COHORT SELECTION AND WORD GRAMMAR EFFECTS FOR SPEAKER RECOGNITION

J. M. Colombi, D. W. Ruck
S. K. Rogers, M. Oxley

Air Force Institute of Technology
Wright-Patterson AFB, OH, 45433

T. R. Anderson

Armstrong Laboratory/CFBA
Wright-Patterson AFB, OH, 45433

ABSTRACT

Automatic speaker recognition systems are maturing and databases have been recently designed to specifically compare algorithms and results to target error rates. The LDC YOHO Speaker Verification Database was designed to test error rates at the 1% false rejection and 0.1% false acceptance level. This work examines the use of speaker-dependent (SD) monophone models to meet these requirements. By representing each speaker with 22 monophones, both closed-set speaker identification and global-threshold verification was performed. Using four combination lock phrases, speaker identification error rates are obtained at 0.19% for males and 0.31% for females. By defining a test hypothesis, a critical error analysis for speaker verification is developed and new results reported for YOHO. A new Bhattacharyya distance is developed for cohort selection. This method, based on the second order statistics of the enrollment Viterbi log-likelihoods, determines the optimal cohorts and achieves an equal error rate of 0.282%.

1. INTRODUCTION

The LDC YOHO Database is one of the largest speaker verification databases in existence. Based on Campbell's recent overview [1], only a few text-dependent verification experiments have been conducted on this set. While HMMs have enjoyed a long success for speech recognition, their effectiveness is only recent for speaker modeling [2, 3]. This research examines speaker dependent monophone modeling [4] for speaker verification, the various existing methods for cohort selection and presents experimental data on the effects of word grammar and cohort selection strategies.

2. HMM TRAINING

The speech material consists of "combination-lock" phrases. Each phrase consists of three number doublets. The doublets are chosen from a list which includes all the doublets from 21 to 99 with the following exceptions: (1) no exact decades (30, 40, etc.), (2) no double digits (22, 33, etc.), and (3) no numbers ending in "8" (28, 38, etc.) [1]. Thus, an inherent word-pair grammar can be incorporated into the language model using Viterbi [5].

Feature extraction consists of filtering 25 msec windows, analyzed every 10 msec, using 24 equally spaced Mel-scale filters, followed by a liftered cosine transform representing 12 MFCC coefficients, and keeping a normalized log energy. Delta and acceleration features have been added using standard regression practices [5] resulting in a 39 dimensional representation. This choice allowed the bootstrapping of the speaker dependent HMMs by a full set of TIMIT speaker independent (SI) monophone models.

Using these previously trained 3-state TIMIT models, the entire YOHO database was phonetically labeled, using HMM time-alignment [5]. Table 1 lists the possible models, constrained to the YOHO vocabulary, with an additional leading and trailing silence /sil/ and interword space /sp/. Based on the word grammar chosen, the phone /dx/ may or may not be present. This comes from the choice of word grammar - one based on TIMIT (clear read text), Resource Management (conversational) or some combination. Word grammar is exemplified in Table 2. Though comparison of several male test speakers indicate unique bigram probabilities of the /t/ and /dx/ phones, overall recognition differences were not significant when constrained in either the TIMIT or RM grammar. The TIMIT-like dictionaries were used for all reported results.

Table 1: YOHO phonemes, with silence (sil) and interword space (sp).

ah	ax	(dx)	er	f	iy	n	s	th	v	sil
ao	ay	eh	ey	ih	k	r	t	uw	w	sp

Once the YOHO database was phonetically marked, enrollment data was used to train speaker dependent models. The architecture was the same as the TIMIT SI models: 3 state left-to-right with single mixture densities. Enrollment data contains 24 combination lock utterances recorded at 4 different sessions. First the HMMs were reestimated individually using the Baum Welch algorithm, with the initial model being the speaker independent TIMIT monophones. Then, an embedded reestimation of all speakers' models was accomplished by concatenating the individual monophones for each utterance and updating all HMMs simultaneously again using the Baum-Welch algorithm. This method overcomes any limitation of the initial TIMIT alignment process.

Table 2: Example YOHO word grammar. [ ] denotes optional monophone, and | denotes dual path through the word grammar.

Word	Monophones	Grammar
EIGHTY	ey t iy [sp]	TIMIT
	ey dx iy [sp]	RM
	ey dx|t iy [sp]	OPTION
NINETY	n ay n t iy [sp]	TIMIT
	n ay n iy [sp]	RM
	n ay n [t] iy [sp]	OPTION

3. SPEAKER IDENTIFICATION

Speaker identification uses a Bayesian classifier, assuming equal priors, choosing speaker model i from the normalized Viterbi log likelihoods for an utterance (or set of utterances), X. These result provide a reference for speaker separation and model/ feature choice trade-offs.

$$i = \arg\max_k \{ \log p(X|\lambda_k) \}$$

Table 3 are error rates for various numbers of combination lock phrases. For each gender, two different Viterbi constraints were examined, Forced Transcription alignment and Word-Pair Grammar. The latter can be used to check if the prompted text matched the most likely Viterbi label hypothesis. The Word-Pair grammar also catches many confused doublets over a simply word dictionary grammar. For example, for the prompt "75-29-47", Viterbi with word grammar only may hypothesize a transcription of "SEVENTY FIVE ONE NINE FORTY SEVEN" where this label is not valid under a word pair grammar. All remaining results use forced Viterbi alignment based on the prompted transcription.

Table 3: Closed-set error rates(%) for 1,2 and 4 combination phrases.

Method	Males(Females)		
	1	2	4
Word Pair	1.75 (2.19)	0.47 (0.63)	0.38(0.31)
Forced Viterbi	1.75(1.72)	0.47 (0.78)	0.19 (0.31)

A practical pattern recognition concern is the amount of training data for model reestimation. Each speaker is represented by 21 three-state monophone models, resulting in 4914 model parameters per speaker. Based on an average of 38,700 enrollment observations, the ratio of training patterns to model parameters is 7.9. To increase this ratio, feature reduction and shared covariance was performed. Table 4 shows that reducing the model size by removing transitional features increases error rates, and sharing covariance matrices among monophone states shows the opposite effect.

Table 4: Closed-set error rates(%) using forced Viterbi decoding for 1,2 and 4 combination phrases. Base feature is MFCC + Energy.

Feature	Males(Females), Σ/state		
	1	2	4
Base+Δ + $\Delta\Delta$	1.70(1.72)	0.47(0.78)	0.19(0.31)
Base+Δ	2.52(2.34)	0.99(0.94)	0.57 (0.31)
Base	5.83(5.55)	2.55(2.34)	1.60(0.94)
Feature	Males(Females), Σ/monophone		
	1	2	4
Base+Δ + $\Delta\Delta$	1.06(1.48)	0.47(0.78)	0.19(0.31)
Base+Δ	1.37(1.25)	0.57(0.47)	0.28(0.31)
Base	3.56(2.19)	1.46(1.41)	0.85(0.31)

4. SPEAKER VERIFICATION

The log likelihood ratio is a useful tool based on Bayesian decision theory, especially for removing certain non-speaker characteristics of the utterance. The likelihood ratio is defined as

$$LR(X) \triangleq \log p(X|\lambda = \lambda_i) - \log p(X|\lambda \neq \lambda_i) \qquad (1)$$

where λ_i is the claimed speaker and X is an utterance or set of utterances. Classification compares this quantity to a threshold, which accounts for unknown priors and likelihood estimation biases. Often this threshold must be specified, either globally for all speakers or individual thresholds can be used. In order to judge systems as acceptable at a particular significance level, the maximum number of errors allowed is first determined. This statistic is the number of *Critical Errors*.

4.1. Critical Error Analysis

Higgins [6], and more recently Campbell [1], has examined the statistical significance of the YOHO experiments. This section re-examines this analysis by first defining the hypothesis for acceptable systems. Define the null hypothesis, H_0, that the System Error Rate, Ser, does not meet the Target Error Rate Ter,

$$\begin{aligned} H_0 &: Ser > Ter \quad \text{UNACCEPTABLE} \\ H_1 &: Ser \leq Ter \quad \text{ACCEPTABLE} \end{aligned} \qquad (2)$$

Previously, results have been reported at the 75% confidence level for False Acceptance and False Reject target values. However, this method would pass a large percentage of systems that are in reality unacceptable.

The main concern should not be the probability of meeting the Target Error Rate, which a confidence level analysis provides; the main concern should be in the decision to

reject potential candidates taking into account the consequences of a wrong decision. Conjecture all systems are unacceptable and allow the experimental evidence (observed errors) to reject this conjecture [7]. One can also examine the probability of failing acceptable systems, but this is a secondary concern.

Using the Poisson approximation to the binomial (which is good for error rates less than 5% and number of trials greater than 100), the critical error curves developed by Higgins [6] are redrawn using typical numbers for the significance level, Figures 1. The probabilities of accepting a system for various critical errors is given in Figure 2. Using these graphs allows recalculation of critical errors for YOHO in Table 5.

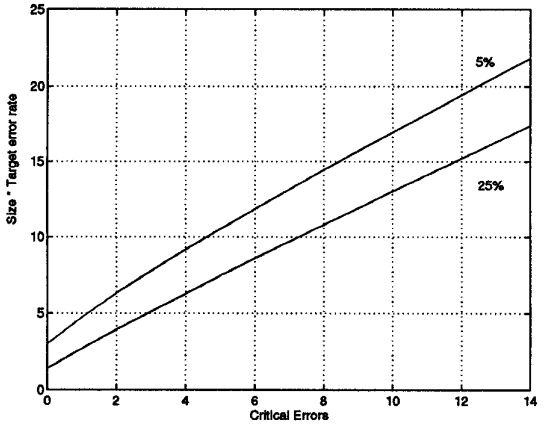

Figure 1: Critical Errors for tests designed at the 5% and 25% significance level.

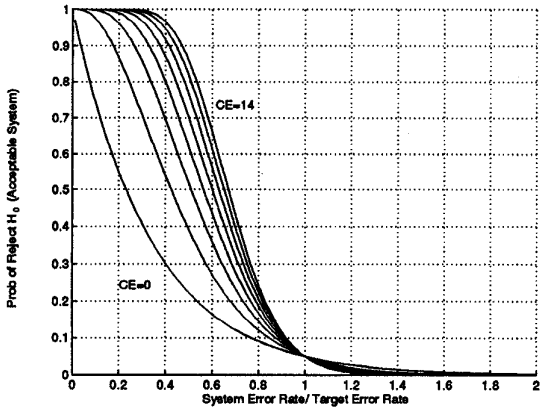

Figure 2: Probability of rejecting H_o - Accepting the system meets or exceeds the Target error rate.

Since the number of false rejects is limited, one cannot report results at the 5% significance level, and these entries are provided from [1]. Also, we chose to use all imposter tests available, counting each session as statistically independent. This amounts to total false acceptance tests of 110240, 106000, and 100700 based on number of cohorts (1,5 and 10) respectively. The rationale for this decision is based on allowing more than one session for false reject testing and counting those as independent.

Table 5: Critical Errors (FA based on maximum number of tests with 5 cohorts).

Test	Target	Sigf	Ppass	e	Size	CE
FR	1.0%	25%	70%	2/3	1,080	8
FR	0.1%	25%	50%	1/2	1,386	0
FA	0.1%	5%	99%	2/3	105,065	88
FA	0.01%	5%	50%	1/2	105,131	5
FA	0.1%	25%	99%	2/3	105,517	98
FA	0.01%	25%	88%	1/2	96,845	7

4.2. Cohort Selection

Recently, a proposal to use cohort speakers provides an efficient method to use likelihood ratios (Equation 1) as a basis for verification [6]. In approximating the last quantity of the log likelihood ratio, Higgins suggests that this latter likelihood is dominated by the density of the nearest reference speakers, called cohorts. If the last expression is assumed to be dominated by the closest reference speaker then the maximum cohort log likelihood can be used for normalization. Furui [2] discusses several measures for cohort selection, each a potentional approximation to the last expression of the log likelihood. Some of these approximations include the logarithm of the summation of cohort likelihoods or the summation (average) of log likelihoods [3]. This latter method was for these experiments.

Creating cohort sets is accomplished in one of three ways, each a sorted list of "close" speakers. Define the *Difference of Means* log ratio as

$$d_{DOM}(\lambda_i, \lambda_j) \triangleq \log p(X|\lambda_i) - \log p(X|\lambda_j)$$

where X is all enrollment data for speaker i. Reynolds [8] provides a *Symmetric* distortion measure between two models using enrollment utterances from both the target and the potential cohort to determine similarity.

$$d_{SYM}(\lambda_i, \lambda_j) \triangleq \log \frac{p(X_i|\lambda_i)}{p(X_i|\lambda_j)} + \log \frac{p(X_j|\lambda_i)}{p(X_i|\lambda_j)}$$

These approaches are examples of a first order statistical analysis of the output distributions. Several researcher's have examined the issue of measuring "distances" between HMMs [9] for measuring model similarity. The goal then is to search for the set of cohort HMMs which are close to the claimant's HMM in some probabilistic distance. If enough training sequences from each speaker are evaluated against

each HMM, a distribution of log likelihoods begins to form, where a sample mean and variance can be extracted.

Higher order statistics can be used in conjunction with the *Bhattacharyya* distance for measuring the separability between the output distributions of a pair of HMMs [10]. The Bhattacharyya distance is derived from an analysis of determining an upper bound on the Bayes error rate of a two class problem. The form of this distance, for the 1-dimensional log likelihoods, is

$$d_B(\lambda_i, \lambda_j) \triangleq \frac{1}{4}\frac{(m_i - m_j)^2}{\sigma_i^2 + \sigma_j^2} + \frac{1}{2}\log\left(\frac{\frac{\sigma_i^2+\sigma_j^2}{2}}{\left(\sigma_i^2\sigma_j^2\right)^{\frac{1}{2}}}\right)$$

where m_i represents the enrollment mean and σ_i^2 represents the enrollment variance. The first term is a measure of the class separability due to the difference in the means while the second term is a measure of separability due to the variance difference.

To avoid statistical dependence between phrases with-in each verification session, all 4 combination phrases are taken as a test sample. Standard procedure is not performing inter-gender tests or testing with cohort speakers, where the number of cohorts is 10. Figure 3 demonstrates the effectiveness of the Bhattycharyya distance when used in conjunction with the log ratio normalization compared to other methods.

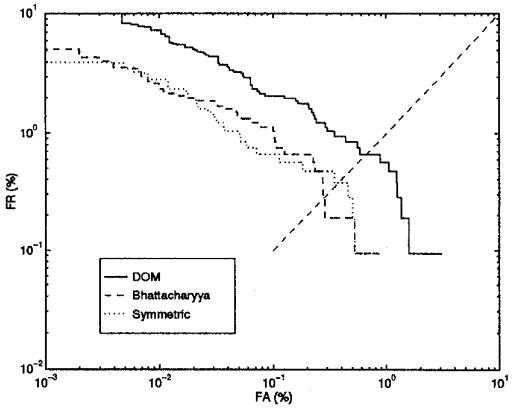

Figure 3: Verification error rates(%) using DOM, Bhattycharyya and Symmetric cohort selection.

5. CONCLUSION

This effort examined speaker dependent monophones for verification using the YOHO database. Initial speaker identification examined the effect of transitional features on speaker HMM modeling indicating, like speech recognition, first and second order regression coefficients contain useful information. In addition, estimation of the covariance matrices, while bootstrapping from speaker independent models, shows better performance with shared covariance models, especially with fewer combination test phrases.

A critical error analysis was conducted by first defining a hypothesis test with a significance level (Type I error) of 5%. This allowed a recalculation of the False Accept critical errors as previously published to account for more imposter tests and a more constrained passing of unacceptable systems. A forced Viterbi alignment was used in conjunction with cohort normalization schemes to remove the greatly varying effects of the transcription itself. Three cohort selection schemes were tested showing the best equal error rate of 0.282% using a second order Bhattacharyya distance.

6. REFERENCES

[1] J. P. Campbell, "Testing with the YOHO CD-ROM voice verifcation corpus," in *ICASSP*, pp. 341–344, May 1995.

[2] S. Furui, "An overview of speaker recognition technology," in *ESCA Workshop on Automatic Speaker Recognition, Identification, Verification*, pp. 1–9, April 1994.

[3] T. Matsui and S. Furui, "Similarity normalization method for speaker verification based on a posteriori probability," in *ESCA Workshop on Automatic Speaker Recognition, Identification, Verification*, pp. 59–62, April 1994.

[4] A. E. Rosenburg, C.-H. Lee, and F. K. Soong, "Subword unit talker verification using hidden Markov models," in *ICASSP*, vol. 1, pp. 269–272, 1990.

[5] Entropic Research Laboratory, Inc., Washington,DC, *HTK - Hidden Markov Model Toolkit*, v1.5 ed., 1993.

[6] A. Higgins, L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting," *Digital Signal Processing*, vol. 1, pp. 89–106, 1991.

[7] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York N.Y.: McGraw-Hill, Inc., 3rd ed., 1991.

[8] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[9] B. H. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden markov models," *AT&T Technical Journal*, vol. 64, pp. 391–408, February 1985.

[10] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic Press, second ed., 1990.