# TESTING WITH THE YOHO CD-ROM VOICE VERIFICATION CORPUS

Joseph P. Campbell, Jr. <jpcampb@alpha.ncsc.mil>
US Department of Defense, R2
Fort Meade, Maryland, USA 20755-6000

## ABSTRACT

A standard database for testing voice verification systems, called YOHO, is now available from the Linguistic Data Consortium (LDC). The purpose of this database is to enable research, spark competition, and provide a means for comparative performance assessments between various voice verification systems. A test plan is presented for the suggested use of the LDC's YOHO CD-ROM for testing voice verification systems. This plan is based upon ITT's voice verification test methodology as described by Higgins, et al., but differs slightly in order to match the LDC's CD-ROM version of YOHO and to accommodate different systems. Test results of several algorithms using YOHO are also presented.

## INTRODUCTION

The YOHO voice verification corpus was collected while testing ITT's prototype speaker verification system in an office environment [1]. This database is the largest supervised speaker verification database known to the author. The number of trials and the number of test subjects were chosen to allow testing at the 75% confidence level to determine whether a system meets 0.1% false rejection and 0.01% false acceptance. The test subjects spanned a wide range of ages, job descriptions, and educational backgrounds. Most subjects were from the New York City area, although there were many exceptions, including some nonnative English speakers. A high-quality telephone handset (Shure XTH-383) was used to collect the speech; however, the speech was not passed through a telephone channel. When the system was used in an enrollment or verification session, a sampled waveform file was created for each phrase-length utterance. A subset of these waveform files comprises the LDC's YOHO CD-ROM.

The LDC release of YOHO was designed, with regard to the quantity and collection of data, to answer the following question: does a speaker verification system perform at 0.1% false rejection and 0.01% false acceptance at 75% confidence with a 50% probability of passing the test? There are 138 speakers (106 males† and 32 females†); for each speaker, there are 4 enrollment sessions of 24 utterances each and 10 verification sessions of 4 utterances each. In a text-dependent speaker verification scenario, phrases are prompted and the claimant is requested to say them. The syntax used in the YOHO database is "combination lock" phrases. For example, the prompt might read: "Say: twenty-six, eighty-one, fifty-seven." Where the claimant is to speak the phrase as three doublets. The LDC YOHO CD-ROM can be summarized as

- "Combination lock" phrases
- 138 subjects: 106 males†, 32 females†
- Collected over 3-month period
- Approximately 3-day verification intervals
- Real-world office environment
- 4 enrollment sessions per subject
- 24 phrases per enrollment session
- 10 verification sessions per subject
- 4 phrases per verification session
- Total of 1,380 verification sessions†
- 8 kHz sampling with 3.8 kHz bandwidth
- 1.2 gigabytes of data (when uncompressed)

† Contrary to the CD's 0readme.txt file.


0-7803-2431-5/95 $4.00 

The speech data is divided into two directories to separate enrollment and verification sessions. The enrollment and verification (or identification) phases should only use data from their respective directories.

## ENROLLMENT

Speaker enrollment models should be constructed from enrollment sessions 1 through 3. Session 4 can be used to determine cohort [2] (also known as ratio or likelihood [1]) set speakers and used for building a speech segmenter.

Unlike some text-dependent speaker verification systems, not all possible verification phrases are available from enrollment (this would lead to excessive enrollment time). Enrollment does, however, cover the acoustic space of all possible speech that could be prompted during verification. For example, during enrollment, models for a given speaker's "fif"-"tee"-"three" can be obtained without actually collecting "Fifty-Three" by using subwords from other prompts; e.g., 51, 52, 63, and 73 (minus coarticulation effects). Because of the difficulty this may cause for some, text-independent test results also can be reported using YOHO.

The enrollment file structure of the disc is `enroll/speaker#/session#/prompted_phrase.wav`. For example, speaker 101's enrollment session 1 phrase "26_81_57" file is

```
enroll/101/1/26_81_57.wav
```

Each session of each speaker's enrollment directory contains 24 `*.wav` files.

There is a total of 138 speakers, numbered from 101 to 277 (there are gaps in the sequence).

## VERIFICATION

A single trial can use all the speech in a given speaker's verification session (i.e., up to four phrases). Each speaker can have 10 verification tests against him/herself. (If the four phrases were used for separate verification tests, the independence of the tests would be weak.)

The verification file structure of the disc is `verify/speaker#/session#/prompted_phrase.wav`. For example, speaker 101's verification session 1320 consists of the following set of 4 speech files:

```
verify/101/1320/41_34_23.wav
verify/101/1320/57_92_26.wav
verify/101/1320/73_61_31.wav
verify/101/1320/86_79_65.wav
```

There is a total of 1,380 sessions†, numbered from 528 to 2527 (there are gaps in the sequence).

False-rejection measurements are based on the 1,380 valid session trials. Impostor trials are simulated by presenting the system with one subject's speech and prompted text (embedded in the file name) under a different subject's hypothesized identity.

### Impostor Selection

To be consistent with ITT, the cohort set speakers should be excluded as impostors and speaker dependent cohort sets should consist of the five "closest" speakers as determined from the 4th enrollment session.

Determining a fair way to compare systems using different size cohort sets is a difficult problem. Cohort set speakers, by definition, are usually good impostors. As the cohort set size increases, excluding cohort set speakers increasingly eliminates good impostors. This may optimistically bias the results in favor of larger cohort set size systems. Furthermore, if the cohort set speakers are not excluded as impostors, they are likely to be rejected (since the system has models for them). Thus, the results again may be optimistically biased toward larger cohort set size systems. Larger cohort set sizes should yield improved real-world unseen impostor performance, but one must be aware of these possible testing biases.

Cross-validation could be used to iteratively partition impostor and cohort sets, but this may reduce the statistical confidence of the tests.

Each of the 138 subjects shall be treated in turn as a claimant. For each claimant, sessions spoken by subjects other than the claimant and his/her cohorts are selected as impostors, with no more

than one session per subject (for independent tests). The sessions are processed using the normal verification procedure, resulting in accept/reject decisions. If 13,862 simulated impostor trials are performed, the most stringent test in Table 1 can be evaluated.

Impostor results should be reported in these three categories (see the speaker.doc file for speaker genders):

- males vs noncohort males
- females vs noncohort females
- each subject vs all other noncohort subjects

The last category is a compromise because the female population is too small to perform high confidence female-only impostor testing. It's also necessary to use all the data to provide the 13,862 trials required for the most stringent test in Table 1.

### Critical Number of Errors

To test the hypothesis that the actual false rejection (FR) rate is less than or equal to 1% at 75% confidence requires 8 or fewer errors in 1,080 tests (for a 70% probability of passing the test if the ratio of the true system error rate to the target error rate (e) = 2/3) [1]. Likewise, as shown in Table 1, to test the hypothesis that the actual false acceptance (FA) rate is less than or equal to 0.1% at 75% confidence requires 8 or fewer errors in 10,802 tests. These tests are based upon the independence assumptions used in the collection and proper use of the YOHO database, Poisson's approximation to the binomial, error rates less than 5%, and sample sizes greater than 100.

**Table 1:** Critical Number of Errors

| Mode | Target | Conf | Ppass | e | Size | Crit Err |
|---|---|---|---|---|---|---|
| FR | 1.0% | 75% | 0.7 | 2/3 | 1,080 | 8 |
| FA | 0.1% | 75% | 0.7 | 2/3 | 10,802 | 8 |
| FR | 0.1% | 75% | 0.5 | 1/2 | 1,386 | 0 |
| FA | 0.01% | 75% | 0.5 | 1/2 | 13,862 | 0 |

## REPORTING RESULTS

In addition to using the critical number of errors tests, a number of other reporting means are of interest:

- Raw error rates (relative frequency)
- Histograms of the number of errors vs number of speakers for each error type (e.g., subject falsely rejected, subject falsely accepted as another, and another falsely accepted as subject)
- Receiver operating curves, preferably bracketed by error bars
- A histogram of the identification rank
- Average identification rank

Fine-grain results on problem speakers can be informative (e.g., 3-D plots of attacker's vs attackee's identification numbers vs number of false acceptance errors).

For text-dependent systems, errors due solely to speech misrecognition should be reported.

In order for the community to make comparative assessments, please explicitly state the options selected and any variations on this suggested test plan used when publishing results. The author welcomes your results.

## TEST RESULTS

The author knows of six tests using the YOHO databases. ITT's CSR [1] and NN [3] results are on the full 186 speaker YOHO database; MIT Lincoln Lab's [4] and Rutgers' NTN [5], HMM [6] and LVQ [6] results are on the LDC's YOHO CD-ROM; and Campbell's results [7] are on an 87 speaker subset of the YOHO database. Equal-error rate (EER) verification and closed-set speaker identification error rates are given in Table 2.

Since these tests were not performed under identical conditions, they cannot be compared directly with each other. They are presented to show a variety of algorithms and their corresponding performance. Please refer to the references for descriptions of the algorithms and test procedures used.

**Table 2:** YOHO Speaker Recognition Error Rates

| | Verification EER | Speaker Id closed-set |
|---|---|---|
| **ITT's CSR** | 1.7% | |
| **ITT's NN** | 0.5% | |
| **MIT/LL's GMM** | 0.51%<br>0.2%m, 1.8%f | 0.8% error<br>1.1 avg rank |
| **Rutgers' NTN** | 0.65% | |
| **Rutgers' HMM** | | 1.36% error<br>1.05 avg rank |
| **Rutgers' LVQ** | | 0.36% error<br>1.03 avg rank |
| **Campbell's divergence** | | 1.15% error<br>1.01 avg rank |

## PROBLEMS WITH YOHO

The following files were not compressed and contain empty headers (the speech data is intact); thus, w_decode is not needed for these files:

```
verify/277/538/29_51_23.wav
verify/277/538/65_56_74.wav
verify/277/538/74_31_67.wav
verify/277/538/96_85_43.wav
```

The LDC promises to provide a script to solve this problem. It should be available via anonymous ftp from ftp.cis.upenn.edu as /pub/ldc/yohosphr.prl.

Note that speaker 101 said "53-73-79" instead of the prompted phrase "56-73-79" in enrollment session 2. Be aware that speaker 240 used a falsetto voice in verification session 969. Please state if any data is excluded from your tests.

The author would appreciate reports of any other errors on the YOHO CD-ROM.

## LDC INFORMATION

For information about the LDC, including obtaining copies of YOHO, please contact the Linguistic Data Consortium, 441 Williams Hall, University of Pennsylvania, Philadelphia, PA 19104-6305, USA. Information about the LDC is also available on their home page and via anonymous ftp from ftp.cis.upenn.edu in the /pub/ldc directory.

## CONCLUSIONS

The LDC's YOHO CD-ROM and its errata were described. A test plan was proposed that will, hopefully, unify the reporting of the performance of speaker verification systems. The performance of a few systems was presented. Future editions of this document will be available from the LDC's home page and by ftp from ftp.cis.upenn.edu.

## ACKNOWLEDGMENTS

The assistance and contributions of Ron Benincasa, Tom Crystal, Mark Forsyth, David Graff, Alan Higgins, Jerry O'Leary, Han-Sheng Liou, Qiguang Lin, Jack Porter, Scott Reider, Doug Reynolds, and Michael Schmidt are gratefully acknowledged.

## REFERENCES

[1] Higgins, A., L. Bahler, and J. Porter. "Speaker Verification Using Randomized Phrase Prompting." *Digital Signal Processing* 1, no. 2 (1991): 89 - 106.

[2] Rosenberg, A., J. DeLong, C-H. Lee, B-H. Juang, and F. Soong. "The Use of Cohort Normalized Scores for Speaker Verification." In *International Conference on Spoken Language Processing in Banff,* University of Alberta, 599 - 602, 1992.

[3] Higgins, A., L. Bahler, G. Vensko, J. Porter, and D. Vermilyea. *YOHO Speaker Authentication Final Report.* ITT Aerospace/Communications Division, 1992.

[4] Reynolds, D. "Speaker Identification and Verification using Gaussian Mixture Speaker Models." Submitted to *Speech Communication,* Spring 1995.

[5] Liou, H-S. and R. Mammone. "A Subword Neural Tree Network Approach to Text-Dependent Speaker Verification." To appear in *International Conference on Acoustics, Speech, and Signal Processing in Detroit,* IEEE, 1995.

[6] Lin, Q. and C. Che. "Personal Communication." December 1994.

[7] Campbell, J. P., Jr. "Features and Measures for Speaker Recognition." Ph.D. Dissertation, Oklahoma State University, 1992.