PERCEPTUAL BENCHMARKS FOR AUTOMATIC LANGUAGE IDENTIFICATION

Yeshwant K. Muthusamy[1] *Neena Jain*[2] *Ronald A. Cole*[2]

[1] Computer Science Laboratory, Texas Instruments, Dallas, Texas 75265, USA
[2]Center for Spoken Language Understanding, Oregon Graduate Institute, Portland, Oregon 97291-1000, USA

ABSTRACT

There has been renewed interest in the field of automatic language identification over the past two years. The advent of a public-domain ten-language corpus of telephone speech has made the evaluation of different approaches to automatic language identification feasible. In an effort to provide benchmarks for evaluating machine performance, we conducted perceptual experiments on 1-, 2-, 4- and 6-second excerpts of telephone speech excised from spontaneous speech utterances in this corpus. The subject population consisted of 10 native speakers of English and 2 speakers from each of the remaining 9 languages. Statistical analyses of our results indicate that duration of the excerpt, familiarity with the language, and number of languages known are important factors affecting a subject's performance on the identification task.

1. INTRODUCTION

Automatic language identification, the problem of recognizing what language is being spoken, is a challenging research problem with important real-world applications. A recent workshop report identified automatic language identification as one of the key research areas needed for multilingual spoken language systems [1]. The National Institute of Standards and Technology (NIST) conducted the first evaluation of language identification algorithms in June 1993, with participation from seven different research laboratories.

Given the surge of interest in language identification, it is important to study human performance on similar tasks. Perceptual studies with listeners from different language backgrounds provide benchmarks for evaluating machine performance. In addition, patterns of confusions between languages provide insights about the salient acoustic characteristics that can be useful for automatic language identification.

In this paper, we report results of perceptual experiments conducted on 1-, 2-, 4- and 6-second excerpts of telephone speech excised from utterances in ten languages. This research is an extension of our earlier work reported in [2].

2. SPEECH EXCERPTS

The speech was taken from the OGI Multi-language Telephone Speech Corpus [2, 3], which has been designated as the standard for evaluating language identification algorithms by NIST. The corpus contains fixed vocabulary and spontaneous speech utterances from 90 different speakers each of English, Farsi, French, German, Japanese, Korean, Mandarin Chinese, Spanish, Tamil and Vietnamese collected over commercial telephone lines.

Excerpts of speech of 1, 2, 4 and 6 seconds duration were excised from the 1-minute spontaneous speech ("story") utterances in the ten languages. Care was taken to ensure that silence constituted less than half of each excerpt. The number of speakers from each language was determined by the language containing the least number of speakers whose stories were long enough to provide all four excerpts. Since Korean had only 76 speakers satisfying this criterion, the excerpts were excised from 76 stories in each language.

3. EXPERIMENT I

In [2], we reported results of a preliminary perceptual experiment using the excerpts of speech described above. This experiment is briefly described below for the sake of completeness.

3.1. Design

Seven female and four male mono-lingual native English speakers participated in the experiment. It was designed such that each subject listened to exactly one excerpt from each speaker, and an equal number of excerpts at each duration from each language. The excerpts were chosen at random, keeping the above constraints in mind. The experiment was conducted using an interactive graphics program that played excerpts of speech of 1, 2, 4 and 6 seconds, chosen at random from each of the 10 languages. The program maintained a log of the subjects' responses.

3.2. Procedure

The subjects first went through a training session of 40 excerpts, to become familiar with the experimental procedure and the languages. They were then presented with 760 different excerpts, 19 at each duration from each language. The subjects could listen to each excerpt as many times as they desired. After responding, they were given feedback on every trial. The subjects could also listen to an excerpt *after* making the choice—a feature that was included to aid in the learning process. In fact, this feature was rarely used by the subjects. Each block of 100 trials was considered a session, and the program automatically quit after every 100 trials, to ensure that the subjects did not get fatigued.

0-7803-1775-0/94 $3.00

3.3. Results

As duration increased from 1 to 2 to 4 to 6 seconds, the average subject performance over all languages rose from 37.0% to 43.0% to 51.2% to 54.6% respectively. The average performance, over all subjects and all durations, was 46.5%, with individual language scores ranging from 14.7% (Korean) to 96.3% (English). As expected, the subjects identified English with high accuracy at all four durations. Excluding English, the average performance at each duration was 20.7%, 37.4%, 45.8% and 49.7%.

Analysis of performance by each quarter of the experiment (190 trials) revealed little evidence of learning during the experiment. Figure 1 displays the average subject performance by language on 6-second excerpts, for the first and last quarters. The subjects found Farsi, Korean, Mandarin and Vietnamese to be the most difficult languages. For example, the average performance on Korean for the first and last quarters was 13.5% and 16.7% respectively.

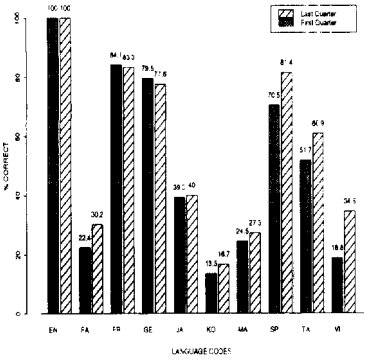

Figure 1. Experiment I: Average Subject Performance on 6-second Excerpts for the First and Last Quarters

4. EXPERIMENT II

This experiment was designed to determine the effect of additional trials on classification performance, and to examine differences in classification performance for speakers of other languages.

4.1. Design

Each subject listened to exactly two excerpts from each speaker, and an equal number of excerpts at each duration from each language. The excerpts were chosen at random, keeping the above constraints in mind. The two excerpts were chosen such that one was from either the 1- or 2-second excerpts, and the other from either the 4- or 6-second excerpts. Thus, compared to Experiment I, there were now twice as many speech excerpts (1520), 38 at each duration from each of the 10 languages.

4.2. Subject Selection

A total of 28 subjects (14 males and 14 females) were paid to participate in this experiment. There were 10 native speakers of English and 2 speakers from each of the other 9 languages. The subjects ranged in age from 16 to 54. All the subjects were literate in English. Prior to the start of the experiment, subjects were asked a number of questions regarding their places of residence during their language formative years and their knowledge of languages. This information proved useful in the analyses of the results. The number of languages known by the subjects ranged from 1 to 4.

4.3. Procedure

The experiment was conducted using the same interactive graphics program as in Experiment I. Before the start of the main experiment, subjects went through an augmented training session in which they (i) listened to two 10-second excerpts of speech from each language (from a male and female speaker) as many times as they wished, and (ii) performed a dry run of the experiment, using 80 excerpts from all ten-languages (8 excerpts per language). The objective was to familiarize the subjects with both the languages and the use of the graphical interface. Care was taken to ensure that there was no overlap in the utterances used for the 10-second excerpts, the dry run and the main experiment.

In the main session, subjects were presented with 1520 different excerpts, 38 at each duration from each language. All the other details of the experimental procedure were identical to those of Experiment I. The average subject completed the experiment in 2 to 3 days.

After completion of the experiment, the subjects were interviewed to determine the strategies they used to identify the excerpts. In particular, they were asked if they had trained themselves to hear any characteristic phonetic or prosodic cues for each language. Their responses were logged and analyzed to determine any consistent trends.

4.4. Results

As duration increased from 1 to 2 to 4 to 6 seconds, the average performance over all languages rose from 44.8% to 52.8% to 62.3% to 65.3% respectively. The average performance, over all subjects and all durations was 56.7%, with individual language scores ranging from 28.3% (Korean) to 93.1% (English). Excluding English, the average performance of the 10 native English speakers at each duration was 33.6%, 41.6%, 51.3% and 54.2% respectively. In comparison, the corresponding figures for Experiment I were 20.7%, 37.4%, 45.8% and 49.7%. The scores are consistently better at all durations in Experiment II. The average (native English) subject performance over all durations was 50.2%. Table 1 displays the confusion matrix corresponding to this result. The entries of the matrix are percentages. The performance on each language is given at the end of each row. Korean, the language with the lowest score (15.4%), is confused more often with Farsi, Japanese, Mandarin, Tamil, and Vietnamese—languages to which native English speakers in the U.S. are rarely exposed, compared to European languages such as German, French and Spanish.

Figure 2 displays the average subject performance as a function of number of languages known. There seems to be a substantial difference in performance between subjects with knowledge of three languages and those with knowledge of four languages. Further, knowledge of a particular language definitely helped in identifying excerpts from it, as

Table 1. Confusion Matrix of Average Performance of the 10 Native English Speakers

Lang.	*EN*	*FA*	*FR*	*GE*	*JA*	*KO*	*MA*	*SP*	*TA*	*VI*	*Perf.*
EN	96.1	0.5	0.5	0.7	0.3	0.4	0.2	0.3	0.7	0.3	96.1%
FA	2.7	28.6	13.2	11.0	7.3	6.4	5.1	8.2	12.3	5.2	28.6%
FR	0.6	6.2	69.3	9.2	3.4	1.8	2.2	2.6	3.0	1.6	69.3%
GE	2.4	5.1	10.1	68.5	3.3	1.8	2.2	2.3	2.6	1.7	68.5%
JA	0.5	7.1	4.1	3.2	40.0	10.8	11.0	6.4	7.4	9.5	40.0%
KO	1.9	11.6	7.6	5.7	18.4	15.4	13.3	4.7	10.0	11.4	15.4%
MA	1.8	7.4	6.4	4.5	18.3	8.8	29.7	1.6	3.0	18.5	29.7%
SP	1.6	5.7	7.0	3.3	4.1	2.4	3.2	63.2	7.2	2.3	63.2%
TA	1.8	10.1	3.1	2.6	5.5	5.7	2.6	10.4	51.8	6.4	51.8%
VI	1.2	6.7	3.2	3.0	13.0	10.3	12.4	2.9	7.2	39.9	39.9%

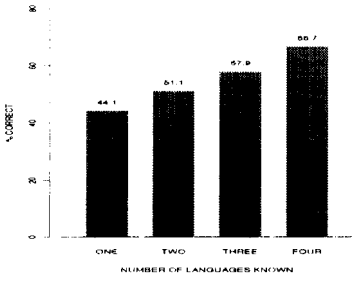

Figure 2. Average Listener Performance by Number of Languages Known

shown by Figure 3, which compares average performance of subjects that know each language with those that do not. The most striking result is for Korean: subjects who knew the language scored 93% correct, while those who did not scored 22.6%.

Unlike Experiment 1, analysis of performance by each quarter of the experiment (380 trials) did reveal evidence of learning during the experiment for all languages. Figure 4 displays the average performance, over all subjects, on the 6-second excerpts in the first and last quarters. Figure 5 shows the corresponding plot for just the 10 subjects whose native language was English. From Figures 1 and 4, we can conclude that human listeners from different language backgrounds perform better on this task than native speakers of English. From Figures 1 and 5, we can conclude that increased training and the larger number of excerpts have contributed to improved identification performance for native English speakers (assuming comparable language proficiency of subjects in both experiments).

4.5. Analysis of Cues

The following is a summary of responses of the subjects in post-experiment interviews:

- As expected, none of the subjects had any problems with English.
- Farsi had several aspirated sounds. Frequent occurrences of the phoneme /sh/.

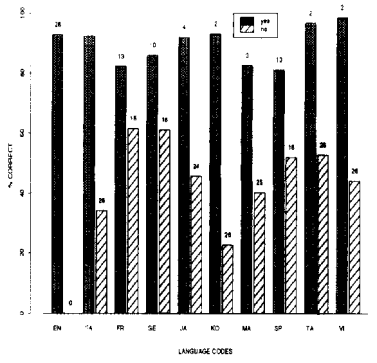

Figure 3. Average Listener Performance by Knowledge of the Language. The number on top of each bar refers to number of subjects that either knew or did not know the language.

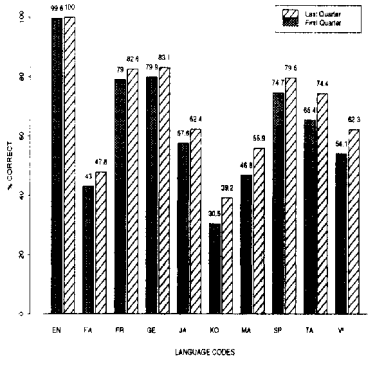

Figure 4. Average Subject Performance on 6-second Excerpts in the First and Last Quarters: All Subjects

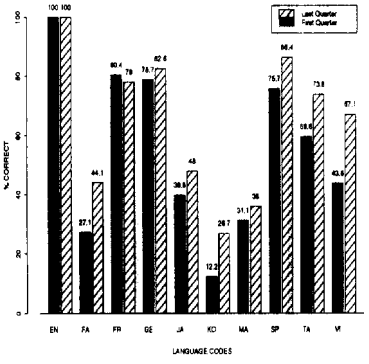

Figure 5. Average Subject Performance on 6-second Excerpts in the First and Last Quarters: Native Speakers of English Only

- French had several nasal sounds. Subjects found its intonation distinct.
- German had the distinctive word *ich* (the velar fricative *kh*). Many subjects found German "harsh" and with frequent aspirated velar sounds. Some subjects confused it with Farsi.
- Japanese had several "crisp" stops, and distinctive words such as *watashiwa* and *mashita*.
- Every subject who was not a native Korean had problems recognizing Korean. Many of the subjects looked for the word *imnida*. Others chose Korean when they were not sure what language it was!
- Subjects found the tones of Mandarin informative. Some of them confused it with Japanese.
- There was no consistent cue that emerged from the responses for Spanish. Some subjects perceived a high speech rate for the Spanish speakers. Others looked for sounds such as the phoneme-pair "eh-s".
- Tamil had distinctive /r/ and /l/. Many subjects confused it with Spanish.
- Vietnamese had a sing-song intonation and several nasals, especially the velar nasal /ng/. Many subjects confused it with Mandarin.

Some subjects also found that for some excerpts, the "voice quality" of the speaker was similar to that of a friend whose native language they knew. This information also helped in the identification process.

These responses indicate that the subjects used a combination of word and phoneme-spotting, and phonetic and prosodic cues to distinguish between the languages. The confusions reported between the languages is consistent with phonological evidence.

4.6. Analysis of Variance

An analysis of variance using *number of languages known* as the between-subjects factor and *language*, *quarter* (every 380 trials),and *duration of the excerpt* as the within-subjects factors, revealed the following observations:

- *number of languages known* was a significant factor in performance ($p < 0.015$), as were *language*, *quarter*, and *duration* ($p < 0.01$)
- the interaction between *language* and *quarter* and that between *language* and *duration* were significant ($p < 0.01$), i.e. some languages were learned sooner than others, and some languages were learned from just the short excerpts
- the interaction between *quarter* and *duration* was significant ($p < 0.023$), i.e., the longer duration excerpts were learned sooner.

5. DISCUSSION

The results indicate that with adequate training, human listeners are able distinguish between languages with accuracies ranging from 39.2% to 100.0% using just 6-second excerpts of speech (Figure 4; average performance: 69.4%). In comparison, the best machine results on the same corpus are 55% using 10-second excerpts from the stories, obtained by K. P. Li at the first NIST language identification evaluation in June 1993. Experiment II showed that increased exposure to each language and longer training sessions contribute to improved classification performance. Subjects who knew more languages tended to perform better, on the average, than subjects who knew just one language. While *a priori* knowledge of the language definitely helped, post-experiment interviews indicate that subjects learned to develop their own cues as the experiment progressed.

While these experiments have provided interesting results, the mix of subjects and multitude of languages makes it difficult to determine the cues that human listeners use to distinguish between two *completely unfamiliar* languages. Perceptual experiments using just pairs of languages and appropriately selected subjects might provide some answers to that intriguing question.

6. ACKNOWLEDGEMENTS

This research was supported by a grant from the Department of Defense to the Center for Spoken Language Understanding at OGI. The authors thank Etienne Barnard for his comments on the analyses of results.

REFERENCES

[1] R. A. Cole et al. Workshop on spoken language understanding. Technical Report CS/E 92-014, Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, 1992.

[2] Y. K. Muthusamy. *A Segmental Approach to Automatic Language Identification*. PhD thesis, Oregon Graduate Institute of Science & Technology, 1993.

[3] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. In *Proceedings International Conference on Spoken Language Processing 92*, Banff, Alberta, Canada, October 1992.