

The American English SALA-II Data Collection

Peter A. Heeman

Center for Spoken Language Understanding
OGI School of Science & Engineering
Oregon Health & Science University
20000 NW Walker Rd., Beaverton OR 97006
heeman@cslu.ogi.edu

Abstract

We discuss the collection of the American English SALA-II speech corpus. We focus on how we designed the prompt sheets to ensure maximum variability and on our strategy for recruiting the required 4000 speakers. We also present results on the effectiveness of the phonetically rich sentence. This paper should benefit others who are interested in using this corpus, or who are planning to collect a speech corpus with a large number of speakers.

1. Introduction

The American English SALA-II corpus collection is part of an ongoing effort to create speech databases for training speech recognition systems (Moreno, 2002). The SALA-II collection is aimed at speech over cellular phones in North, Central and South America. The content and the validation criteria of SALA-II is similar to the EU-funded SpeechDat project. Each speaker is required to say 44 different items, including city names, company names, people's names, credit card numbers, dollar amounts, numbers, phone numbers, typical application words (e.g., stop, play), time phrases, and date phrases. To ensure adequate phonetic coverage, speakers also say 9 different phonetically rich sentences, and 4 phonetically rich words. In order to not bias how they would say a word, speakers read the prompts. A computer system on the other end of the line guides the user through the list of prompts. Also included are several spontaneous items, which consist of asking speakers their first name, the city they grew up in, the current time and date, and several yes-no questions.

The specifications also include distribution requirements. A certain number of calls have to be made (a) in a car, train, or bus, (b) in a public place, (c) next to a busy street, (d) in a car using a speakerphone carkit, and (e) in a quiet office. A certain number of calls have to be made by speakers from each of nine accent regions. The gender and age of the speakers also have to be balanced.

The SALA-II specifications require 4000 American English speakers. The corpus is owned by Loquendo, Microsoft, Natural Speech Communication, and Siemens, each owning a 1000 speaker subset. The complete SALA-II project also includes American Spanish, Canadian French, and Portuguese and Spanish from Latin America. Consortium members have access to each other's corpus, thus reducing the cost of corpus collection for each member. The corpora can also be purchased by non-consortium members from the European Language Resource Distribution Agency (ELDA).

In this paper, we focus on the American English SALA-II project, which was collected by the Center for Spoken Language Understanding at the Oregon Health & Science University (OHSU) under a contract from ELDA. We dis-

cuss how (a) we designed the prompt files and the prompt sheets, (b) our computer system and software for collecting the calls and transcribing them, and (c) our speaker recruitment strategy. An important requirement for training speech recognition is to have good phonetic coverage. Hence, we report the coverage of the phonetically rich sentences.

2. Corpus Collection work at CSLU

The Center for Spoken Language Understanding has been involved in corpus development for a number of years. The corpora we have developed include a telephone corpus (Cole et al., 1995), a speaker recognition corpus with recordings from each subject spanning over two years (Cole et al., 1998), a foreign-accent corpus, a twenty-two language corpus (Lander et al., 1995), and a children's speech corpus (Shobaki et al., 2000).¹ These corpora are available for free to academic institutions and CSLU center members. CSLU also does custom corpus work on a contract basis. We collected the American English SpeechDat-Car corpus through a contract from the ELDA (Heeman et al., 2001). OHSU can use the corpus for internal research purposes but does not have distribution rights. OHSU has the same rights for the American English SALA-II corpus.

3. Design of Prompts and Prompt Sheets

In order to participate, speakers must have a prompt sheet, with all of the prompts they are supposed to say. At the time of designing the prompts and prompt sheets, we wanted to leave as much flexibility in how we would recruit speakers. As some recruiting strategies can have a response rate as low as 1% (Lindberg et al., 1998), we decided we should over-generate prompt sheets. We decided to create 20,000 sheets, and to make each one different. Each sheet has a different order of the prompt types (credit card number, date, time, company name, etc), and uses a different items from each prompt type. We did this to ensure a lot of variability in the responses. If we had, for instance, just

¹The author also collected a corpus of human-human task-oriented dialogues (Heeman and Allen, 1995), distributed through the Linguistics Data Consortium.

made 4000 different sheets, with 5 copies of each sheet, we might have had some sheets being done a large number of times, while others not at all, which would have decreased the coverage.

The specifications required that some of the prompt items be drawn from a small list of alternatives. For instance, the credit card numbers had to come from a list of 150 and the company names from a list of 500. For other items, including telephone numbers, dates, spelling of artificial words, currency amounts, and numbers, the specifications did not have this restriction. For these, we generated 20,000 different items to ensure maximal variability.²

Each prompt sheet included nine phonetically rich sentences. The specifications allowed each sentence to occur a maximum of 10 times in the final corpus. The specifications required each phone (except rare ones) to occur at least 400 times in the final corpus and for hopefully each speaker to have said each phone. To guarantee that each sentence occur at most 10 times, we could have created a set of 18,000 sentences. However, this would have been very time consuming. Instead, we decided to use a feedback strategy. Towards the end of the collection, we would regenerate the sentences on all prompt sheets that had not been distributed yet. As will be explained in Section 5, we created a set of 4412 sentences.

Each prompt sheet also included four phonetically rich words. The specifications required that each phone (except rare ones) occur at least 400 times in the final corpus for this prompt type. Also, each word could not occur more than 5 times in the final corpus. We constructed a list of 5000 words, with the intention of using the same feedback approach for the phonetically rich words.

All information about the sheets was saved in an SQL database. One program decided the ordering of the prompts and which prompt items would be on each sheet. A separate program generated the actual prompt sheets, which was done by using the LaTeX text processing program.

After 3200 speakers were recorded, we examined the frequency of the recorded prompt items. We found that some of our phonetically rich sentences had not even occurred once in the 3200 prompt sheets that had been recorded. The same was true for the phonetically rich words. Hence, we chose a set of 800 sheets that had not been distributed and used the feedback strategy to regenerate these prompt sheets. We changed the phonetically rich words and sentences. We also changed the cities, company names, credit card numbers, pin codes, time and date phrases, PIN codes, and people's names to get a better balance. At this point in our speaker recruitment, we were primarily recruiting speakers directly (see Section 5). Hence, we did not need to worry about over-generating prompt sheets, as we could guarantee that we would use all of the 800 new sheets. Also, the sheets were generated in priority, with the lower numbered ones having the items that occurred the fewest number of times. Hence, we made sure we used the sheets in order.

²For dollar amounts, we ensured that 10% were amounts between 1 and 99 cents to ensure good coverage of smaller amounts. Hence, each amount less than one dollar occurred roughly 200 times each on the 20,000 prompt sheets.

4. Recording and Transcribing Software

The recording platform is a Windows machine, with a Dialogic board that works with the T1 lines in the United States. We used the CSLU speech toolkit to communicate with the T1 board (Sutton et al., 1998). We wrote a script that waits for the phone to ring, plays the initial instructions, asks the user to key in their prompt sheet number, gender, age, and service provider. It then goes through all of the prompts. To ensure speakers say the items the right way, the system prompts them with not only the prompt number, but with its type, e.g. "Prompt 5, credit card number." This is possible because the system knows which prompt sheet that the speaker is using, and so knows the order of the prompt types on their sheet. If the call is terminated part way through, the user can phone back, and after keying in their prompt sheet number, gender, and age, can resume where they left off. Our machine had six incoming phonelines, and so six copies of the script were running simultaneously.

The CSLU toolkit includes a tool for transcribing speech: SpeechView. We used this for transcribing the speech. To speed up transcription, transcribers started with what the speaker had been prompted. For numbers, money amounts, and credit card numbers, we wrote a simple natural language generator to expand the numbers into what we expected the user to say. For instance, "\$10.05" was expanded into "ten dollars and five cents." For company names, we manually converted each company name into an acceptable written form. For instance, for "AT&T", the transcribers started with "A T and T." For the phonetically rich sentences, we stripped out the punctuation and ensured that the case of the words conformed to our transcription conventions. Having the transcribers start with a reasonable transcription of what the person would say greatly speeded up transcription. In fact, part way through the project we improved the initial transcriptions of some of the items, for instance, by writing out money amounts, natural numbers, and centuries in dates, and by fixing up the transcriptions of company names and the case in the phonetically rich sentences. These simple changes increased transcriber output by 30%.

All information was stored in the SQL database running on a Linux machine. This includes all of the information about what items are on the prompt sheet, and the order of the prompt types. It also includes whether the sheets have been distributed, whether and when they have been recorded, and how far the speaker made is on the prompt sheet. It also includes the transcription, and verification results. Using a database made it easy to write scripts for checking the status of the collection.

5. Speaker Recruitment

Our initial strategy was to use non-profit organizations and have them recruit for us as a fund-raiser, receiving \$8-\$12 per completed call and \$13 for carkit calls. However, it was difficult to get organizations from across the country interested. Unsolicited e-mails were rarely answered. We did recruit 17 groups (11 of them through personal contacts), mainly school and church groups, and two choirs. However, the volume from them was disappointing. Three

	Gender		Age					Noise Environment				
	M	F	-15	16-30	31-45	46-60	61-	vehicle	public	street	car/kit	home
North Central	126	196	6	140	98	71	7	83	43	55	0	141
Inland North	262	238	3	380	52	59	6	32	150	119	0	199
Eastern New England	139	168	1	269	12	17	8	6	218	17	0	66
New York City	256	159	7	287	69	47	5	18	125	183	0	89
Western New England	191	196	11	263	64	40	9	8	222	43	1	113
North Midland	168	145	4	255	26	26	2	11	138	73	0	91
South Midland	368	254	7	475	64	70	6	25	263	142	0	192
South	244	311	76	178	211	87	3	106	66	240	0	143
West	254	422	20	216	236	188	16	313	10	51	160	142
Foreign	11	14	1	4	12	6	2	5	3	5	0	12
TOTAL	2019	2103	136	2467	844	611	64	607	1238	928	161	1188

Table 1: Distribution of Sessions

groups did not even get 20 speakers. A church group mailed out 1800 sheets to their members and only 22 people responded (this was mailed right before Christmas). Only seven groups recruited more than 100 speakers, and three of these were located in our local area. In all, the non-profits recruited 1363 speakers. The one bright spot was that they recruited 463 speakers between the ages of 31 and 45 and 351 between 46 and 60, and 430 speakers in the car, train and bus environments, sessions that we found difficult to do with direct recruiting.

The remaining 2760 speakers were recruited by us personally. We purchased six cell phones and went on four recruiting trips to other parts of the country. We mainly focused on recruiting college students, in public places, in a quiet office, and on the street. Subjects were paid \$5 each. Recruiters carried dialectic maps so as to properly categorize where subjects were from.

Table 1 gives the distribution of the sessions in the corpus. In all, there were 4122 sessions recorded by 4090 speakers. Twenty-two of the speakers, accounting for 25 sessions, had a foreign accent.³

6. Phonetically Rich Sentences

The specifications required that each speaker say 9 phonetically rich sentences, and that no sentence occur more than 10 times in the corpus. We created a set of 4412 sentences. We used the 1300 sentences from the Harvard corpus and the Timit corpus. Additional sentences were gathered from children’s stories, including “Pinocchio,” and “Beauty and the Beast.” Some of these sentences were modified to simplify their syntactic structure, and ensure that they were not too long (the optimal size being between 5 and 10 words). To ensure good phone coverage, we also constructed a number of sentences. Using the sentences we had already gathered, we determined which phones were rare. We then combed a phonetic dictionary for words that contained these rare phones, creating a list for each word. One of our staff, who has a degree in teaching English as a second language, constructed sentences that contained between two and four words from the different lists. We then

³One of our non-profits accounted for practically all of the foreign and duplicate speakers, which was contrary to the instructions we gave them.

did a simulation to ensure we would get a good distribution of phones for each speaker: we randomly created 4000 sets of nine sentences, such that each sentence occurred the same number of times. We modified the set of phonetically rich sentences until we reached good phone coverage.

Table 1 shows how often each sentence was successfully completed by our subjects.⁴ There were 145 sentence types that occurred 11 times (one more than allowed), 37 occurred 12 times, 18 occurred 13 times and 6 occurred 14 times, 1 occurred 15 times, 2 occurred 16 times, 3 occurred 17 times, 1 occurred 18 times and 1 occurred 25 times. This gives a total of 358 sentences that were over the allowed amount. Even if these sentences were thrown out, we would still have 35,967 allowable sentences, which would still give us a success rate of 99.9% on the required 9 sentences for 4000 speakers, exceeding the 95% required in the specifications.

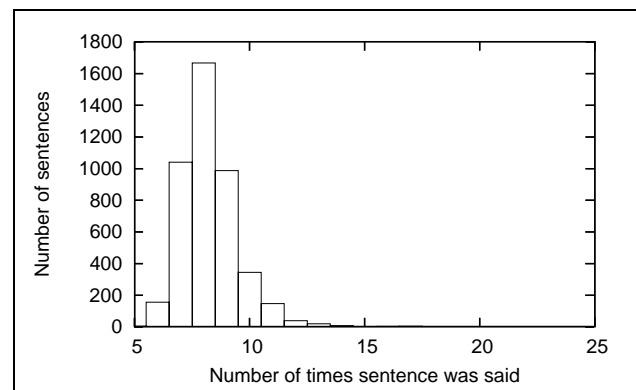


Figure 1: Frequency of sentences

We then measured the number of occurrences of each phone in the phonetically rich sentences. This was based on what the speaker said, rather than what the speaker was prompted to say. We excluded any words that were marked with as having a mispronunciation, cutoff, or cellphone distortion. Across all 36325 sentences collected, the rarest phones were [Z], [OI], [T] and [U] (in the SAMPA pho-

⁴By successfully completed, we mean that the audio file exists and the transcription consists of something more than noise symbols or ‘**’. It might, however, just consist of words marked with ‘*’ (mispronunciation), ‘~’ (cutoff) or ‘%’ (cellphone distortion).

Phone	Instances	Number of speakers
Z	3092	2074
OI	3925	2473
T	5987	3130
U	6166	3151
tS	7384	3367
dZ	7415	3364
aU	7421	3412
S	7772	3431
j	8171	3409
O	8543	3544
g	11288	3819
N	12030	3837

Table 2: Rare phones in phonetically rich sentences

netic alphabet). Table 2 gives the distribution of the rarer phones. We also give the number of different speaker sessions that contained each phone. The rarest phone [Z] occurred 3092 times, but in only 2074 different speaker sessions. As almost all of the rare phonemes occurred only once in a sentence, we could have improved these figures by taking more care in distributing the sentences that had the rare phonemes: ensure that each session contained at most one sentence for each of the rare phonemes.

We also measured the number of different phones that were said in each speaker session in the phonetically rich sentences. The results are shown in Figure 2. Of the 40 SAMPA phones, approximately 3500 of the 4122 speaker sessions contained at least 36 phones at least once. Although it is good sentences with rare phones so that as many different speakers say each phone as possible, it is also good to have more than one example of each phone per speaker. We found that in the phonetically rich sentences, 3500 sessions contained at least 32 phones at least twice, and 28 phones at least three times.

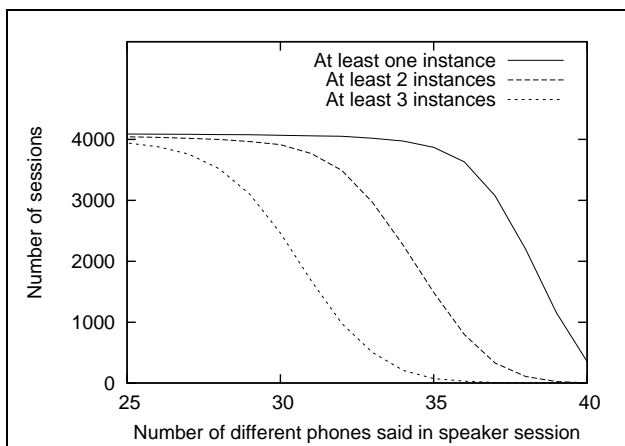


Figure 2: Phone coverage in phonetically rich sentences

Lastly, we measured the effect of the phonetically rich sentences on the overall phonetic richness of the corpus. Figure 3 gives the results. We see that 4060 of the sessions contained at least 36 different phones; excluding the sentences would result in only 3009 sessions having 36 different phones. For sessions containing at least 39 phones, without the sentences, the number drops from 2937 to 151.

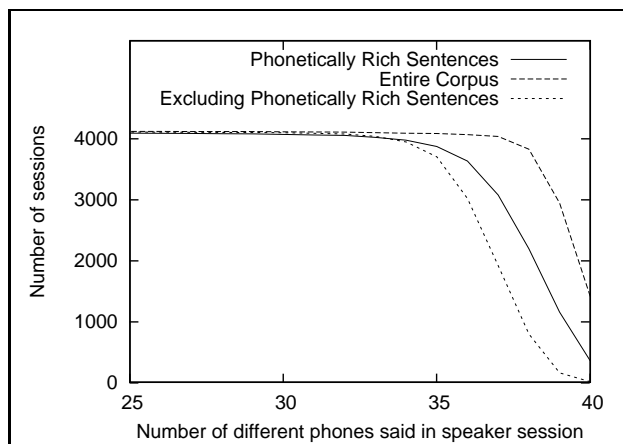


Figure 3: Effect of the sentences on overall phone coverage

Thus, the phonetically rich sentences served their purpose of ensuring good phone coverage.

7. Conclusion

This paper described the American-English SALA-II data collection. Speaker recruitment turned out to be the most challenging part of this project. Our initial plan to use non-profit organizations was not as successful as we had hoped it would be. Instead, we armed ourselves with six cellphones and went on recruiting trips across the country. The paper also showed that the phonetically rich sentences served their purpose of ensuring good phone coverage.

8. Acknowledgments

I would to thank everyone who worked on the project. In particular, I thank Hannah Hadfield who went on three recruiting trips and Pavel Chytil for setting up the database and scripts.

9. References

- Cole, R., M. Noel, T. Lander, and T. Durham, 1995. New telephone speech corpora at CSLU. In *Eurospeech*.
- Cole, R., M. Noel, and V. Noel, 1998. The CSLU speaker recognition corpus. In *ICSLP*.
- Heeman, P. and J. Allen, 1995. The Trains spoken dialog corpus. CD-ROM, Linguistics Data Consortium.
- Heeman, P., D. Cole, and A. Cronk, 2001. The U.S. SpeechDat-Car data collection. In *Eurospeech*.
- Lander, T., R. Cole, B. Oshika, and M. Noel, 1995. The OGI 22 language telephone speech corpus. In *Eurospeech*.
- Lindberg, B., R. Comeyne, C. Draxler, and F. Senia, 1998. Speaker recruitment methods and speaker coverage — experiences from a large multilingual speech database. In *ICSLP*.
- Moreno, A. The complete SALA II project specifications. Technical report, Universitat Politècnica de Catalunya. Version 1.50. Available at www.sala2.org.
- Shobaki, K., J. Hosom, and R. Cole, 2000. The OGI kids' speech recognizers and corpus. In *ICSLP*.
- Sutton, S. et al., 1998. Universal speech tools: the CSLU toolkit. In *ICSLP*.