

# **Eliciting natural speech from non-native users: collecting speech data for LVCSR**

**Laura Mayfield Tomokiyo and Susanne Burger**

Interactive Systems Laboratories  
Language Technologies Institute  
Carnegie Mellon University  
Pittsburgh, PA 15213 USA  
{laura,sburger}@cs.cmu.edu

Contribution Type: Paper

Special equipment needs: none

# Eliciting natural speech from non-native users: collecting speech data for LVCSR

## Abstract

*In this paper, we discuss the development of a database of recorded and transcribed read and spontaneous speech of semi-fluent, strongly-accented non-native speakers of English. While many speech applications work best with a recognizer that expects native-like usage, others could benefit from a speech recognition component that is forgiving of the sorts of errors that are not a barrier to communication, and in order to train such a recognizer a database of non-native speech is needed. We examine how collecting data from non-native speakers must necessarily differ from collection from native speakers, and describe research we did to develop an appropriate scenario, recording setup, and optimal surroundings during recording.*

## 1 Introduction

In this work, we try to address the question of what the best way to collect samples of natural, spontaneous speech from non-native speakers is. In particular, we are interested in the sort of data that is needed to train an LVCSR (large-vocabulary continuous speech recognition) system; (Byrne98) describes a similar collection project. There are many databases available for LVCSR training, most of which contain almost exclusively samples from native or near-native speakers of the language being collected. We feel that the techniques that have been successful in collecting native speech data may be less appropriate when the speakers are not completely comfortable in the language, and propose an alternate protocol, pointing out assumptions made by the conventional data collection methodology that other researchers may wish to consider when designing their own data collection methods.

The subject of data collection should be of interest to researchers in language learning technologies because many kinds of natural language processing systems require training on large corpora of exactly the same type of input that will be seen from the user. Especially

as corpus-based techniques become popular, a training database that accurately represents the speech patterns of the end user is necessary for developing NLP applications. Computer-aided language learning tools that have NLP components, then, will benefit greatly from both a good database of non-native language and an understanding of how collection of data from non-native speakers differs from collection from native speakers.

## 2 Target Data

Speech data that is commonly collected for speech recognition can be broken down into three major stylistic categories: read, planned/careful, and spontaneous. Although there are many types of spontaneous speech, the conversation and the query being two examples, spontaneous speech is subconscious; the speaker is not paying attention to the act of speaking itself. Obviously, this is not the case for semi-fluent non-native speakers. Although he might be in the same conversation as a native speaker speaking spontaneously, it is not clear that the non-native speaker's speech could be characterized as spontaneous. In fact, it is not clear that there is a distinction between planned and spontaneous speech at all.

With native speakers, spontaneous speech contains disfluencies, filler words, conversational devices, and other vocalizations that are rarely present in read speech. Read speech contains reading errors and stumbling that may not occur in spontaneous speech. These differences are likely to be more extreme with non-native speakers than with native speakers. However, word choice and syntax, which can vary greatly depending on speech style for native speakers, may vary much less for non-native speakers, who may not have developed a distinctive conversational style.

Our primary objective is to collect speech of the sort that would be used in a tourist information task, in which users can walk up to a terminal and ask questions about local sights,

restaurants, and the like. This sort of speech would be characterized as spontaneous for native speakers, and so with our discussion of the convergence of spontaneous and planned speech for non-native speakers in mind, we will call this spontaneous speech for non-natives as well. Recognizing that our speakers' speaking skills may limit their ability to speak spontaneously for as long as native speakers could, we have chosen to collect samples of read speech from each speaker as well. This will allow us to make a detailed analysis of the difference between read and spontaneous non-native speech.

### 3 Collection Methodology

Here we describe the physical environment for recording, discuss some pilot experiments, and present our final protocol for data collection from non-native speakers.

#### 3.1 Recording Setup

All recordings were taken by a DAT recorder; speakers wore a Sennheiser headset. Recordings were done in a small computer lab with some incidental noise but no excessive outside noise. On some occasions there were other people in the room when the recording was being done; this will be discussed further below. In non-interactive recordings, users were seated at a table with the instruction sheets, pen or pencil, and water.

#### 3.2 Pilot Experiments

We did two pilot experiments which greatly helped us to understand the needs of our speakers and how we could make them more comfortable, in turn improving the quality of our data. For these experiments, we recorded native speakers of Japanese.

##### 3.2.1 Pilot experiment one

In the first experiment, we drew from a human-machine collection task that we had had success with for native speakers in another domain. Speakers were provided with prompts such as the following:

- Ask how to get to the museum
- Find out where you can exchange money
- Ask where to get a ticket for the subway

Speakers came in on two different occasions and gave us feedback after both. The first time they came in, they were given the prompts in English. As we had predicted, they were strongly influenced in their word choice by the phrasings used in the prompts. The second time they came in, they were given the prompts in their native language. They felt that this task was much harder; they perceived it as a translation task in which they were expected to give a correct answer, whereas with the English prompts they were basically given the correct answer. Their productions, however, were more varied, different both from each other and from the original English prompt.

In addition to the prompt-based task, we had speakers read from a local travel guide, specifically about the university area so that the context would be somewhat familiar. We found that there were indeed reading errors of the type that would not occur in spontaneous speech (transcribers were instructed to spell the words they way they sounded to them). Examples will be provided in the final paper.

In addition to the prompt-based task, we had speakers read from a local travel guide, specifically about the university area so that the context would be somewhat familiar. We found that there were indeed reading errors of the type that would not occur in spontaneous speech. We observed that some speakers were stumbling over words that they obviously didn't know. We attempted to normalize for this by re-using utterances from dialogues that had been previously transcribed, hoping that they would be more likely to be familiar with words that other speakers of similar fluency had used. We still found that they had some difficulty in reading. Our speakers were native speakers of Japanese, though, which has a different writing system, and this would have some influence.

##### 3.2.2 Pilot experiment two

In the second pilot experiment, we attempted a wizard-of-oz collection using an interactive map; the speakers could ask for locations and routes to be highlighted in the map, and there was a text screen to which the wizard could send messages to answer speaker queries. Instead of a list of prompts, the speakers were given a sheet of paper listing some points of interest in the city, hotel names, some features that they could ask

about (business hours, location, etc.) and the dates that they would be in the city. Their task was to plan a weekend, finding hotels, restaurants, and things to do. Our thought was that perhaps speakers would speak more naturally in an information-gathering task, where they are actually trying to communicate instead of simply producing sentences.

Our general impression was that although the visual highlighting of the locations was a feature that the users enjoyed, and which helped them to become involved in the task, the utterances could not be characterized as more natural than those given in the prompted task. Examples will be given in the full paper. It took time to read and understand the responses from the wizard; also, speakers were aware that someone (the wizard) was listening in. Both of these factors were a source of self-consciousness.

### 3.3 Final Protocol

The final data collection protocol that we settled on has two parts. The first is a series of scenarios, in each of which a situation is described in L1 and a list is given, in bullet form, of things relevant to the situation that the speaker is to ask about. For instance, if the situation is a Pittsburgh Steelers game, the speakers would see the bullets **arena location, ticket price, seat availability, transportation, and game time**. The bullets are made as short as possible so that the speakers absorb them in a glance and can concentrate on formulating an original question instead of on translating a specific phrase or sentence.

The second part is a read task. There was no doubt left after the pilot experiments that the amount of patience speakers had with the prompted task was limited; after the novelty wore off speakers tired quickly. Although spontaneous data would be better than read data, read data would be better than no data, and speakers seemed willing to continue at least as long again reading as they had with the prompted task. We considered two types of material for the reading. We rejected a phonetically balanced text in favor of a version of the story of Snow White, which was familiar to all the speakers.

Finally, we ask speakers to read a selection of previously recorded and transcribed utterances in the same task, both by native speak-

ers and non-native speakers, randomly selected and with small modifications made to preserve anonymity. Our objective here was threefold: to quantify the difference between read dialogues and spontaneous dialogues; to quantify the difference between read dialogues and read prose; and to compare the performance of the end recognizer on native grammar with non-native pronunciation with performance on non-native grammar with non-native pronunciation.

We have recorded 23 speakers so far in the post-pilot phase of data collection, and all have expressed satisfaction with the setup.

## 4 Discussion

In the course of our pilot experiments, we became aware of a number of assumptions that are commonly made which do not necessarily hold for non-native speakers, and which it is important to address when designing a data collection protocol.

**There is little risk of alienating the community.** Local communities of non-native speakers are not always large, and if it is close knit, word can quickly spread if the task is too hard or embarrassing. Also, it is important to de-emphasize the fact that we are basically interested in speaker's speech because it is imperfect, or risk offending the community.

**The act of speaking is not difficult.** When recording native speakers speaking spontaneously, it is assumed that the the act of speaking does not in and of itself represent a major cognitive load for the speaker. This can be very untrue of non-native speakers, and we had several speakers ask to quit in the middle of the recording because they felt unable to continue. The researcher needs to make a decision about what to do in such a situation, and possibly prepare an alternate task.

**The task is not perceived as a test.** Again, when speaking spontaneously, few native speakers of nonstigmatized varieties of English would feel that they are being evaluated on the correctness of their speech. Many non-native speakers will feel tested, and as this can make them nervous and affect their speech, it is important to reassure them as far as possible that they are not being tested and that the data is being anonymized.

**The speaker knows what to say.** Most spontaneous collection tasks are chosen because they are tasks speakers can be expected to have done before and be comfortable with. Although a non-native speaker has probably made an airplane reservation in his native language before, it is entirely possible that he has never done so in the target language, and does not have a good idea of what he should say in that situation. If he were really planning to make an airplane reservation in the target language, he would probably think about what to say in advance and might even ask someone, which he may not have a chance to do during the data collection. This undermines the representativeness of the database.

## References

William Byrne et al. Is Automatic Speech Recognition Ready for Non-Native Speech? A Data Collection Effort and Initial Experiments in Modeling Conversational Hispanic English. In *Proceedings of Speech Technology in Language Learning (STiLL)*, 1998.