Two Protocols Comparing Human and Machine Phonetic Recognition Performance in Conversational Speech

Wade Shen¹, Joseph Olive², and Douglas Jones¹

¹ MIT/Lincoln Laboratory, 244 Wood St., Lexington, MA 02420
² Defense Advanced Research Projects Agency, 3701 Fairfax Dr., Arlington, VA 22203
{swade,daj}@ll.mit.edu, jpo@darpa.mil

Abstract

This paper describes two experimental protocols for direct comparison of human and machine phonetic discrimination performance in continuous speech. These protocols attempt to isolate phonetic discrimination while eliminating for language and segmentation biases. Results of two human experiments are described including comparisons with automatic phonetic recognition baselines. Our experiments suggest that in conversational telephone speech, human performance on these tasks exceeds that of machines by 15%. Furthermore, in a related controlled language model experiment, human subjects were better able to correctly predict words in conversational speech by 45%.

Index Terms: phonetic discrimination, speech perception, phonetic recognition

1. Introduction

Speech recognition technologies have made significant advances in the last 40 years. During that time a number of studies have been conducted that compare the ability of machines and humans to recognize speech. Some of these studies have focused on end-to-end comparisons of the recognition process, i.e. how do humans and machines compare in terms of the word errors they make when listening to a speech utterance [1][2][3][4]. Other studies have attempted to dissect the performance of human subjects on controlled phonetic discrimination tasks [5][7][8] and word prediction/ASR correction tasks [9].

Prior studies that have focused on phonetic discrimination have used controlled phonetic contexts (e.g. VCV or CVC tasks, DRTs, etc.). Often the purpose of these studies is to estimate the relative degradation associated with vocoder processing in which a fixed baseline of human intelligibility is needed, but an absolute measure of human performance is not required. As such the input speech is often presented to subjects with a segmentation bias (when compared to continuous conversational speech) – subjects know that they are listening to particular classes of segments in fixed contexts. In many of these studies the speech used is read and controlled; as a consequence the difficulties of conversational speech (e.g. reduction, approximate articulation, etc.) are often absent. As a result, these protocols may produce an inflated estimate of human performance when recognizing spontaneous conversational speech.

In this paper we present two protocols for the direct comparison of human and machine phonetic recognition. These protocols attempt to isolate phonetic discrimination from language and segmentation biases using natural continuous conversational speech. In sections 2 and 3, we describe each of the protocols and controls that we used. Section 4 describes the data and experimental facility used for both experiments. In section 5, we present results comparing machine and human performance on each of these tasks.

2. Phonetic Transcription of Foreign Language Speech

Our goal in this protocol is to design a task that allows human subjects to make phonetic discrimination decisions without providing segmentation or language-specific phonotactic biases. In [10][11], Greenberg et al showed that trained phoneticians disagreed at rates greater than 25% in their native language. Their study of interannnotator agreement suggests that even with proper training, detailed annotation guidelines, and a native language phonotactic model, human subjects have difficulty making fine-grained phonetic decisions in real conversational speech.

Although that study was not intended to measure human phonetic discrimination capability, it can serve as a correlate to an upper bound on human performance (as subjects made offline decisions and the task did not control for language biases). In our experiments, we modify the phonetic transcription task control for lexical and phonotactic biases by 1) selecting two languages L₁ and L₂ with a similar phonetic inventory and 2) ask monolingual speakers of L₁ to phonetically transcribe L₂. Assuming that both languages share a phonetic inventory, this task forces subjects to make phonetic decisions without a priori segmentation or lexical constraints. Subjects do this task using a natural, orthographic coding scheme eliminating the need for special training.

For this study, we chose data from conversational Japanese and Spanish speech (L₂) and asked native speakers of Italian (L₁) to transcribe this data. The choice of languages was designed to maximize phonetic overlap, and minimize cross-language phonotactic mismatches. For this study we define phonotactic mismatches as systematic cross-language biases that lead to errors when subjects of one language attempt to transcribe data from another. Examples of these phenomena for Japanese speakers transcribing Italian speech include certain complex consonant clusters (e.g. sförza, scussi), and implausible coda consonants (e.g. res.ta.to, lec.tu.ra). Because of this, we did not attempt to use Japanese subjects to transcribe Italian speech. Instead, all transcribers used for this experiment were native speakers of Italian without knowledge of Japanese. In order to minimize subject workload we presented subjects with short segments of continuous, conversational speech data and we allowed subjects to use native orthography during transcription. This was possible for subjects due to the near 1-1 phoneme grapheme mapping in Italian. These constraints were arrived...
at in earlier pilot experiments conducted at MIT with English native subjects.

Subjects were then asked to transcribe speech drawn from the OGI stories corpus [12] and their performance on this task was measured in terms of Phone Error Rate (PER) which is computed as shown in eqn. 1.

\[
PER = \frac{N_{sub} + N_{del} + N_{ins}}{N_{ref}}
\]  

(1)

The performance of human subjects is then compared against that of an automatic phone recognition system. For this, we trained an HMM-based context-independent phonetic recognizer using Japanese speech. The details of this system are described in section 4.

3. Word Transcription of Ambiguous Regions

In pilot studies of the phonetic transcription protocol described in section 2, many subjects reported difficulty in doing phonetic transcription tasks due to the intense listening and concentration requirements. As a result, we sought a different experiment design that makes use of an easier word transcription task.

For these experiments we extended a protocol developed by Vasilescu et al [13]. We asked subjects to transcribe seven-word-long speech segments extracted from conversational telephone speech and we measured their word accuracy for the center (4th) word. These segments have the special property that the 4th word could be replaced by a near-homophone and the lexical bias from the left and right contexts is neutral (i.e. subjects don’t prefer one alternative vs. another). We define near-homophone as either 1) a minimal pair, or 2) an exact homophone. We find these segments using the procedure shown below:

1) Scan transcribed corpus for words with near-homophones.
2) If a word \( w \) has near-homophones \( n_h \) ... \( n_h \), we replace \( w \) in its seven-word context with each of \( n_h \) ... \( n_h \). The resulting \( N \) seven-word sequences are then scored by a 4-gram language model. An example of this is shown below:

\[
\begin{align*}
\text{w}_1 \text{ orig: } & \text{ know that Bob had a house in} \\
\text{w}_2 \text{ alt: } & \text{ know that Bob has a house in} \\
\end{align*}
\]

3) Near-homophone alternatives with LM scores close to that of the original transcript are retained for human judgment. We define close via the equation below.

\[
\log \frac{P(w_{i-3}w_{i-2}w_{i-1}n_h,w_{i+1}w_{i+2}w_{i+3})}{P(w_{i-3}w_{i-2}w_{i-1}w_{i+1}w_{i+2}w_{i+3})} \geq -0.25
\]

4) Regions extracted in step 3 are then presented to human subjects. Subjects perform a word selection task choosing between possible alternates of the original sequences. For this task subjects are given only the transcript without audio. Regions that exhibit near-chance performance are labeled as ambiguous with respect to their lexical bias and presented to subjects for transcription with audio.

5) Subjects transcribe ambiguous regions at the word level listening to force-aligned audio corresponding to the original seven-word sequence. Their performance is assessed for the 4th word.

As the resulting regions are ambiguous with respect to human judgment, we can use these regions to assess the error rate of subjects based on acoustic information alone (without language biases).

We compare human subject performance on these segments against the performance of machines in two ways: 1) we use an ASR system to force-align near-homophone alternatives and choose the maximum likelihood hypothesis and 2) we perform full recognition on the seven-word sequence. In both cases, the results are compared to human performance on the transcription of the same audio.

4. Experimental Setup

4.1. Foreign Language Phone Transcription

For this experiment we used 15 native Italian speakers visiting MIT as students or researchers, with no knowledge of Japanese. Subjects were naïve, in that they had no linguistic experience – specifically, they had no experience in phonetics or phonetic transcription.

4.1.1. Data

The data that we used for these experiments were drawn from Japanese and Spanish portions of OGI Stories Corpus, a collection of multilingual conversational telephone speech recordings with phonetic annotations. We constructed two conditions for Italian subjects to transcribe:

1) Japanese – High phonetic overlap, limited phonotactic match
2) Spanish – High phonetic and phonotactic overlap

Figure 1: Annotator display for phonetic transcription
Subjects were each given 27 (19 Japanese, 8 Spanish) utterances to transcribe with an average of 1.73 seconds (22 phones) per utterance. Utterances were selected for gender balance and were limited to 2 seconds in length. In total, each subject transcribed ~600 phones amounting to ~45 seconds of total speech. On average, this took subjects 3.4 hours (~272x RT), similar to numbers reported in [14] and [11].

4.1.2. Presentation and Guidelines

Figure 1 shows the transcription interface used by subjects on this task. The tool was built on WaveSurfer [15] but designed to work with non-computer-savvy users. To transcribe segments, subjects simply highlight regions of the speech waveform using a mouse and then type the letter corresponding to the sound that they hear. As we used native orthography for this experiment, each phonetic unit corresponded to exactly one sound. Subjects were given a short set of samples to practice using the transcription tool. Prior testing of the tool during early pilot experiments suggested that subjects were able to learn the use of the tools features and start transcribing audio within 10-15 minutes after a short tutorial.

We did not ask Italian subjects to make long/short vowel distinctions and we asked subjects annotate /sh/ using special notation, as this was not represented in Italian as a single letter. Japanese glides /w/ and /y/ were normalized to Italian spelling “u” and “i” respectively.

4.1.3. Automatic Phone Recognition Baseline

In order to compare with automatic speech recognition techniques, we constructed a baseline HMM-based phonetic recognition system. We used phonetically transcribed data from OGI Stories to initialize standard 3-state HMM-based phonetic models. These models were then adapted via MAP to word-labeled Italian speech data from the telephone portion of the CLIPS corpus. We used standard PLP cepstral features (with their 1st, 2nd and 3rd order deltas) rotated via HLDA as a front end to this model. Observation probabilities were modeled using 16 gaussians per state. During decoding, no phonotactic language models were used. Instead a simple open phone-loop grammar was applied.

4.2. Word Transcription of Ambiguous Regions

We used 41 English native speakers from the MIT community for this experiment. All subjects had no prior linguistic knowledge or experience with ASR systems.

4.2.1. Data

We selected data from 400+ hours of the transcribed English FISHER corpus [16] using the procedure described in section 3. The speech data used for this task was conversational speech spoken over a telephone. From the corpus transcripts, we extracted 2,066 potentially ambiguous regions with near-homophone alternatives that met the aforementioned likelihood ratio criterion using a 4-gram LM.

4.2.2. Presentation and Guidelines

Subjects were presented with each region in one of two conditions:

1) Word Selection – As shown in Figure 2, subjects were asked to read a seven-word transcript in which the 4th word has been excised. Subjects were asked to make a multiple choice decision as to the word that best completes the sequence. Data from this condition was used to find lexically ambiguous regions for analysis of the subjects’ word transcription performance.

2) Word Transcription – Subjects were presented with audio from a subset of 1,180 items from the 2,066 potentially ambiguous regions. The corresponding audio was found through forced alignment with reference transcriptions. The order of presentation was randomized and all items were presented in a latin square design to ensure that each subjects would only see a particular item in one of the two conditions.

4.2.3. Automatic Speech Recognition System

The system we used for this task was designed to work with conversational telephone speech. We trained context dependent triphone models using 600+ hours of transcribed FISHER data from LDC. Triphone states were clustered into 5,746 state clusters for use during recognition. Each state cluster was trained with up to 32 gaussians per state using ML training followed by 9 iterations of MPE training. The resulting models perform at 28% WER on the rt02 test set.

For all experiments we apply a 2-gram language model to generate lattices that are then rescored with a 4-gram language model trained using the transcripts from training.

5. Results

In this section we present results from both protocols and compare these to those of the ASR baselines described above.

5.1. Foreign Language Phonetic Transcription

Tables 1 and 2 summarize the results of Italian subjects transcribing conversational telephone speech from the OGI Japanese and Spanish corpora respectively. Results from automatic phonetic recognition are also shown in these tables using the baseline automatic systems trained as described in 4.1.3. For direct comparison, results are shown in terms of Phone Error Rate (PER) as computed by NIST’s sclite.

<table>
<thead>
<tr>
<th>System</th>
<th>Subst</th>
<th>Del</th>
<th>Ins</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR</td>
<td>10.6</td>
<td>8.6</td>
<td>10.2</td>
<td>37.5</td>
</tr>
<tr>
<td>Human</td>
<td>16.6</td>
<td>10.7</td>
<td>10.2</td>
<td>37.5</td>
</tr>
</tbody>
</table>

Table 1: OGI Japanese transcription results (c.i. for subject means ±2.8%, p=0.05)

<table>
<thead>
<tr>
<th>System</th>
<th>Subst</th>
<th>Del</th>
<th>Ins</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR</td>
<td>22.0</td>
<td>8.8</td>
<td>7.8</td>
<td>38.7</td>
</tr>
<tr>
<td>Human</td>
<td>10.6</td>
<td>3.9</td>
<td>4.8</td>
<td>19.3</td>
</tr>
</tbody>
</table>

Table 2: OGI Spanish transcription results (c.i. for subject means ±2.3%, p=0.05)

From these results, we can readily see that there is significant performance variation across human subjects. The performance of our Italian subjects differed significantly on Spanish vs. Japanese data. This is likely due to the matched phonotactics and many shared cognate words between Spanish and Italian. Comparing with ASR-based phone recognition, human subjects are 15% more accurate. The best human subjects are more than two times more accurate.
Table 3 shows the most common mistakes made by both human subjects and ASR phonetic recognition. Interestingly, the human subjects and the ASR system tended to make very different errors. Whereas human transcribers had more difficulty with consonants, all but one of the top-3 errors made by the ASR system related to vowels or semivowels.

<table>
<thead>
<tr>
<th>Type</th>
<th>Japanese</th>
<th>Spanish</th>
<th>ASR(JP)</th>
<th>ASR(SP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subst</td>
<td>/b/ − /d/</td>
<td>/b/ − /v/</td>
<td>/b/ − /b/</td>
<td>/b/ − /d/</td>
</tr>
<tr>
<td></td>
<td>/v/ − /v/</td>
<td>/v/ − /v/</td>
<td>/v/ − /v/</td>
<td>/v/ − /v/</td>
</tr>
<tr>
<td>Del</td>
<td>/n/ − /m/</td>
<td>/n/ − /n/</td>
<td>/n/ − /n/</td>
<td>/n/ − /n/</td>
</tr>
<tr>
<td>Ins</td>
<td>/a/ − /a/</td>
<td>/a/ − /a/</td>
<td>/a/ − /a/</td>
<td>/a/ − /a/</td>
</tr>
</tbody>
</table>

Table 3: Top-3 errors made by subjects and ASR

5.2. Near Homophone Experiments

Tables 4 shows subject results from 2,066 regions of the word selection task used to filter lexically biased items. By design these items have a word prediction error rate near chance when scored with a 4-gram LM (because of the pre-filtering process described in section 3). As such, Table 4 shows the ability of human subjects to decrease the 4-gram LM’s word error rate by 45%.

Subjects transcribed a 1,180-item subset of the total set of near-homophone items that contained at least one phone difference (minimal pairs). Table 5 shows results for this subset including word selection performance and transcription performance. The addition of acoustic information allows subjects to reduce the number of errors by an additional 44% over the word selection baseline. Again, the performance of human subjects varies widely.

Table 4: Word selection performance (c.i. for subject means ±2.0%, p=0.05)

<table>
<thead>
<tr>
<th>LM Baselines</th>
<th>4-gram LM</th>
<th>Word Select</th>
<th>Average Subject</th>
<th>Best Subject</th>
<th>Worst Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>46%</td>
<td>25%</td>
<td>14%</td>
<td>9%</td>
<td>22%</td>
</tr>
</tbody>
</table>

Table 5: Word transcription performance (c.i. for subject means ±1.2%, p=0.05)

From the transcribed items, we selected a subset that had, on average, chance performance (50%) for human subjects doing word selection without audio to minimize lexical bias effects. Table 6 shows the performance of human subjects in comparison to our ASR baselines.

Table 6: Transcription performance on near chance items

<table>
<thead>
<tr>
<th></th>
<th>4-gram LM</th>
<th>ASR forced alignment</th>
<th>ASR recognition</th>
<th>Average Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>62.1%</td>
<td>17.1%</td>
<td>34.8%</td>
<td>14.4%</td>
</tr>
</tbody>
</table>

Interestingly, subject performance doesn’t differ much for this subset of items despite the lack of language bias information. Here the human subjects perform 15.7% better than the forced alignment baseline (in which segmentation and context are provided to the ASR system) and 58.6% better than speech recognition. This is, in part, explained by the inferior performance of the 4-gram language model used the ASR system. As these items have exactly one phone difference, each word error corresponds to one phone error. As such, the word error rate shown for subjects in Table 6 is a lower bound of the phonetic confusion rate for human subjects given the data set.

6. Discussion

In both experiments we found that human subjects were at least 15% better than an ASR baseline when doing phonetic recognition in either protocol. These results reflect phone recognition performance alone, without segmentation and lexical bias. To date, we know of no other studies that have attempted to measure this with natural, conversational speech. Using the near-homophone protocol, we also found that subjects were better able to make use of word context information than a standard 4-gram LM (45% better).

Using the transcription protocol required significant subject effort and, though we tried to minimize this through the use of native orthography, short segments, and simplified tools, more data is needed to verify the protocol. Similarly, baseline ASR systems could be improved. Currently, the best conversational telephone systems can achieve a WER < 20% and the best phonetic recognizers report PER < 25% on TIMIT data in matched train/test conditions. In the case of phonetic recognition, our test baselines are not well matched to training because of language differences. More work would be needed to test newer recognizers with this mismatched condition.

Interestingly, machine phonetic recognition, though different in means, seems to be quite close to human performance. We also know from prior interannotator studies, human transcribers have error rates < 10% for word recognition [2]. As our study suggests, human listeners seem better able to make use of lexical, syntactic or semantic information during speech understanding.

7. References