

Creating a speech corpus with semi-spontaneous, parallel conversational and clear speech

Tech Report: CSLU-11-003

Alexander Kain, John-Paul Hosom, Sarah Hargus Ferguson, Brian Bush

August 18, 2011

1 Desired properties of a clear speech corpus

Our goal is to collect a speech corpus for the purpose of studying intelligibility and acoustic differences between the conversational and clear speech styles. The ideal corpus has the following properties: (1) speech has been produced spontaneously as part of a communicative interaction, as opposed to having been read to an imagined interlocutor; (2) entire identical utterances, or large parts of utterances, are available in both conversational and clear speaking styles, also known as *parallel* recordings; and (3) utterances comprehensively and systematically cover the space of prosodic and phonetic features. We call the spontaneous (i. e. non-read) elicitation of speech with highly anticipated content (established through a given task) *semi-spontaneous*. We now discuss these desirable properties in more detail.

1.1 Spontaneous speech in a communication setting

Much research on clear speech has been performed using written text materials read by a talker in either conversational or clear styles. The two styles have been elicited via instructions: for example, when reading conversational speech, talkers have been instructed to speak in the way they normally communicate in their daily lives; when reading clear speech, talkers have been instructed to speak clearly, as they would when communicating with hearing-impaired listeners [5, 8].

Arguably, speech that has been produced by reading cannot be truly conversational. Moreover, recent research has shown that *read* clear speech is not the same, acoustic-phonetically, as *spontaneously* elicited clear speech [7]. Specifically, read clear speech is a more extreme version of spontaneous clear speech, featuring greater speaking rate decreases and greater expansion of F0 median, F0 range, and vowel space. (No intelligibility experiments have been reported on the effects of read clear speech versus spontaneous clear speech.) Finally, since most speech acts occur in a communicative setting, it follows that speech should be collected in such realistic settings, in place of (or in addition to) reading text materials in isolation [11].

We propose a communicative setting with a talker and a listener who has either normal hearing or a simulated hearing loss. The interlocutors are acoustically and optically isolated from each other and communicate via an intercom using headsets consisting of microphones and *supra-aural* (open-back) earcups. This setup allows simultaneous two-way transmission between the interlocutors (also known as full-duplex). It also allows talkers to naturally hear themselves while speaking (in contrast to circumaural earcups or insert earphones), and allows listening sound pressure levels to be controlled.

Hearing impairment can be simulated by digitally processing the audio signal from the talker to the listener. For example, previous work added multi-talker babble noise to the talker's speech or used (low-fidelity) vocoding techniques [6]. We propose to simulate hearing loss by implementing the effects of recruitment [10] and spectral masking [1], and optionally adding of various levels of background noise, such as multi-talker babble noise [8].

Phonetic contrast	Acoustic correlates	Word pair examples
front-back vowels	range of F2	feed/food
high-low vowels	range of F1	feet/fat
long-short vowels	duration of vowels	beat/bit
voiced-voiceless initial consonants	voice onset time	bat/pat
voiced-voiceless final consonants	duration of preceding vowel	bat/pat
alveolar-palatal fricatives	noise spectra	see/she

Table 1: Example phonetic contrasts used for systematically studying intelligibility.

1.2 Parallel recordings in conversational and clear speech styles

To study acoustic-phonetic and prosodic differences between the conversational and clear speech styles, it is desirable to keep the linguistic message constant, such that a written transcript of a conversational and a clear utterance would be identical. Previously, this has been achieved by having talkers read printed text materials [e. g. 8]. When eliciting spontaneous speech as described in Section 1.1, the task must be designed to elicit specific utterances in both styles.

1.3 Controlled content

The purpose of controlling the utterance content for even a single style is two-fold. First, to study all existing acoustic and prosodic phenomena, we need to collect examples that provide good coverage of the available space. In terms of phonetic coverage this means obtaining a good sampling of all phonemes in various contexts (diphone, triphone, or larger). Prosodic variety can be achieved by eliciting a number of different syntactic constructs, such as statements, questions, continuations, and lists.

Second, to increase the likelihood of recognition errors, it is desirable to use phonetically confusable words. We propose to use a word pair list designed to exhaustively test phonetic contrasts (both vowels and consonants) using word pairs [9]. A few examples can be found in Table 1.

2 Task

Recently, spontaneous laboratory speech has been elicited via the Diapix task, a collaborative spot-the-difference game involving a pair of pictures and a pair of participants [4, 2]. This task is highly successful in eliciting conversational speech, and the electronic versions of the pictures have been produced in a way permitting keywords to be changed to target desired content. However, since a new picture pair is needed for each new dialog, this task is unlikely to yield parallel content across dialogs. In addition, since interlocutors move on once a difference has been identified, each keyword is only produced once or twice in a given dialog.

To maximize the desired properties given in Section 1, we propose a new speech elicitation task [3] using a network-based computer application. The collection involves two subjects at a time, one of whom takes the role of the instructor/talker, and the other the role of the executor/listener. One or more *keywords* and an *instruction* are presented to the talker, while the listener is shown a set of words that include the keywords and alternate, highly confusable words. These words are sourced from the word pair list discussed in Section 1.3. For example, the keywords “ship” is associated with the confusable alternates “chip, tip, sip, shop”. Speech is expected to be mostly unidirectional, but the listener may ask questions to clarify or confirm instructions.

The talker is given a list with codified instructions (either using very brief statements, or entirely pictorial), designed to elicit specific utterances. Example instructions are:

1. Select A.
2. Select A, B and C.
3. Select any word. Did you select A?

4. Unselect A and B. (shifts emphasis, causing keywords to be less emphasized)
5. Select A and something that rhymes with it. Did you choose B or C? (rhyming words in question prosody)
6. Select A and B. (where both A and B are rhyming words, stated with statement prosody)

During test familiarization, subjects will use the same interface, with a short representative instruction list. Both subjects will take turns being the talker. After having taken turns, the procedure is repeated but this time the listener will experience simulated hearing loss.

The main experiment will be carried out in three conditions:

1. The normal hearing condition, yielding semi-spontaneous conversational speech
2. The simulated hearing loss condition, yielding semi-spontaneous clear speech (only the listener will be experiencing the hearing-loss condition)
3. Subjects reading their transcribed conversational speech utterances, yielding read clear speech

In each condition, subjects will take turns being the talker and the listener. Each condition will run the same sequence of tasks, but the order of the individual instructions will be randomized. The motivation for collecting *both* semi-spontaneous clear and *read* clear speech is that we can perform phonetic-acoustic analysis and novel intelligibility difference testing on these two forms of clear speech. The order of conditions 2 and 3 will be counter-balanced, with half of the talker pairs performing the tasks in each order. This will not only guard against possible ordering effects but allow these possible effects to be tested directly. In studies that found read clear speech to be an extreme form of spontaneous clear speech, read clear speech was always elicited *after* spontaneous clear speech [7].

It is likely that conditions 1 and 2 will not provide exactly parallel utterances for every utterance. Therefore, we will either use an algorithm that will explicitly take insertions and deletions into account by aligning phonemes via relevant representative features (for utterances with a small number of differences) [8], or we will simply discard non-parallel utterances (for utterances with a large number of differences). The read speech of condition 3 is expected to be highly parallel with that of condition 1.

3 Preliminary data collection experiment

To test the feasibility of the proposed data collection method, we recorded one male talker eliciting speech in response to a given instruction list. (See Appendix A for a complete list of keywords and their alternates.) The talker was communicating with a male listener, at first for 15 minutes in the normal hearing condition and then for 15 minutes in the simulated hearing loss condition. The latter was created by listening to an infinitely-looped 12 second segment of babble noise delivered via headphones, set at a loudness level such that intelligibility was approximately at 50%. We omitted the third condition (read speech).

Upon a cursory examination and comparison of words spoken in the two different conditions, we found that the acoustic-phonetic properties of keywords in the normal hearing condition corresponded to conversational speech, while those of keywords in the hearing-loss condition displayed properties of moderately clear speech. This is in agreement with previous work [7].

References

- [1] T. Baer and B. Moore. Effects of spectral smearing on the intelligibility of sentences in noise. *Journal of the Acoustical Society of America*, 94:1229–1241, 1993.
- [2] R. Baker and V. Hazan. DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, 2011.

- [3] M Beckman. A typology of spontaneous speech. in: Sagisaka y, n campbell and n higuchi, eds, computing prosody. computational models for processing spontaneous speech. In *Department of Scandinavian Languages, Lund University*, pages 7–26. Springer, 1997.
- [4] K. J. Van Engen, M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow. The Wildcat Corpus of native- and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*, 2010.
- [5] S. H. Ferguson. Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *Journal of the Acoustical Society of America*, 116(4):2365–2373, 2004.
- [6] J. Gryn timer, R. Baker, and V. Hazan. Clear speech strategies and speech perception in adverse listening conditions. *ICPhS*.
- [7] V. Hazan and R. Baker. Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style? 2010.
- [8] A. Kain, A. Amano-Kusumoto, and J.-P. Hosom. Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility. *Journal of the Acoustical Society of America*, 124(4):2308–2319, 10 2008.
- [9] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek. Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, 54:482–499, 11 1989.
- [10] B. Moore and B. Glasberg. Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in background of speech. *Journal of the Acoustical Society of America*, 94(4):2050–2062, 1993.
- [11] Yi Xu. In defense of lab speech. *Journal of Phonetics*, 38(3):329–336, 2010. URL <http://linkinghub.elsevier.com/retrieve/pii/S0095447010000318>.

A Keywords and their alternates

Keyword	Alternates
ache	cake, ape, ate, aches
add	had, at
air	hair, fair, chair, are
at	hat, add, fat
ate	hate, aid, fate, ache
bad	bat, bed, pad
bat	bad, pat
beat	bit, boot, meat
bill	dill, gill, mill
bit	beat, pit
blend	bend, lend, end
blow	low, bow, bloat
bunch	punch, bun, munch
cake	ache, take, cakes
cash	cat, gash, catch
chair	tear, air, share
cheer	tear, sheer
chop	top, chap, shop
coat	code, goat
dock	docks, mock, knock
dug	duck, tug, bug
ease	is, peas, cheese
eat	heat, sheet
fat	feet, hat, at
feed	feet, seed, fee, food
feet	fat, feed, sheet, heat, fit
fill	pill, hill, full
fork	cork, four, forks
geese	gas, guess, goose
had	pad, hid, add
hail	tail, ail, sail
hall	tall, ball, all
hand	pant, and, sand
harm	arm, farm, charm
hash	dash, ash, hatch
hat	pat, at, fat, that
hate	heat, ate
heat	hate, eat, feet
him	hem, ham, hum
hold	cold, old, fold

knew	know, gnaw, knee, know
knot	nod, nut, dot
know	knew, knew
leak	reek, league, lick, lee, luke, leaks
leap	reap, lip
lip	rip, leap, slip, lit
low	row, blow
meat	me, neat, meats, beat
much	mut, muck, mush
nice	night, knife, dice
pad	had, pat, bad
pat	hat, pit, pad, bat, pot
pit	pet, pat, bit, spit
rake	lake, ray, rakes
reap	leap, rip, weep
reed	lead, rid, weed
rip	lip, reap
rise	lies, eyes, wise
rock	lock, rocks, walk
row	low, owe, woe
see	she, tea, he, seed
seed	feed, see, seeds
sell	shell, tell, fell
sew	show, toe, foe
sheet	seat, feet, eat, shoot
ship	sip, tip, sheep, chip
shoot	suit, shot, sheet
side	sight, sigh, sign
sigh	shy, tie, thigh, side
sin	shin, tin, in
sink	pink, ink, sing
sip	ship, tip, zip, slip
slip	sleep, sip, lip
spit	pit, sit, it
steak	take, sake, snake
sticks	six, ticks, stick
take	steak, cake
tear	chair, cheer
tile	dial, pile, mile
tip	sip, ship
wax	lax, wack, racks
witch	wit, wish, rich
write	light, ride, white