A NEW SPEECH DATABASE FOR WITHIN- AND BETWEEN-SPEAKER VARIABILITY

Patricia Keating, Jody Kreiman, Abeer Alwan

Departments of Linguistics, Head & Neck Surgery, Electrical Engineering, UCLA, Los Angeles CA
keating@humnet.ucla.edu, jkreiman@ucla.edu, alwan@ee.ucla.edu

ABSTRACT

We describe a new speech corpus designed to sample variability in speaking within individual speakers and across a large number of speakers. The public version of the database comprises audio recordings of 201 speakers performing 12 brief speech tasks over three recording sessions. Most of the tasks are unscripted, and include a phone call and pet-directed speech. The recordings have been orthographically transcribed, and dictionary broad transcriptions have been force-aligned. The database can be downloaded for free.

Keywords: voice quality, speech corpus, phonetic variation, speaker characteristics

1. INTRODUCTION

Voices vary both within and between speakers. Any one speaker's voice varies constantly, so that no speaker's voice is fixed, but rather exhibits a distribution of qualities. But at the same time, each speaker sounds different from other speakers. It is often assumed that within-speaker variability is markedly and reliably less than between-speaker variability [22, 11, 8], and to the extent this is the case, automatic speaker discrimination, recognition, or verification by machine is easier. How can this assumption be tested? A database of speech samples that includes extensive variation both within and between speakers is needed. Here we describe a new speech database collected for that purpose.

2. REQUIREMENTS

Our first requirement was that the speech in the database be in English, both for our own purposes and to make the database useful to other researchers interested in American voices. To provide between-speaker variability, the database should include speech from many (at least 100) speakers, both men and women. To provide within-speaker variability, the database should include a variety of speech tasks that each speaker performs; speakers should not only read, or not only give an interview or monolog. Sturim et al. [20], a detailed consideration of issues in database design for speaker verification research, also suggest at least two recording sessions per speaker, each session at least 30 seconds; intrinsic variability of speaker state, such as emotion, plus variability of recording conditions. We followed these recommendations except for the last one; we wanted consistently high-quality audio for all recordings. Other recording conditions can be simulated post-hoc.

While we did not want to focus on read speech, we did want some materials to be spoken by all of the speakers. It was also desirable to have some material repeated verbatim by each individual speaker. Together these conditions allow text-dependent analysis (in which the phonetic content is controlled) both within and across speakers.

None of the existing, public English databases that we are aware of offers the desired combination of a large number of speakers (male and female), multiple recording sessions per speaker, multiple speech tasks per speaker, repeated text, and high quality audio. Some well-known English databases have too few speakers (BU Radio News [14], Buckeye Corpus [15], RedDots Challenge [9]), widely-variable audio quality across the speakers (Speakers in the Wild [10]), too few speech tasks (TIMIT [4], Switchboard [5], GMU Speech Accent Archive [21], Canadian Maritimes [7]), just one recording session (Intonation Variation in English [13]), or just one sex (Dynamic Variability in Speech [12]).

In short, having decided that there is a need for a new database to fill this gap, we designed and produced the UCLA Speaker Variability Database.

3. DESIGN AND METHODS

3.1. Speakers

The goal was to record 100 men and 100 women, drawn from the UCLA student population. Reflecting the demographics of this group, we recorded speakers with diverse language backgrounds: monolingual native speakers of American English; bilinguals who are L1 speakers of American English; L2 speakers of English (most started English before age 10); and a few native speakers of other dialects of English. The speaker breakdowns in the public version of the database are given in Section 4 below.

On the other hand, our speakers are similar in age and current university community, such that many

other potential dimensions of speaker differentiation have not been incorporated into the database design. While providing between-speaker variability is a key goal of the database, too much speaker variability is a minus, not a plus, because the voices are then too easily distinguished. We believe that two dimensions of high variability – here, sex and language background – is about right.

3.2. Recording conditions

3.2.1. Equipment

Recordings were made in a sound-attenuated booth using a Bruel & Kjaer microphone suspended from a baseball cap worn by the speaker. Recordings were direct-to-disk at 22 kHz sampling rate using PCQuirerX and its hardware.

3.2.2. Recording sessions

Recordings were made in three sessions on separate days. The speaker sat in front of a computer screen on which prompts were presented via Matlab. For some tasks, printed sheets were available on the table in front of the speaker, providing suggestions about what to talk about (see specific tasks below). For tasks in which participants were asked to speak for 30 seconds, a countdown-bar was shown on the screen.

3.3. Speech tasks

Each recording session comprised four speech tasks, two repeated across sessions and two unique to each session.

3.3.1. Vowels task

Each recording session began with the speaker producing the vowel /ɑ/ (in isolation) three times, 1-2 seconds per vowel. Thus a total of 9 tokens were produced across the three sessions. This task was included to allow for comparisons with an existing large library of pathological voice samples from our campus voice clinic.

3.3.2. Read sentences task

In each recording session, the next task after the vowels task was reading five Harvard sentences [3] twice each in a random order. Thus 10 sentence tokens were recorded in each session, with a total of 30 tokens across the three sessions. This task was included to provide repeated text both within and across recording sessions and across speakers. Read speech is usually clear speech, though some of our participants were not fluent readers. This task is the only reading task in the database.

3.3.3. Instructions task

The first unscripted speech task came after the sentences task in the first recording session. Participants were prompted to give 30 seconds of instructions about how to do something, as if to the research assistant visible outside the sound booth. A list of possible topics was provided, but participants could talk about anything they wanted to. This task generally resulted in fairly clear speech.

3.3.4. Conversation report tasks

In each recording session, one task involved participants recalling and recounting a conversation they had had in recent days. Participants were asked to report the dialog of both speakers in the remembered conversation, in a "s/he said – I said" (or: "she was like" – "I went") style. Speech comprising "reported speech" or "constructed dialog" can show significant voice quality variation, because it can re-create multiple voices, and/or because the speaker conveys information about stance or evaluation [16, 17, 6]. However, not all of our speakers fully complied with instructions to repeat the original dialog, and instead simply described the conversation's content.

The conversations recounted in the three recording sessions had different prompts, intended to encourage different affects. In the first session, the conversation was one that the speaker had viewed as unimportant, not exciting, not upsetting: "neutral". In the second session, it was one that had made the speaker really happy. In the third session, it was one that had really annoyed the speaker. For each prompt, a list of example topics was provided, to help jog the participants' memories of suitable conversations they had had. Most speakers produced some degree of affect difference between the "happy" and "annoyed" tasks.

3.3.5. Phone call task

The last task in the second recording session was a phone call: participants used their own cell phones to call an unidentified friend or relative and talk for at least two minutes. Only our participants' side of the conversation was recorded, not via the phone signal, but directly from the speaker's mouth, as in all our other speech tasks. This task generally produced the most casual speech in the database.

3.3.6. Pet video task

The last task in the third recording session was included to elicit a very different speaking style from all the other tasks: pet-directed speech. Participants

chose to watch an approximately 2-minute video of either cute kittens or cute puppies (their choice). They were asked to talk aloud to the pets as they watched the video.

While infant-directed speech is more-studied than pet-directed speech, we thought that undergraduate students are more likely to have produced pet-directed than infant-directed speech. Pet-directed speech is known to show the exaggerated prosody of infant-directed speech, but with less affect [1]. Many but not all of our participants did produce pet-directed speech in this task (for example, speakers' mean F0s are much higher in this task than in the Sentences task); however, some speakers produced relatively little speech. The phonetic content of the utterances in this task tended to be very limited, with many participants saying things like "Oh, so cute!" repeatedly.

Table 1 at the end of the paper summarizes the tasks in the database, and gives an estimate of how much speech each task yielded.

3.4. Transcription and alignment

3.4.1. Orthographic transcription

All audio recordings are accompanied by transcriptions in the form of Praat textgrids. The Vowels speech task was manually segmented and labelled. Otherwise, the corpus is fully orthographically transcribed at the sentence or utterance level (intervals delimited by breaks). For example, one interval in the orthographic textgrid could contain the transcription "all right so this weekend i was at cal my best friend goes to cal so i was staying with her which was really awesome".

3.4.2. Automatic forced-alignment

These orthographic transcriptions were used as input to a forced-alignment program to derive a new Praat textgrid containing word-level orthographic transcriptions plus segment-level phoneme transcriptions. Most of these were obtained using the University of Pennsylvania P2FA forced-aligner FAVE [23, 19], but some were obtained using the Dartmouth DARLA program [18]. Both of these aligners produce phoneme strings in ARPABET from dictionary look-up of orthographic words in the CMU Pronouncing Dictionary [2], which gives lexical stress, and has multiple pronunciation entries for some words.

For example, the orthographic word tier could contain the aligned transcription "CAL", while the corresponding phone tier could contain the aligned ARPABET symbols "K", "AE1", "L".

Forced alignments are of course error-prone. While the total corpus is too big for manual correction of all alignments, all outputs have been visually examined and gross errors (e.g. transcriptions located in the wrong part of the audio file) have been corrected. Furthermore, we are in the process of manually checking and correcting alignments for the Read Sentences task.

4. RELEASE OF PUBLIC DATABASE

While the full database is available internally to the UCLA project team for our own research, not all speakers have consented to public release of their recordings. The public version of the database comprises 201 speakers (96 men, 105 women), as follows:

- 77 (35 men, 42 women) self-reported monolingual speakers of American English who sound like native speakers
- 47 (20 men, 27 women) self-reported bilingual speakers of American English plus some other language, who sound like native English speakers
- 77 (41 men, 36 women) self-reported bilinguals or L2 speakers who do not sound like native English speakers

This public version of the database will be available for free download in 2019.

5. ACKNOWLEDGMENTS

This work has been supported by NSF grants IIS-1450992 and IIS-1704167 to Abeer Alwan. We thank the student research assistants who helped record and prepare the database.

6. REFERENCES

[1] Burnham, D., Kitamura, C., Vollmer-Conna, U. 2002. What's New, Pussycat? On Talking to Babies and Animals. *Science* 296 (5572), 1435.

[2] CMU dictionary http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[3] IEEE Subcommittee on Subjective Measurements. 1969. IEEE Recommended Practices for Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics* 17 (297), 227–246.

[4] Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S. 1993. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon technical report 93. Linguistic Data Consortium https://catalog.ldc.upenn.edu/LDC93S1.

[5] Godfrey, J.J., Holliman, E. 1997. Switchboard-1 Release 2. Linguistic Data Consortium https://catalog.ldc.upenn.edu/LDC97S62.

[6] Günther, S. 1999. Polyphony and the "layering of voices" in reported dialogues: An analysis of the use of prosodic devices in everyday reported speech. *Journal of Pragmatics* 31, 685-708.

[7] Kiefte, M., Nearey, T.M. 2017. Modeling consonant-context effects in a large database of spontaneous speech recordings. *J. Acoust. Soc. Am.* 142, 434-443.

[8] Kinnunen, T., Li, H. 2010. An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication* 52, 12–40.

[9] Lee, K.A., Larcher, A., Wang, G., Kenny, P., Brummer, N., Van Leeuwen, D., Aronowitz, H., Kockmann, M., Vaquero, C., Ma, B., Li, H., Stafylakis, T., Alam, J., Swart, A., Perez, J. 2015. The RedDots data collection for speaker recognition. *Proc. Interspeech-2015*, 2996-3000.

[10] McLaren, M., Ferrer, L., Castan, D., Lawson, A. 2016. The Speakers in the Wild (SITW) Speaker Recognition Database. *Proc. Interspeech-2016*, 818-822.

[11] Nolan, F. 1983. *The Phonetic Bases of Speaker Recognition*. Cambridge UK: Cambridge U. Press.

[12] Nolan, F., McDougall, K., De Jong, G., Hudson, T. 2009. The DyVIS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language & the Law* 16, 31-57.

[13] Nolan, F., Post, B. 2014. The IViE Corpus. In: Durand, J., Gut, U., Kristoffersen, G. (eds), *The Oxford Handbook of Corpus Phonology*. Oxford Handbooks in Linguistics. Oxford, UK: Oxford U. Press, 475-485.

[14] Ostendorf, M., Price, P.J., Shattuck-Hufnagel, S. 1996. The Boston University Radio News Corpus. LDC https://catalog.ldc.upenn.edu/LDC96S36.

[15] Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W, Hume, E., Fosler-Lussier, E. 2007. Buckeye Corpus of Conversational Speech (2nd release). Ohio State University https://buckeyecorpus.osu.edu/.

[16] Podesva, R. 2013. Gender and the social meaning of non-modal phonation types. *Proc. 37th Annual Meeting Berkeley Linguistics Society (2010)*, 427-448.

[17] Podesva, R., Callier, P. 2015. Voice quality and identity. *Annual Review of Applied Linguistics* 35, 173-194.

[18] Reddy, S. & Stanford, J. 2015. A web application for automated dialect analysis. *Proc. NAACL-HLT-2015*, 71-75.

[19] Rosenfelder, I., Fruehwald, J., Evanini, K., Yuan, J. 2011. FAVE (Forced Alignment and Vowel Extraction) Program Suite.

[20] Sturim, D.E., Torres-Carrasquillo, P.A., Campbell, J.P. 2016. Corpora for the evaluation of robust speaker recognition systems. *Proc. Interspeech-2016*, 2776-2780.

[21] Weinberger, S. 2013. Speech Accent Archive. George Mason University. http://accent.gmu.edu/.

[22] Wolf, J. 1972. Efficient acoustic parameters for speaker recognition. *J. Acoust. Soc. Am.* 51, 2044–2056.

[23] Yuan, J., Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *Proc. Acoustics 2008*, 5687-90.

Table 1: Summary of the speech tasks and recordings that comprise the new database.

	Session A	Session B	Session C	Amount of speech (total of 3 sessions)
isolated vowels	3 tokens (3 sec speech)	3 tokens (3 sec speech)	3 tokens (3 sec speech)	~10 sec
read sentences	10 sentences (~25 sec speech)	10 sentences (~25 sec speech)	10 sentences (~25 sec speech)	~75 sec
other speech task (unscripted)	instruction narrative (25-30 sec speech)	phonecall (60-120 sec speech)	talk to pet video (60-120 sec speech)	~145-270 sec
reported conversations (unscripted)	neutral (25-30 sec speech)	happy (25-30 sec speech)	annoyed (25-30 sec speech)	~75-90 sec
TOTAL	~75-90 sec per speaker	~110-180 sec per speaker	~110-180 sec per speaker	**~300-450 sec per speaker**