LINGUISTIC PROPERTIES OF NON-NATIVE SPEECH

Laura Mayfield Tomokiyo

Language Technologies Institute
Carnegie Mellon University
laura@cs.cmu.edu

ABSTRACT

As speech recognition systems improve in quality, they become attractive as components in applications which will be used by non-native speakers, both applications designed specifically for language learners and those intended for general use. Recognizer performance on speakers who are not fluent in the language they are speaking, however, is often poor. A number of factors contribute to recognition failure for non-native speakers; pronunciation, lexical choice, and syntactic structure are a few of the elements of speech that set native and non-native speakers apart. In this paper we examine the character of non-native speech, both spontaneous and read, describing how features that are known to be important in recognition system development compare with those of native speakers.

1. INTRODUCTION

Recognition performance for state-of-the-art systems is currently very high even for large-vocabulary connected-word tasks . It has been shown, however, that recognizer performance degrades significantly for non-fluent speakers ([1], e.g.). Systems that are designed specifically for non-native speakers, such as language tutoring systems, are often able to constrain the recognition component in a way that will bring recognition accuracy to an acceptable level without diminishing the effectiveness of the application ([3],[2], e.g.). Multi-lingual systems such as [6] in which new languages can be quickly added with little training data must answer difficult questions about how to map existing knowledge about speech to unseen languages, and can take advantage of regularities in native speech. The problem of adapting a full LVCSR system to non-native speech, however, requires an understanding of what makes non-native speech unique. In this paper we examine how non-native speech differs from native speech, and ask what effects idiosyncrasies of non-native speech have on recognition.

While native speakers do vary tremendously in their use of language, many similarities between native speakers are systematic enough to be well captured by statistically derived acoustic and language models, particularly when the speech is constrained in some way (such as speaker, style, or domain). Individual idiosyncrasies and preferences notwithstanding, native speakers have a command of their language that guides them to speak with a certain flow in order to convey meaning. Pauses and disfluencies, which occur often in natural speech, are not merely speech errors; they play an important role in discourse, allowing processing time for both speaker and listener, and marking the introduction of new or difficult information. Pronunciation is consistent; while accents may vary, they are distinguished primarily by systematic differences in the realizations of individual phones (which can be addressed with speaker adaptation).

Speakers who are not fluent in the language they are speaking face obstacles to language production that native speakers do not. They may have difficulty in articulating certain phonemes, or producing appropriate allophones in context. They may not know the right words, or have mastered the right syntax, to express what they want to say; they may know these things, but need time to formulate meaningful sentences. They may be worried that they will not be understood. Obstacles like these increase the cognitive load required to speak, causing speakers to stumble and pause. Non-native speakers are also limited by their exposure to, and grasp of, the language they are speaking, meaning that the distribution of both lexical items and disfluencies can be quite different from that of native speakers, and among different non-native speakers.

In this paper, we describe and contrast features that are found in native and non-native read and spontaneous speech. We examine such linguistic properties as speaking rate, lexical distribution, disfluency distribution, and perplexity in spoken English for native speakers of Japanese, Chinese, and English.

2. EXPERIMENTAL DATA

2.1. Speaker characteristics

We examined three sets of speakers, at three different levels of proficiency in English. All speakers were between the ages of 20 and 40 and had at least 2 years of college education (all college graduates or current college students).

Group 1 consisted of 12 native speakers of Japanese, all of whom had lived in the United States less than 1 year. All had studied English for a minimum of 8 years in Japan, but experienced difficulty making themselves understood. All self-reported a speaking confidence level of 3 or lower on a scale of 1 to 5 (5 being high confidence).

Group 2 consisted of 6 native speakers of Mandarin Chinese, all of whom had lived in the United States for approximately one year. All had studied English for a minimum of 10 years before coming to the United States. Speakers in group 2 reported speaking confidence levels of 3 or

4, but seldom experienced difficulty making themselves understood.

Group 3 consisted of 6 native speakers of English.

2.2. Task description

Recordings were done in a quiet room with a close-talking microphone. Speakers were allowed full control of the recording and were alone in the room while recording. Speakers were asked to do two tasks. The first was a prompted task designed to elicit spontaneous utterances in tourist information domain. Speakers were given a short description of a situation and several short prompts for questions, all in their native language. (The decision to prompt the speakers in their native language came out of our observation that when the prompts were given in English, speakers depended heavily on the words used in the prompt, whereas with native-language prompts, speakers came up with unique expressions which may better represent what would be seen in a real-world situation). This prompted task is described more completely in [5]. The second task was a read task, in which speakers read from two different texts. The first text consisted of transcriptions of utterances produced by both native and non-native speakers in earlier recordings. The transcriptions were cleaned of nonlexical items. The second text was a restricted-vocabulary, phonetically-enhanced version of the story of Snow White, a story which was familiar to all speakers.

Data was fully anonymized and the anonymization process was clearly explained to speakers.

3. SPEAKING RATE AND PAUSE DISTRIBUTION

An LVCSR system that is trained on native speech may expect very specific behavior at word boundaries. In fluent native speech, coarticulation across word boundaries can be very pronounced, and the presence or absence of certain allophonic alternations can be semantically meaningful (pronouncing *the one that he sent me* as [ðəwənðəthisɛʔmi] instead of [ðəwənðərisɛʔmi] can emphasize the word *he*, for example). Excessive pausing between words in an utterance, then, may be a factor in poor recognizer performance.

Speaking rate and pause distribution statistics for non-native speakers are shown in table 1. While it appears that there is not a significant difference in speaking rate between the conversational and read speech for the Japanese speakers, there is a marked difference in the average pause duration: speakers are pausing longer and more often between words in the read task.

4. LEXICAL DISTRIBUTION

Although non-native speakers of the proficiency level we are examining do not have the range of vocabulary and expression available to them that native speakers do, it is not clear that their speech, either individually or in the aggregate, could be described as more *restricted* than that of native speakers. In the context of a certain task, native speakers often rely on standard words and phrases, whereas non-native speakers, perhaps performing the task for the first time, may each come up with a unique way to ask the same question. For example, when prompted to ask about dress, most native speakers responded with "what should I wear," while non-native speakers were more creative with their queries:

Do we need to wear the formal dress or we can wear the casual one?
What kind of clothes do I have to wear for there?
In what kind of dresses should I go there?
Should I oh should I go formal with formal style?
What should I wear to go there?

A comparison of the vocabulary growth rates of the non-native responses to prompts in the tourist information domain with a native database of similar size and content shows similar behavior, although the non-native vocabulary growth rate is slightly higher (see Fig. 1). The use of contracted forms is higher in the native speech; *I am* and *I would*, the most common contractable expressions in the native database, were contracted in 43% and 69% of eligible instances respectively, while *I would* did not appear at all in the non-native database and *I am* appeared only once (and was contracted). Conversely, the most common contractable forms in the non-native database, *what is* and *where is* (contracted in 30% and 13% of cases respectively), did not appear frequently in the native database, in which *what is* was contracted 2 out of 9 eligible times, and *where is* did not appear at all. The difference in distribution of these common expressions can probably be attributed to a fundamental difference in the question format used by the two groups of speakers; native speakers made heavy use of embedded questions such as *Can you tell me where the museum is?* where non-native speakers favored shorter ones like *Where is the museum?* which require subject-verb inversion and therefore provide opportunities for contraction.

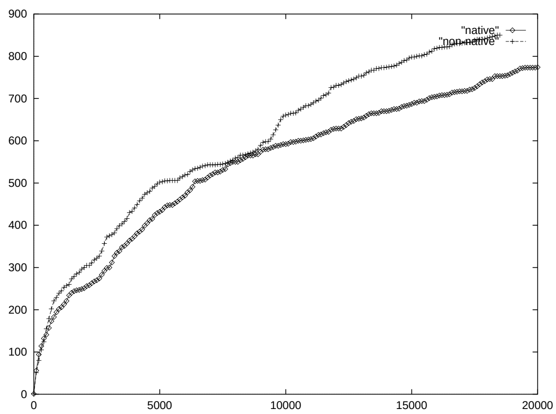

Figure 1: Vocabulary growth of native and non-native speakers in spontaneous tourist information queries. The number of unique word types is shown as a function of the number of word tokens in the corpus.

5. DISFLUENCIES

It has been observed that native speech contains many instances of abandoned words, stutters, restarts, filler words,

	word rate		silence rate		phone duration		pause duration	
speaker	prompt	story	prompt	story	prompt	story	prompt	story
Japanese	2.68	2.72	0.14	0.40	0.14	0.11	0.10	0.19
Chinese	2.90	2.50	0.15	0.41	0.12	0.14	0.10	0.12
Native	4.43	3.93	0.08	0.20	0.08	0.09	0.09	0.12

Table 1: Speaking rate and pause distribution statistics for non-native speakers. The word rate is reported in terms of words per second. The silence rate is a silence-to-word ratio. Average phone duration and pause duration are measured in seconds.

and other disfluencies, some of which occur systematically enough to warrant incorporation in the language model ([7],e.g.). Disfluencies often occur when the speaker is searching for the right word or expression, or is pronouncing a word that is difficult to articulate; such situations arose more often for the non-native speakers in our study than for the native speakers, and examination of their disfluencies shows a high incidence of both incomplete words and filler words.

5.1. Disfluencies in spontaneous speech

In general, the non-native speech in our study contained more disfluencies than the native speech, although there was a much greater variance between speakers in the two non-native groups. While native speakers were fairly similar in their use of filler words, some non-native speakers did not make use of filler words at all, and others relied heavily on them. The behavior of the Chinese speakers was much more similar to that of the native speakers than that of the Japanese speakers was with respect to disfluencies. In the Japanese group, the speakers that had the highest percentages of partial words actually did not use any filler words. Additionally, some speakers used filler words from their native language, although primarily between queries to the system as they were formulating their next utterance.

Table 2 shows the average percentages of stumbles (word fragments and abandoned words) and filler words such as *um* and *uh* for the different speaker groups.[1]

	% of stumbles		% of filler words	
speaker group	prompt	story	prompt	story
Japanese	1.46	2.48	4.37	0.25
Chinese	0.83	0.99	1.46	1.31
Native	0.53	0.14	0.94	0.04

Table 2: Average percentage of stumbles and filler words for each speaker group in spontaneous speech.

Another feature that can be used to compare disfluencies in different speech types (native vs. non-native, or spontaneous vs. read) is retrace behavior. It was reported in [4] that English and Swedish speakers show similar patterns in the number of words they "rewind" after an interruption. We found similar behavior in terms of average retrace length (2.4 words for non-native speakers vs. 2.25 for native speakers), but the incidence of retrace events was much higher for the non-native speakers. Table 3 compares

[1] Percentages are calculated for each speaker using the formula *# of {stumbles or filler words} / total word tokens* and then averaged.

	retrace rate		repeat rate	
speaker group	prompt	story	prompt	story
Japanese	0.90	0.43	0.28	0.48
Chinese	0.22	0.20	0.06	0.04
Native	0.13	0.20	0.08	0.00

Table 3: Retrace and repeated word rates. Individual rates are calculated for each speaker as a percent of the total number of word tokens and then averaged.

retrace behavior, showing retrace and repeated word rates averaged over all speakers. (In these calculations, we distinguished retrace events from single-word repetitions. For example, the repeated *so* in the utterance *so so so if I turn right on Beacon Street* is used as a filler, and is counted as a repeated word. In the utterance *okay so* /um/ `pause` *so I'll go to Harvard first i guess*, the speaker is momentarily distracted from speech production and uses the repetition of the word *so* to smooth the return; this is an example of a single-word retrace event.) In some cases, particularly with longer retraces, speakers do not repeat exactly what they said before the interruption. Sometimes this is a conscious repair, and sometimes not, and it can be difficult to judge after the fact. We have therefore combined exact retraces and repaired retraces in our calculations. The ratio of exact to repaired retraces was nearly identical for both native and non-native speakers (5:3).

5.2. Disfluencies in read speech

The load required of speakers to formulate meaningful sentences is lifted when they are reading aloud and not speaking spontaneously. However, reading errors, which are not an issue in spontaneous speech, may be introduced. Our data shows quite different behavior for the Japanese and Chinese speaker groups. For the Japanese group, while the incidence of filler words decreases for non-native speakers in read speech, the number of stumbles increases substantially. The Chinese group shows a similar trend, but only a slight one. A breakdown is shown in Table 2.

The words which speakers most commonly stumbled on were long or phonetically complex ones which they had probably had little occasion to pronounce before, such as *bitterly*, *dwarves*, and *stepmother*. Several speakers confused *he* and *she* when reading, sometimes correcting themselves and sometimes not. Substitution errors, in which the speaker reads a word that is different from the one that is on the page, were not common in our data.

It should be noted that our speakers had all had many years of formal education in English but little practice speaking it conversationally. A speaker with similar conversational ability but less experience reading may make significantly more reading errors. It was suggested by several speakers that the lower disfluency rates of Chinese speakers might be attributed to the practice common among Chinese university students of reading aloud from texts as a strategy for learning English, something that the Japanese speakers did not report doing frequently in their studies.

6. PERPLEXITY

It is difficult to make a statement about the grammaticality of the non-native speech in our study. Certainly, there were many times that speakers used an incorrect tense or article, and these errors were flagged during transcription. Ungrammatical events, however, are not limited to non-native speech, and it is difficult to quantify correctness in a useful way. A feature of non-native speech that can be quantified is its predictability, both inherent and with respect to a language model trained on native speech.

The language model we used for our comparisons was trained on a broad corpus of native data which included both read and conversational speech, the latter comprising primarily conversations between a traveler and a travel agent or information booth agent. Perplexities are shown in table 4 (OOV rates for all groups were under 0.5%).

Speaker group	Perplexity	Trigram hit rate
Japanese	66.5	55.8 %
Chinese	74.4	52.9 %
Native	102.6	48.6 %

Table 4: Perplexities and trigram hit rates of native and non-native test corpora measured with respect to a broad native language model

Both non-native speaker groups show low perplexities when compared with the native speakers, with similar trigram hit rates. The trigrams that the native and non-native speakers use frequently, however, are quite different. Both non-native groups show a preference for queries formed around the word *can*, for example, which were relatively rare in the native corpus. Common trigrams are shown in table 5.

Japanese	Chinese	Native
can i get	the name of	i need to
do you know	can i go	you tell me
how can i	can i get	i'd like to

Table 5: Most common trigrams for the different speaker groups

7. SUMMARY

Non-native speech differs from native speech in measurable ways and in elements of speech that are known to affect recognizer performance. The overall speaking rate for non-native speakers was 2/3 that of native speakers. The silence rate, or the ratio of silence elements to words, is double that of native speakers. This suggests that non-native speakers are pausing more between words that would undergo cross-word coarticulation in native speech, which may mean that contextual modeling based on native speech is inappropriate for non-native speech. Average phoneme length for non-native speakers is approximately 1.5 times that of native speakers for both spontaneous and read speech, while the average silence duration is similar for spontaneous speech but much longer in native read speech than non-native read speech. This may indicate that native speakers are pausing longer at semantically meaningful points in the utterance, while non-native speakers are pausing more often between words but because of difficulty in reading the text aloud, not to support the content.

There were many more disfluencies of all types in the non-native speech samples. Perplexity was lower than native perplexity for both non-native groups, and vocabulary growth rates were similar, but trigram distribution was very different. This may mean that although the phrases the non-native speakers are using are common, they are not the ones native speakers would choose to express the same idea. This would be an encouraging result from the point of view of recognition, in which long-term dependencies and semantic content are less important than collocational distribution, but may be evidence that natural language understanding of non-native speech will be a challenging problem.

8. REFERENCES

[1] 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.

[2] Jonathan Dalby, Diane Kewley-Port, and Roy Sillings. Language-Specific Pronunciation Training Using the HearSay System. In *Proceedings of Speech Technology in Language Learning (STiLL)*, 1998.

[3] Farzad Ehsani, Jared Bernstein, Amir Najimi, and Ognjen Todic. SUBARASHII: Japanese Interactive Spoken Language Education. In *Proceedings of Eurospeech*, 1997.

[4] Robert Eklund and Elizabeth Shriberg. Crosslinguistic Disfluency Modeling: A Comparitive Analysis of Swedish and American English Human-Human and Human-Machine Dialogs,. In *Proceedings of ICSLP*, 1998.

[5] Laura Mayfield Tomokiyo and Susanne Burger. Eliciting Natural Speech from Non-Native users: Collecting Speech Data for LVCSR. In *Proceedings of the ACL-IALL joint workshop in Computer-Mediated Language Assessment and Evaluation in Natural Language Processing*, 1999.

[6] Tanja Schultz and Alex Waibel. Adaptation of pronunciation dictionaries for recognition of unseen languages. In *Speech and Communication*, St. Petersburg, Russia, October, 1998.

[7] Elizabeth Shriberg and Andreas Stolcke. Word Predictability after Hesitations,. In *Proceedings of ICSLP*, 1996.