# The CHAINS Speech Corpus: CHAracterizing INdividual Speakers

Fred Cummins, Marco Grimaldi, Thomas Leonard, Juraj Simko
School of Computer Science and Informatics
University College Dublin
*Fred.Cummins@ucd.ie, Marco.Grimaldi@ucd.ie*
*Thomas.Leonard@ucdconnect.ie, Juraj.Simko@ucd.ie*

2006-06-16

## 1 Introduction

The CHAINS corpus is a speech database expressly designed to help to characterize speakers as individuals. The corpus contains the recordings of 36 speakers obtained in two different sessions with a time separation of about two months.

The design goal of the corpus is to provide a range of speaking styles and voice modifications for speakers sharing the same accent. Other existing corpora, in particular the CSLU Speaker Identification Corpus, the TIMIT corpus, and the IViE corpus served as referents in the selection of material. This design decision was made to ensure that methods designed and evaluated on the CHAINS corpus might be directly testable on these other corpora, which were recorded using quite different dialects and channel characteristics.

## 2 Distribution

The CHAINS corpus will initially be made available at little or no charge to the research community. It is released under a Creative Commons license which allows free use for both commercial and non-commercial research. Redistribution is permitted as long as the redistributed corpus remains under the same license, and that copyright of University College Dublin is acknowledged. Details are available at the project website: `http://chains.ucd.ie`.

## 3 Recording Methods and Speaking Styles

Two recording sessions provided speech in six different speaking styles. The first recording session (SOLO, SYNCHRONOUS, RETELL) was carried out in a professional recording studio (December 2005) and speakers were recorded in a sound-attenuated booth. The recordings in the released corpus were done using a Neumann U87 condenser microphone. Additional tracks using other microphones (near and far-field) were recorded and may be made available upon request.

The second recording session (RSI, WHISPER, FAST) (from March 2006 to May 2006) was carried out in a quiet office environment, using an AKG C420 headset condenser microphone.

Across the two sessions, each speaker provided recordings in six different speaking styles:

- SOLO reading
- SYNCHRONOUS reading
- Spontaneous speech (indicated in the followings as RETELL condition)
- Repetitive Synchronous Imitation (hereafter RSI)
- WHISPERed speech reading
- FAST speech reading.

In two of the speaking conditions adopted, speakers modified their speech in a constrained fashion towards a known target; in the SYNCHRONOUS condition, the speech of the co-speaker served as a target, while in RSI, there was an explicit known static target. The presence of a known target which speakers aim to copy raises the bar in the discovery and design of procedures for automatic speaker identification, as the target speech provides a potentially highly confusing foil.

The WHISPER and FAST speech conditions are also well defined speaking styles which require substantial voice modification by the speaker.

**SOLO Reading** In this condition, subjects simply read a prepared text at a comfortable rate.

**SYNCHRONOUS Reading** In this method, two subjects read a prepared text in synchrony with one another, using a methodology described further in [2]. At the start of each reading, subjects were given a countdown by the experimenter, and were asked to commence reading when the count reached zero. Dysfluencies were caught and the sentence restarted with a new countdown. In the final corpus, any such dysfluencies have been omitted, leaving only the fluent readings.

> *Note: because dysfluencies have been omitted, the pauses in the resulting corpus are not to be taken as actual pauses for the purposes of research.*

**RETELL: Spontaneous Speech** After reading the Cinderella fable in the SOLO condition, subjects were asked to retell the story in their own words. No time limit was set and retold versions range from about 20 seconds to about 90 seconds in duration.

**RSI: Repetitive Synchronous Imitation** For this condition, only the Cinderella fable was used. It was divided into 19 individual phrases. Subjects listened to a repeating loop in which a single phrase was uttered, and after the second repetition, they joined in, repeating the phrase six times in synchrony with the target. They wore headphones through which a binaural mixture of their own voice and the target were heard. This methodology was originally developed as a pedagogical tool in teaching prosody, especially pitch accents, to learners of Swedish as a second language [3]. It ought to produce a close match in timing and intonation to the given target. The recording released is the penultimate repetition, together with the target (left and right channels). In the case of dysfluencies, a fluent production preceding the penultimate was used.

**WHISPER** Subjects read all texts in a whisper. Any involuntary switch to modal voicing was interpreted as a dysfluency and led to a restart of the phrase.

**FAST** Two recordings of one sentence were prepared. The first was read by the principal investigator at a comfortable rate. The second was read at a substantially greater rate of articulation. These were played to each subject to illustrate the degree of rate increase they should aim for, and subjects then read all texts at this self-controlled fast rate.

## 4 Corpus Texts

We recorded a selection of short fables and sentences. The selected four fables are familiar from many experimental studies, and include the first paragraph of the Rainbow Text, The Members of the Body Text, and the North Wind and the Sun. The longest fable is the version of Cinderella used, inter alia, in the IViE corpus. This latter text is the only text used in the RSI condition, and forms the basis for the spontaneous speech condition (RETELL), in which subjects provide an unscripted retelling of the fable. Otherwise, all texts were used in all conditions.

In order to provide good phonetic coverage, there are 33 individual sentences: nine selected from the CSLU Speaker Identification corpus, and 24 from the TIMIT corpus. In selecting sentences, those felt to be likely to induce speech errors were avoided, as were those which were judged to be over long or over short.

In the following we present the full text of all fables and sentences. A code uniquely identifying each text is provided in parentheses.

**(f01) The Cinderella Story**

(fs01) Once upon a time there was a girl called Cinderella.
(fs02) But everyone called her Cinders.
(fs03) Cinders lived with her mother and two stepsisters called Lily and Rosa.
(fs04) Lily and Rosa were very unfriendly and they were lazy girls.
(fs05) They spent all their time buying new clothes and going to parties.
(fs06) Poor Cinders had to wear all their old hand-me-downs!
(fs07) And she had to do the cleaning!
(fs08) One day, a royal messenger came to announce a ball.
(fs09) The ball would be held at the Royal Palace, in honour of the Queen's only son, Prince William.
(fs10) Lily and Rosa thought this was divine.
(fs11) Prince William was gorgeous, and he was looking for a bride!
(fs12) They dreamed of wedding bells!
(fs13) When the evening of the ball arrived, Cinders had to help her sisters get ready.
(fs14) They were in a bad mood.
(fs15) They'd wanted to buy some new gowns, but their mother said that they had enough gowns.
(fs16) So they started shouting at Cinders.
(fs17) "Find my jewels!" yelled one.

(fs18) "Find my hat!" howled the other.
(fs19) They wanted hairbrushes, hairpins and hair spray.

**(f02) The Rainbow Text (First paragraph only)**

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.

**(f03) The North Wind**

The North Wind and the Sun were arguing one day about which of them was stronger, when a traveller came along wrapped up in an overcoat. They agreed that the one who could make the traveller take his coat off would be considered stronger than the other one. Then the North Wind blew as hard as he could, but the harder he blew, the tighter the traveller wrapped his coat around him; and at last the North Wind gave up trying. Then the Sun began to shine hot, and right away the traveler took his coat off. And so the North Wind had to admit that the Sun was stronger than he was.

**(f04) The Members of the Body**

One fine day it occurred to the Members of the Body that they were doing all the work and the Belly was having all the food. So they held a meeting, and after a long discussion, they decided to go on strike until the Belly agreed to do its proper share of the work. So for a day or two, the Hands refused to pick up food, the Mouth refused to receive it, and the Teeth had no work to do. But after a few days the Members began to find that they themselves were not in a very active condition: the Hands could hardly move, and the Mouth was all parched and dry, while the Legs were unable to support the rest. And so they realised that even the Belly in its dull quiet way was doing necessary work for the Body, and that all must work together or the Body will go to pieces.

**CSLU's Phonetically Rich Phrases - 9 sentences in total**

(s01) If it doesn't matter who wins, why do we keep score?
(s02) Stop each car if it's little.
(s03) Play in the street up ahead.
(s04) A fifth wheel caught speeding.
(s05) It's been about two years since Davey kept shotguns.
(s06) Charlie, did you think to measure the tree?
(s07) Tina got cued to make a quicker escape.
(s08) Joe books very few judges.

(s09) Here I was in Miami and Illinois.

**TIMIT Sentences - 24 sentences in total**

(s10) She had your dark suit in greasy wash water all year.
(s11) Don't ask me to carry an oily rag like that.
(s12) A boring novel is a superb sleeping pill.
(s13) Call an ambulance for medical assistance.
(s14) We saw eight tiny icicles below our roof.
(s15) Each untimely income loss coincided with the breakdown of a heating system part.
(s16) Jeff thought you argued in favor of a centrifuge purchase.
(s17) The sermon emphasized the need for affirmative action.
(s18) Kindergarten children decorate their classrooms for all holidays.
(s19) Cory and Trish played tag with beach balls for hours.
(s20) The frightened child was gently subdued by his big brother.
(s21) The tooth fairy forgot to come when Roger's tooth fell out.
(s22) Alice's ability to work without supervision is noteworthy.
(s23) Special task forces rescue hostages from kidnappers.
(s24) If Carol comes tomorrow, have her arrange for a meeting at two.
(s25) Military personnel are expected to obey government orders.
(s26) Laugh, dance, and sing if fortune smiles upon you.
(s27) The fish began to leap frantically on the surface of the small lake.
(s28) The easygoing zoologist relaxed throughout the voyage.
(s29) Brush fires are common in the dry underbrush of Nevada.
(s30) How much will it cost to do any necessary modernizing and redecorating?
(s31) Was she just naturally sloppy about everything but her physical appearance?
(s32) Is a relaxed home atmosphere enough to help her outgrow these traits?
(s33) The same shelter could be built into an embankment or below ground level.

# 5 Speakers

A total of 36 subjects participated to the collection of the corpus. We recorded the bulk of the corpus (28 speakers) within a single dialect, in order to raise the bar for forensic speaker identification. We also included a few out-of-dialect speakers (8) for comparison. The 36 speakers are divided as follows:

- 1 female from the United Kingdom
- 3 females from the USA
- 2 males from the United Kingdom
- 2 males from the USA
- 12 females from the Eastern part of Ireland (Dublin and adjacent counties)

- 16 males from the Eastern part of Ireland (Dublin and adjacent counties)

Participants were recruited through the university, and were paid for their participation. No participant had any known speech or hearing deficit. Table 1 provides the dialect of each participant, gender, and the speaker ID.

| ***SpeakerID*** | ***Gender*** | ***Dialect*** | ***SpeakerID*** | ***Gender*** | ***Dialect*** |
|---|---|---|---|---|---|
| frf01 | female | California - USA | frm01 | male | East Anglia - UK |
| frf02 | female | California - USA | frm02 | male | Bedfordshire - UK |
| frf03 | female | Indiana - USA | frm03 | male | Virginia - USA |
| frf04 | female | Midlands - UK | frm04 | male | New England - USA |
| irf01 | female | Co. Dublin - IE | irm01 | male | Co. Dublin - IE |
| irf02 | female | Co. Dublin - IE | irm02 | male | Co. Dublin - IE |
| irf03 | female | Co. Dublin - IE | irm03 | male | Co. Dublin - IE |
| irf04 | female | Co. Dublin - IE | irm04 | male | Co. Dublin - IE |
| irf05 | female | Co. Dublin - IE | irm05 | male | Co. Dublin - IE |
| irf06 | female | Co. Dublin - IE | irm06 | male | Co. Dublin - IE |
| irf07 | female | Co. Dublin - IE | irm07 | male | Co. Dublin - IE |
| irf08 | female | Co. Dublin - IE | irm08 | male | Co. Dublin - IE |
| irf09 | female | Co. Dublin - IE | irm09 | male | Co. Dublin - IE |
| irf10 | female | Co. Dublin - IE | irm10 | male | Co. Dublin - IE |
| irf11 | female | Co. Dublin - IE | irm11 | male | Co. Dublin - IE |
| irf12 | female | Co. Dublin - IE | irm12 | male | Co. Dublin - IE |
| | | | irm13 | male | Co. Dublin - IE |
| | | | irm14 | male | Co. Dublin - IE |
| | | | irm15 | male | Co. Dublin - IE |
| | | | irm16 | male | Co. Dublin - IE |

Table 1: Speakers ID, Gender and Dialect

In the SYNCHRONOUS condition two subjects read a prepared text in synchrony with one another. Speakers were paired as described in table 2.

| ***Non-Irish Females*** | ***Non-Irish Males*** | ***Irish Females*** | ***Irish Males*** |
|---|---|---|---|
| frf01-frf02 | frm01-frm02 | irf01-irf02 | irm01-irm02 |
| frf03-frf04 | frm03-frm04 | irf03-irf04 | irm03-irm04 |
| - | - | irf05-irf06 | irm05-irm06 |
| - | - | irf07-irf08 | irm07-irm08 |
| - | - | irf09-irf10 | irm09-irm10 |
| - | - | irf11-irf12 | irm11-irm12 |
| - | - | - | irm13-irm14 |
| - | - | - | irm15-irm16 |

Table 2: Speakers Coupling in the Synchronous Condition

In the RSI condition, each male speaker used the same target: the recordings of irm05 obtained in the SOLO condition. Similarly, each female speaker used the recordings of irf05 (obtained in the SOLO condition) as target.

## 6 Data Organization and Naming Schema

The sound files contained in the corpus are .WAV files sampled at 44100 Hz with a resolution of 16 bits. The content is organized according to the following hierarchy:

```
/data/[SpeakingStyle]/[SpeakerID]/[filename].wav
```

where:

```
[SpeakingStyle] := fast|retell|rsi|solo|sync|whsp
[SpeakerID]  := the string identifying the speaker
```

the label `sync` indicates SYNCHRONOUS recordings, while `whsp` indicates WHISPERed recordings. In order to maximise the readability of the `[filename]`, it is constructed as follows:

```
[filename] := [SpeakerID]_[TextID]_[SpeakingStyle]_[TargetID]
```

where:

```
[SpeakerID] := the string identifying the speaker
[TextID] := the code uniquely identifying the text
[SpeakingStyle] := fast|retell|rsi|solo|sync|whsp
[TargetID] := id of the target/co-speaker
```

The tag `TargetID` is used only in the relevant speaking styles: `sync` and `rsi`. As conditions SOLO, SYNCHRONOUS and RETELL were recorded in a studio, additional tracks recorded on different microphones are available if needed. These additional tracks are not being released in this initial version of the corpus, but may be obtained by applying to the authors.

**Example 1** in the file:

```
irf07_s27_whsp.wav
```

speaker `irf07` reads sentence `s27` in the `whsp` condition; this recording was carried out during the second recoding session.

**Example 2** in the file:

```
irm08_f01_fs01_rsi_irm05.wav
```

speaker `irm08` reads sentence `fs01` of fable `f01` in the `rsi` condition using speaker `irm05` (SOLO) as a target.

**Example 3** in the file:

```
frf03_f03_sync_frf04.wav
```

speaker `frf03` reads fable `f03` in the `sync` condition with co-speaker `frf04`.

# 7 Detailed Recording Notes

## 7.1 Post-processing

The following describes the post-processing stages adopted to convert the raw-recordings into the speech files included in the DVDs.

**SOLO** The raw recordings of the SOLO condition consisted of a mono file per speaker and microphone. Each mono file contained the whole read material (short fables and sentences). Speech samples of the short fables have been obtained by editing out dysfluencies or portions of speech not adhering to the above criteria. Pause durations are thus largely byproducts of the file trimming procedure, and do not represent actual pause behavior.

**SYNCHRONOUS** The raw recordings of the SYNCHRONOUS condition consisted of a stereo file per speaker pair (one channel per speaker). Each stereo file contained the whole read material (short fables and sentences), with one speaker on each channel. In order to preserve the time alignment between the two speakers' recordings, left and right channels have been edited in time aligned fashion before splitting into individual files. Again, pauses are often the result of editing operations, and are not as produced.

**WHISPER, FAST** The raw recordings of the WHISPER and FAST conditions consisted of a mono file per speaker. Each mono file contained the whole read material (short fables and sentences). Speech samples of the short fables have been obtained by editing out dysfluencies.

**RSI** The raw-recordings of the RSI recordings consisted of a stereo file per speaker. The left channel of each file contains the target phrase (`irm05_f01_solo` for male speakers and `irf05_f01_solo` for female speakers) as recorded during the SOLO session, while the right channel contains the actual utterance of the subject. The files were chopped to the same length, including approximately 150 ms of silence before the acoustic onset of the phrase and 150 ms after its offset.

**RETELL** No post-processing has been applied to the recordings obtained in this condition.

**Intensity normalization** The selected speech files recorded in WHISPER condition have been normalized to 60 dB. All the other speech files have been normalized to 70 dB. This procedure was carried out using *Praat* [1]

# References


[1] Paul Boersma and David Weenink. Praat: doing phonetics by computer. http://www.praat.org.

[2] Fred Cummins. Practice and performance in speech produced synchronously. *JP*, 31(2):139–148, 2003.

[3] Gabor Harrer. Effektivare språkundervisning med ny metod. http://www.diu.se/nr1-97/nr1-97.asp?artikel=rsi, 1997.