The CHAINS corpus: CHAracterizing INdividual Speakers

Fred Cummins, Marco Grimaldi,
Thomas Leonard and Juraj Simko

School of Computer Science and Informatics
University College Dublin, Dublin 4, Ireland
{fred.cummins,marco.grimaldi,thomas.leonard,juraj.simko}@ucd.ie

Abstract

We present a novel speech corpus collected with the primary aim of facilitating research in speaker identification. The corpus features approximately 36 speakers recorded under a variety of speaking conditions, allowing comparison of the same speaker across different well-defined speech styles. Speakers read a variety of texts alone, in synchrony with a dialect-matched co-speaker, in imitation of a dialect-matched co-speaker, in a whisper, and at a fast rate. There is also an unscripted spontaneous retelling of a read fable. The bulk of the speakers were speakers of Eastern Hiberno-English. The corpus will be made freely available for research purposes.

1. Introduction

This paper describes the CHAINS Recorded Speech Corpus. The CHAINS project (REF) is motivated by the dual goals of identifying those properties of a voice which are *unique* to an individual, and those which are potentially *shared* when speakers adopt a similar speaking style.

Idiosyncratic, or unique, characteristics which serve to identify an individual are of potential importance in biometric applications, including speaker verification and identification. Open set speaker identification, as arises, for example in forensic and security application scenarios, is still an unsolved problem, despite many recent advances (REFS). Difficulties arise as a result of many sources of variability which are still only partly understood. Intrinsic variability of speech is compounded by changes that arise as a result of intentional disguise [1], physiological variation (colds, stress, etc), and the adoption of distinct speaking styles, such as whispering, rapid speech, etc. If seasonal effects can be normalized [2] by recording speakers in well time-separated (different months) sessions, some other effects are unavoidable. As discussed in [3, p.10]: "it is a phonetic truism that no-one ever says the same thing in exactly the same way". On the other hand, *deliberate* variability of the human voice is a key issue in a (forensic) speaker identification scenario. In practice, criminals often *disguise* their voice with the obvious goal to making identification very hard. As discussed by Rodman et al. [1], within the voice disguise area there is extreme richness and variety. Falsetto, creaky voice, whispering and variation of speech tempo are few examples of possible disguise that affect the phonetics of the human voice in very many different ways.

In fact, speech provides a rich signal which is widely believed to contain sufficient information to uniquely identify a person. This information may be roughly grouped into two separate classes: static information arising from, e.g. the details of vocal tract anatomy, and dynamic information which is available through the act of speaking. It is thus important to understand how the human ability to intentionally modify the voice through disguise or alteration in speaking style impacts the speech which is produced. By the same token, we want to know which characteristics of an individual's voice remain invariant despite such large-scale modification. Much of phonetics has been characterized by the search for *linguistic* invariants, and so large speaker databases have been used. However, with this approach, the idiosyncrasies of individuals has been relegated to noise. A different approach is required in studying those characteristics which distinguish an individual from all others.

A second issue in speaker identification that affects the usability of a speech corpus is the availability of recordings done in different environments and with different equipment. It is widely recognized that equipment changes significantly affect the performance of automatic speaker recognition systems. As reported in [4], the influential "KING" corpus included data which was recorded before and after a minor equipment change, resulting in a "Great Divide" within the corpus [5]. Results based on data collected exclusively before or after the equipment change have proven to be much "better" than results which included data from both before and after. Moreover, recordings done in a single environment may not be optimal for speaker identification. As discussed by Campbell et al. [6], the TIMIT corpus [7] and its derivatives (e.g. FFMTIMIT, NTIMIT, ...) are poorly suited for evaluating speaker recognition systems. This fact is due to the pristine conditions of the corpus collection: no intersession variability and wideband recordings in a sound booth.

The CHAINS corpus is the result of an effort to provide a speech database expressly designed to characterize speakers as individuals. The corpus contains the recordings of approximately 36 speakers (section 2) obtained in two different sessions with a time separation of about two months.. The first recording session (section 2.3) was carried out in a professional recording studio; speakers were recorded in a sound-proof booth. The second recording session is currently being carried out in a quiet office environment. Across the two sessions, each speaker provides recordings in six different speaking styles: SOLO reading, SYNCHRONOUS reading, spontaneous speech (indicated in the followings as RETELL condition), repetitive synchronous imitation (RSI), WHISPERED speech reading and FAST RATE speech reading. They also provide some unscripted speech in a retelling of one fable. Details of each recording session and each speaking style are provided in section 2.2 and 2.3.

The second goal of the corpus pertains to the characterization of speaking styles. In several of the conditions we used, speakers modify their speech in a constrained fashion to-

wards a known target; e.g. in the SYNCHRONOUS condition (section 2.2.1), the speech of the co-speaker serves as a target, while in RSI (section 2.2.4), there is an explicit known static target. The WHISPERED and FAST RATE speech conditions are also well-defined speaking styles which nonetheless require substantial voice modification by the speaker. The study of variability in speech requires the identification of a variety of well-defined speaking styles, beyond the laboratory speech which characterizes much of experimental phonetics (REF: Brink/Harnsberger). By studying intentional modification as a result of style, we hope to obtain a window, not just on invariant characteristics of a single speaker, but on shared characteristics of multiple speakers adopting a similar style. Mixdorff et al [8] demonstrated great differences between speakers in their implementations of several speech styles. This novel database will provide a rich resource for the study of inter and intra-speaker variability across speaking styles.

2. Corpus Structure

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, especially the CSLU Speaker Identification Corpus, the TIMIT corpus, and the IViE corpus served as referents in the selection of material. This design decision has been made to ensure that methods designed and evaluated on the CHAINS corpus, might be directly testable on those corpora, which were recorded using quite different dialects and channel characteristics.

We chose to record the bulk of the corpus within a single dialect, in order to raise the bar for forensic speaker identification [3, p. 97]. We also included a few out-of-dialect speakers for comparison. There are approximately 36 speakers in total. Of these, 28 (14 male, 14 female) are from the Eastern part of Ireland (Dublin and adjacent counties), thus providing a substantial amount of dialectal homogeneity in the body of the corpus. We call this dialect Eastern Hiberno-English. A further 8 (4 male, 4 female) are from the UK or USA. Participants were recruited through the University, and were paid for their participation. No participant had any known speech or hearing deficit.

In what follows, we describe the corpus texts, the speaking conditions and the recording session details.

2.1. Corpus Texts

We recorded both short fables and sentences. The 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 (section 5.3). The longest fable is the version of Cinderella used, *inter alia*, in the IViE corpus [9, 10]. 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. The full text of all fables and sentences used is provided in section 5.

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.

2.2. Speaking Conditions

There were six speaking conditions in total. In the SOLO condition, subjects read the materials at a comfortable pace. Dysfluencies were dealt with by re-reading the sentence within which they occurred. The SOLO condition serves as a baseline referent with respect to all the other conditions.

2.2.1. Synchronous Condition

In the SYNCHRONOUS condition, subjects were enjoined to speak in synchrony with a co-speaker. On the direction of the experimenter, a countdown was provided (3..2..1..) after which subjects started speaking together. The requirement of speaking in synchrony typically causes little problems to speakers, and it has been demonstrated that speech elicited in this fashion is relatively unmarked (REF: J. Phon paper). However it has been shown [11] that speech recorded in this condition exhibits a marked reduction in the inter-speaker varability for some variables associated with global timing and phrasing. Specifically, when speaking synchronously, speakers show marked agreement on the division of a text into phrases, pause placement and pause duration, all of which are highly variable across speakers when they read alone. A few speakers exhibited a little difficulty with this task, producing speech with idiosyncratic segmental prolongations at times.

One advantage of this condition in the present context, is that both speakers must aim at a common timing, and they typically match their intonation contours closely as well. This makes speaker identification hard (we hope), and provides a useful test bed for any candidate index to be used in speaker identification. In a speaker identifiaction scenario, a potential test to be passed is to see if the candidate index recognizes speech samples of speaker A obtained in SOLO and SYNCHRONOUS as more similar to each other than samples obtained in the synchronous condition from A and a co-speaker B. More formally, let X be a feature vector extracted from one reading, so that $X_{A_{\text{solo}}}$ is X taken from speaker A reading alone, and let '–' indicate some appropriate distance metric such as Euclidian distance or similar, then if the following relation holds, X is a likely candidate for inclusion in a speaker identification procedure:

$$|X_{A_{\text{synch}}} - X_{A_{\text{solo}}}| < |X_{A_{\text{synch}}} - X_{B_{\text{synch}}}| \qquad (1)$$

2.2.2. Whisper Condition

In the WHISPER condition, subjects read the entire set of texts in a whisper. Instructions were provided to ensure that subjects did not lapse into breathy voice, or even modal voicing: readings were redone where this occasionally happened. Whisper phonation is distinguished from other phonation types, glottal creak, breathy voice etc., by a constricted glottis resulting in turbulent airflow and "a characteristic hissing quality", as described by Laver in [12, p. 190]. The relevance of whispered speech to the field of speaker identification lies in the difficulties and challenges it presents for most traditional means of forensic speaker analysis. For example, the utility of fundamental frequency (F_0) as an indicator of speaker identity is called into question as it is very much a characteristic of voiced/modal speech. Whisper phonation also has the effect of reducing the "available information about vocal intensity, voice quality and, to a lesser degree, prosody or speech timing", as pointed out by Hollien in [13, p.49].

2.2.3. Fast Rate Condition

In the FAST RATE condition, subjects were played a short example of a speaker reading two novel sentences at a relatively fast rate, and were asked to attempt to read the texts at that rate. Increased speech rate is well known to introduce complex changes to the speech signal [14], and thus presents a particular challenge to speaker identification, where samples to be compared may be collected under very different real world conditions.

2.2.4. Repetitive Synchronous Imitation

In the RSI - Repetitive Synchronous Imitation - condition, one speaker per sex was selected from the SOLO readings to act as a target. Speakers selected were fluent readers with relatively unmarked accents within the pool of Eastern Hiberno-English speakers. That speaker's recording of the Cinderella fable was split up into 19 individual phrases, each of which had been read as a single intonational phrase. For each phrase, subjects listened to a repeating loop, in which the phrase was played 8 times in total, with a gap of 0.5 sec between phrases. After the first two instances, subjects joined in and spoke in synchrony with the recording. This technique was originally developed to teach Swedish prosody to learners of a second language [REF: Gabor], and has been found to rapidly produce a very close match to target in prosody (timing, stress patterns, intonation).

Although this degree of target approximation is unlikely to be employed in any real situation where a person tries to mimic another, it provides a limiting case. If a candidate variable has been found to identify a speaker, rather than the speaking condition itself, any variable which still picks out the speaker and not the condition here is certainly worthy of further testing on more substantial databases.

2.2.5. Retell Condition

Finally, in the RETELL condition, subjects who had recently read the Cinderella fable were asked to retell the content of the story in their own words. This method was used in the IViE corpus as well [9, 10] to elicit spontaneous speech, which nonetheless was assured of containing specific lexical items, and had a high probability of containing specific short phrases. It thus provides an opportunity to test methods for speaker identification on a radically different style, which still exhibits considerable overlap with the read data.

2.3. Corpus Recordings

The SOLO, SYNCHRONOUS and RETELL conditions were recorded in a professional recording studio. Each speaker sat in a sound treated booth, and, in the SYNCHRONOUS condition, could see the other speaker through a thick glass partition. Recordings were done using three microphones per subject. The principal track was recorded using a Neumann U87 Condenser microphone. Most users of the corpus will use these recordings. However, for those who might be interested in the challenges provided by alternative channels, we also recorded from a Neumann K184 condenser mic positioned above the subject's head, and from a B & K 4006 omnidirectional condenser microphone located to the rear of the subject. These additional tracks will be made available upon request.

The remaining conditions (RSI, WHISPER, FAST) were recorded in a quiet, but not sound-treated office environment using a Shure SM50 head-mounted microphone connected to a Marantz PMD 670 Compact Flash recorder.

All recordings will be made available as 16 bit PCM encoded WAV files with a sampling rate of 44.1 kHz.

3. Distribution and Releases

At the time of submission, the second set of CHAINS recordings is ongoing. We anticipate a final release of the corpus by May 2006. The corpus will be made freely available for research purposes. The principal release will comprise a set of DVDs including only the principal microphone recordings from the first recording session. Recordings from the other microphones will be provided to interested parties. We have no plan to release time aligned transcriptions together with the recordings, however the full texts of the sententces and the fables will be provided in the corpus documentation.

4. Acknowledgements

The CHAINS project is funded by Principal Investigator Grant 04/IN3/I568 to the first author. The authors gratefully acknowledge this support.

5. Appendix

The following sentences are used in all conditions except RSI and RETELL. They have all been used in other corpora, including the CSLU Speaker Identification Corpus and TIMIT.

5.1. CSLU's Phonetically Rich Phrases

1. If it doesn't matter who wins, why do we keep score?
2. Stop each car if it's little.
3. Play in the street up ahead.
4. A fifth wheel caught speeding.
5. It's been about two years since Davey kept shotguns.
6. Charlie, did you think to measure the tree?
7. Tina got cued to make a quicker escape.
8. Joe books very few judges.
9. Here i was in Miami and Illinois.

5.2. TIMIT Sentences

1. She had your dark suit in greasy wash water all year.
2. Don't ask me to carry an oily rag like that.
3. A boring novel is a superb sleeping pill.
4. Call an ambulance for medical assistance
5. We saw eight tiny icicles below our roof
6. Each untimely income loss coincided with the breakdown of a heating system part.
7. Jeff thought you argued in favor of a centrifuge purchase.
8. The sermon emphasized the need for affirmative action.
9. Kindergarten children decorate their classrooms for all holidays.
10. Cory and Trish played tag with beach balls for hours.
11. The frightened child was gently subdued by his big brother.
12. The tooth fairy forgot to come when Roger's tooth fell out.

13. Alice's ability to work without supervision is noteworthy.
14. Special task forces rescue hostages from kidnappers.
15. If Carol comes tomorrow, have her arrange for a meeting at two.
16. Military personnel are expected to obey government orders.
17. Laugh, dance, and sing if fortune smiles upon you.
18. The fish began to leap frantically on the surface of the small lake.
19. The easygoing zoologist relaxed throughout the voyage.
20. Brush fires are common in the dry underbrush of Nevada.
21. How much will it cost to do any necessary modernizing and redecorating?
22. Was she just naturally sloppy about everything but her physical appearance?
23. Is a relaxed home atmosphere enough to help her outgrow these traits?
24. The same shelter could be built into an embankment or below ground level.

5.3. Short Fables

The Cinderella Story: Once upon a time there was a girl called Cinderella. But everyone called her Cinders. Cinders lived with her mother and two stepsisters called Lily and Rosa. Lily and Rosa were very unfriendly and they were lazy girls. They spent all their time buying new clothes and going to parties. Poor Cinders had to wear all their old hand-me-downs! And she had to do the cleaning! One day, a royal messenger came to announce a ball. The ball would be held at the Royal Palace, in honour of the Queen's only son, Prince William. Lily and Rosa thought this was divine. Prince William was gorgeous, and he was looking for a bride! They dreamed of wedding bells! When the evening of the ball arrived, Cinders had to help her sisters get ready. They were in a bad mood. They'd wanted to buy some new gowns, but their mother said that they had enough gowns. So they started shouting at Cinders. 'Find my jewels!' yelled one. 'Find my hat!' howled the other. They wanted hairbrushes, hairpins and hair spray.

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.

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.

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.

6. References

[1] R. Rodman and M. Powell, "Computer recognition of speakers who disguise their voice," in *Proceedings of the International Conference on Signal Processing Applications and Technology (ICSPAT2000)*, Dallas, Texas, USA, October 2000.

[2] R. Cole, M. Noel, and V. Noel, "The cslu speaker recognition corpus," 1998.

[3] P. Rose, *Forensic Speaker Identification*. Taylor and Francis, 2002.

[4] J. Godfrey, D. Graff, and A. Martin, "Public databases for speaker recognition and verification," in *Proceedings of the ESCA Workshop Automatic Speaker Recognition, Identification, Verification*, Martigny, Switzerland, April 1994.

[5] King Corpus. [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95S22

[6] J. P. Campbell and D. A. Reynolds, "Corpora for the evaluation of speaker recognition systems," in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing*, vol. 2, 1999, pp. 829–832.

[7] TIMIT Corpus. [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

[8] H. Mixdorff, H. R. Pfitzinger, and K. Grauwinkel, "Towards objective measures for comparing speaking styles," in *Proc SPECOM*, Patras, Greece, 2005, pp. 131–134.

[9] IViE Corpus. [Online]. Available: http://www.phon.ox.ac.uk/~esther/ivyweb/

[10] E. Grabe and B. Post, "Intonational variation in the british isles," in *Corpus Linguistics: Readings in a Widening Discipline*, S. G. and M. D., Eds. Continuum International Publishing Group, 2004.

[11] F. Cummins, "Synchronization among speakers reduces macroscopic temporal variability," in *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, 2004, pp. 304–309.

[12] J. Laver, *Principles of Phonetics*. Cambridge: Cambridge University Press, 1994.

[13] H. Hollien, *Forensic Voice Identification*. Academic Press, 2001.

[14] J. Trouvain, "Tempo variation in speech production," Ph.D. dissertation, Institut für Phonetik, Universität des Saarlandes, 2004, published as Forschungsbericht Nr. 8.