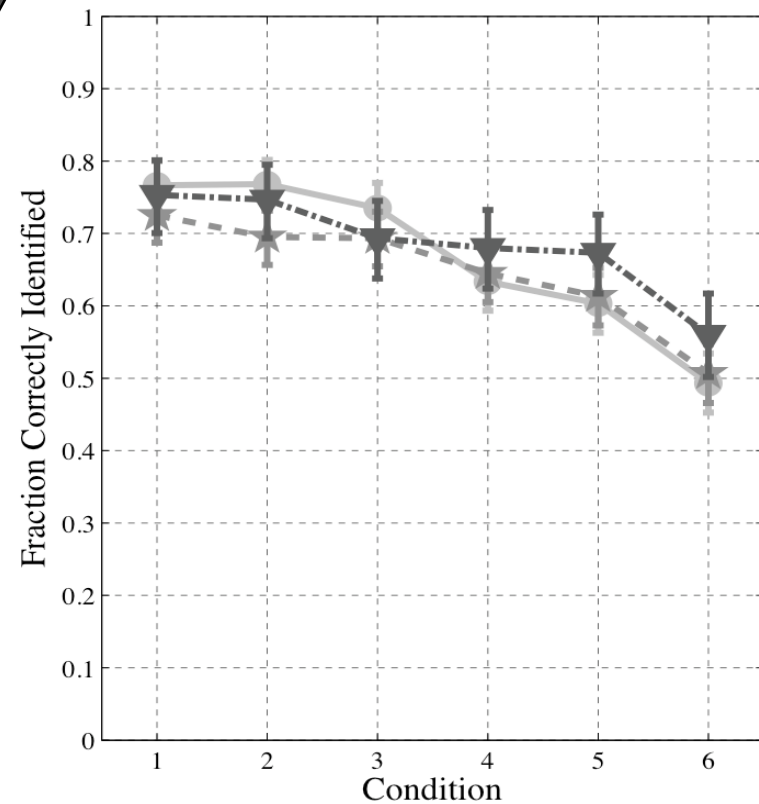
- Confusion Matrix
 - Male-female confusion is very low
 - Males 2 and 3 most often confused
 - Females 2 and 3 most easily recognized

	M1	M2	M3	F1	F2	F3
M1	0.67	0.22	0.11	0.00	0.00	0.00
M2	0.15	0.57	0.22	0.01	0.03	0.01
M3	0.12	0.34	0.54	0.00	0.00	0.00
F1	0.00	0.003	0.001	0.65	0.19	0.16
F2	0.00	0.004	0.001	0.17	0.74	0.08
F3	0.001	0.003	0.005	0.07	0.12	0.80

Table 2. Confusion Matrix: rows indicate the actual speaker, columns indicate the speaker selected by listeners. “M” indicates male, “F” indicates female. Shaded cells indicate a fraction of correct SID, unshaded cells indicate a fraction of confused SID.

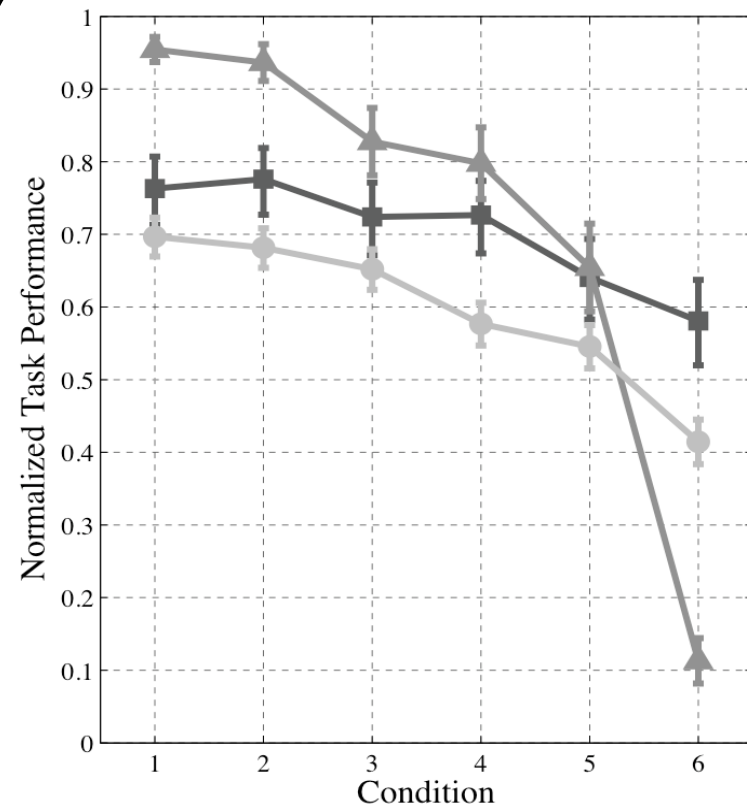
SID: Results

- Per Length Results
 - Interesting outcome: no length is significantly easier!
 - Consistent with prior research, but unintuitive
 - Experimental order (sentence, four digits, two digits) may have had an effect



SID: Results

- SID Vs. Intelligibility and Stress Detection
 - SID is not as robust as dramatized urgency (DU) detection
 - About 3 times more robust than intelligibility
 - Light gray: SID, medium gray: intelligibility, dark gray: DU detection

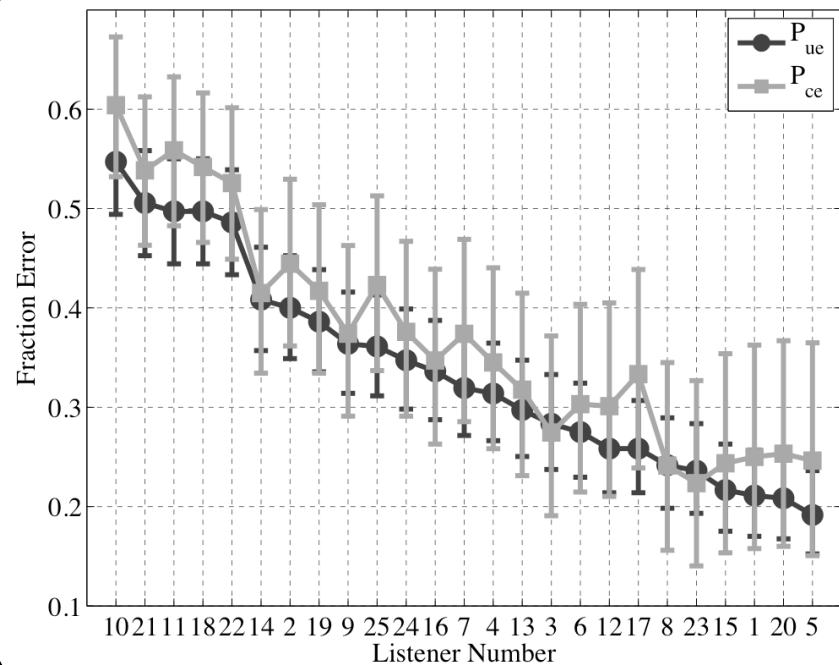


SID: Post-Hoc Work

- We had these questions while we were conducting the test:
 - Is an “event” causing temporary mistraining?
 - How often does a “confusion” result in a more permanent mistraining?
 - How often is a speaker assigned a similar memory aid?
 - How often are clips replayed?

SID: Post-Hoc Work

- Many listeners showed a slight tendency towards “bursty” errors
- Clearly not enough data
- Can’t say anything about permanent mistraining either



SID: Post-Hoc Work

M1



3 listeners



3 listeners



3 listeners



3 listeners

M2



5 listeners



4 listeners



3 listeners

M3



3 listeners



2 listeners



2 listeners

SID: Post-Hoc Work

F1



7 listeners



4 listeners



2 listeners

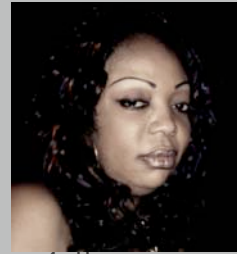
F2



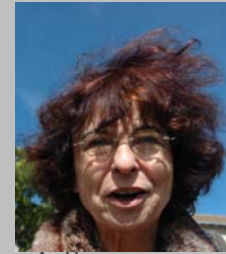
5 listeners



4 listeners



4 listeners



4 listeners

F3



5 listeners



3 listeners



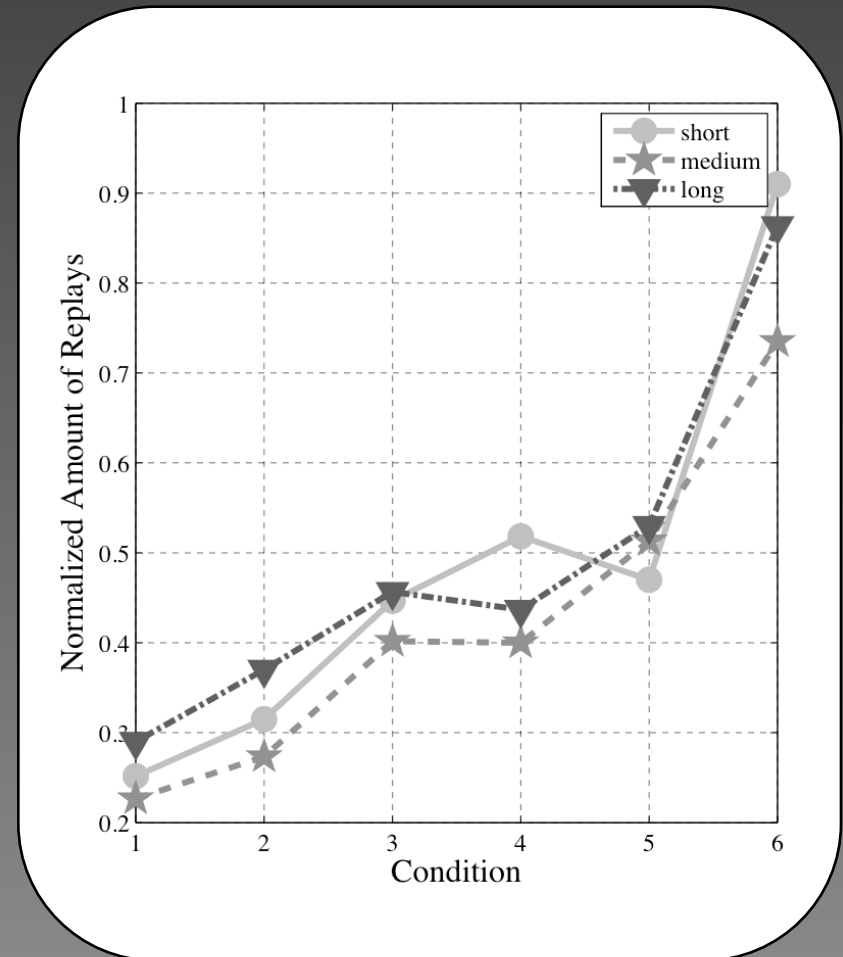
3 listeners



3 listeners

SID: Post-Hoc Work

- C1 replayed 20-30% of the time, on average
- C6 replayed 70-90% of the time, on average
- Number of replays goes up with difficulty
- Amount of prosodic information might have been a source of listener confusion



SID: Open Questions

- Consult with experts in psychology and neurology to design lab tests that more closely model real world situations
- Attempt an experiment with better controlled recordings and familiar speakers

SID: In depth

- Paper covering results published in the conference proceedings of MESAQIN 2008:
<http://wireless.feld.cvut.cz/mesaqin/contributions.html>