CHANNEL AND NOISE COMPENSATION FOR TEXT DEPENDENT SPEAKER VERIFICATION OVER TELEPHONE

William Y. Huang[†‡]

[†]ITT Aerospace Communications
10060 Carroll Canyon Rd.
San Diego, CA 92131
wyhuang@raman.ucsd.edu

Bhaskar D. Rao[‡]

[‡]Dept. of Elec. and Comp. Eng.
University of California, San Diego
La Jolla, CA 92093
brao@ucsd.edu

ABSTRACT

The performance of text dependent, short utterance speaker verification systems degrades significantly with channel and background artifacts. We investigate maximum likelihood and adaptive techniques to compensate for a stationary channel and noise. Maximum likelihood channel and noise compensation was introduced by Cox and Bridle in 1989, and has been shown to be effective in many other speech applications. For adaptive estimation, a Bussgang like algorithm is developed which is more suitable for real-time implementation. These techniques are evaluated on a speaker verification system that uses the nearest neighbor metric. Our results show that for telephone speech with channel differences, channel compensation can provide substantial performance improvement. For un-cooperative speakers, background compensation resulted in a 35% improvement.

1. INTRODUCTION

This paper addresses the problem of short utterance, text dependent speaker verification over telephone. It is assumed that performance degradation is due to a stationary channel and/or background artifact. This is a reasonable assumption since the utterance durations considered here are on the order of two seconds. This paper examine solutions to this problem using maximum likelihood and adaptive techniques. Maximum likelihood channel and background adaptation was introduced by Cox and Briddle [2], and has been shown to be effective in unsupervised microphone adaptation for speech recognition [7], keyword spotting [10], speech recognition over telephone [9], and text independent speaker identification [11]. However, maximum likelihood methods are iterative and difficult to implement in real time. Therefore, adaptive techniques are also explored. In particular, Bussgang algorithms have been applied successfully to communication [1] and seismic signal processing [4]. Both the maximum likelihood and the adaptive techniques require a statistical model for speaker independent speech. The model used here is a speaker independent HMM trained from the first 50 speakers of the YOHO corpus [5]. In Sec. 2 we describe the speaker verification system, as well as our model for channel/noise corruption of the data. Sec. 3 describes the maximum likelihood estimation of the channel and noise corruption vectors. Sec. 4 describes the adaptive technique applied to this problem. Sec.'s 5 and 6 discuss the database used and the experimental results. Sec. 7 contains our conclusions

2. DESCRIPTION OF THE VERIFIER

The preprocessing for the speaker verification system consists of a mel–scaled filterbank followed by an estimate of the Olano observations given by [8],

$$O_i[t] = \sqrt{\frac{\sqrt{p_i[t]}}{\sum_j \sqrt{p_j[t]}}}. \tag{1}$$

p_i is the spectral power from the i-th component of the filterbank, and $O_i[t]$ is the i-th component of the Olano observation at time frame t.

The nearest neighbor speaker verification algorithm used here is based on Higgin's system [6]. Let $T_j[i]$, $i = 1...N_{\text{train}}$ be frames of speech from speaker j, and $U[i]$, $i = 1...N_{\text{test}}$ be frames of a test utterance. Both the training and test utterances are segmented into speech classes via Viterbi segmentation using an HMM. The nodes of the HMM define the classes. The forward score between the test utterance and speaker j is

$$F_j = \frac{1}{N_{\text{test}}} \sum_{k=1}^{N_{\text{test}}} \min_{\{i:\mathcal{E}(i,k)\}} d(T_j[i], U[k])$$

0-7803-2431-5/95 $4.00 © 1995 IEEE

where $d(\cdot,\cdot)$ is the squared Euclidean distance operator, and $\mathcal{E}(i,k)$ is the *event* that frame k of the test utterance (U) and frame i of T_j have the same HMM node class. The backward score is

$$B_j = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \min_{\{k:\mathcal{E}(i,k)\}} d(T_j[i], U[k])$$

The distance score between the test utterance and speaker j training utterance is $S_j = F_j + B_j$ and the speaker identity is chosen according to

$$\hat{j} = \arg\min_j S_j$$

Other details regarding scores normalization, frames selection, and so forth are discussed in [6].[1]

Our model for stationary noise and channel corruption at the t-th time window is as follows:

$$X_t(\omega) = H(\omega)S_t(\omega) + N(\omega) \qquad (2)$$

where $X_t(\cdot)$ is the acquired speech, $S_t(\cdot)$ is the uncorrupted speech, $H(\cdot)$ and $N(\cdot)$ are stationary channel and noise artifacts. Let

$$c_i =< H(\omega), F_i(\omega) >^2, \quad \text{and} \quad n_i =< N(\omega), F_i(\omega) >^2$$

where $< \cdot,\cdot >$ is the inner product operator, and $F_i(\omega)$ is the filter that corresponds to the $i-th$ component of the filterbank. Our model for the corrupted Olano observation is

$$O_i[t] = \sqrt{\frac{\sqrt{c_i p_i[t] + n_i}}{\sum_j \sqrt{c_j p_j[t] + n_j}}} \qquad (3)$$

Other details: Sampling rate is 8kHz. Frame rate is 50Hz. Filterbank window size is 32msec. 14 filters per frame.

3. MAXIMUM LIKELIHOOD ESTIMATION

Our goal is to estimate the channel and noise compensation vectors, c_i's and n_i's, from the corrupted Olano observations of Eq. 3 via maximum likelihood; i.e., find

$$\{\hat{\underline{c}}, \hat{\underline{n}}\} = \arg\max_{\underline{c},\underline{n}} f(\mathbf{O}|\underline{c}, \underline{n}) \qquad (4)$$

where $\mathbf{O}$ is the set of Olano observations for all times and components, and $\{\underline{c}, \underline{n}\}$ are the channel and noise compensation vectors (c_i's and n_i's). For this short utterance text dependent task, the probabilistic model we use is an HMM, and $\underline{c}, \underline{n}$ is estimated using Generalized Expectation Maximization (GEM) [3].[2] Define the EM functional

$$Q(\underline{c}; \underline{c}') = \sum_{\mathbf{q}} p(\mathbf{q}|\mathbf{O}, \underline{c}') f(\mathbf{q}, \mathbf{O}|\underline{c}) \qquad (5)$$

where $\underline{c}'$ is the estimated channel at the previous iteration, $\mathbf{q}$, the "hidden" variable, is a sequence of states in the Baum–Welsch algorithm. Ignoring the initial probabilities and the transition probabilities, we have

$$f(\mathbf{O}, \mathbf{q}|\underline{c}, \underline{n}) = \prod_{t=1}^{T} f(\underline{O}[t]|\underline{c}, \underline{n}, q_t)$$

where $\underline{O}[t]$ is the vector of Olano components at time t, q_t is the HMM state at time t corresponding to $\mathbf{q}$, and the observation probability is Gaussian with a diagonal covariance matrix. Under the Gaussian model, the above equations become

$$Q(\underline{c}, \underline{c}') = \mathcal{C}_1 + \mathcal{C}_2 \sum_{t,s} P_{t,s} \sum_{i=1}^{14} \frac{1}{\sigma_i^2} (O_i[t] - \mu_{s,i})^2 \qquad (6)$$

where i is an index to the components of the Olano observation, σ_i is the (i,i)-th element of the global diagonal covariance matrix, $\mu_{s,i}$ is the Gaussian mean for HMM state s component i, $\mathcal{C}_1$ and $\mathcal{C}_2$ are constants, and $P_{t,s}$ is $p(q_t = s|\mathbf{O}, \underline{c}')$, computed using the forward backward algorithm. Substituting the distorted Olano observations of Eq. 3 for the observation in Eq. 6, we obtain the following gradients for channel and noise updates:

$$\begin{aligned} &\frac{\partial}{\partial n_i} Q(n_i; n_i') \propto \\ &\quad \sum_{s,t} P_{t,s} \left(\sum_k \frac{1}{\sigma_k^2} (O_k[t] - \mu_{s,k}) \frac{\mu_{s,k}}{\mu_{i,k}} \right. \qquad (7) \\ &\quad \left. - \frac{1}{\sigma_i^2} (O_i[t] - \mu_{s,i}) \frac{1}{\mu_{s,i}^3} \right) \end{aligned}$$

Since $\mathbf{O}$ is compensated at each iteration, we used the simplifying assumption that $n_i \approx 0$ and $c_i \approx 1$ in the above. For channel compensation, we again assume that $n_i \approx 0$ at each iteration and simplify Eq. 3 to

$$O_i[t] = \sqrt{\frac{\tilde{c}_i \tilde{p}_i[t]}{\sum_j \tilde{c}_j \tilde{p}_j^t}}$$

[1] For scores normalization, Higgins used test samples to estimate normalization factors [6]. Here, we estimate normalization factors from training data and apply them to the test speech without further modification.

[2] The GEM algorithm modifies the EM algorithm by using gradient descent instead of maximum likelihood in the M–step (of the EM algorithm).

where $\tilde{p}$ and $\tilde{c}$ are the squares of their corresponding components in Eq. 3 and solve

$$\begin{aligned} &\tfrac{\partial}{\partial \tilde{c}_i} Q(\tilde{c}_i; \tilde{c}'_i) \propto \\ &\quad \tfrac{1}{\tilde{c}_i} \textstyle\sum_{s,t} P_{t,s} \left(- \sum_k \left(\tilde{O}_k[t] - \tilde{\mu}_{s,k} \right) \tilde{\mu}_{s,k} \tilde{\mu}^2_{s,i} \sigma^2_i \right. \\ &\quad \left. + \left(\tilde{O}_i[t] - \tilde{\mu}_{t,i} \right) \tilde{\mu}_{s,i} \left(\textstyle\sum_j \tilde{\mu}^2_{s,j} \sigma^2_j \right) \right) \end{aligned} \tag{8}$$

where

$$\begin{aligned} \tilde{O}_k[t] &= O_k[t]/\sigma^2_k \\ \tilde{\mu}_{s,k} &= \mu_{s,k}/\sigma^2_k \end{aligned}$$

The validity of the above technique has been verified from synthetic data experiments.

4. ADAPTIVE ESTIMATION

The adaptive procedure implemented for this problem was motived by the Bussgang algorithm [1]. The update equation is

$$c_i = c'_i + \delta \left(\frac{\hat{Y}_i}{Y_i} - W \right) \tag{9}$$

where c'_i is the current channel estimate, δ is the gradient descent parameter, Y is the channel compensated Olano input (via Eq. 3), $\hat{Y}_i$ is the MMSE estimate of the uncorrupted speech, and W, defined below, is approximately 1. A real–time approximation for $\hat{Y}_i$ is determined during recognition by using node probabilities ($\hat{P}(n)$'s) computed during recognition, and then applying the heuristic that probabilities below 0.2 are discarded.[3] I.e.,

$$\underline{\hat{Y}} = \frac{\sum \mu_n \hat{P}(n)}{\sum \hat{P}(n)} \tag{10}$$

where μ_n is the mean vector for node n, $\hat{P}(n)$ is described above, and all summations take place over the event $\hat{P}(n) > 0.2$. W in Eq. 9 is $\sum \hat{P}(n)$ over all $P(n)$'s greater than 0.2. The deconvolved speech vectors after a few milliseconds are used for deconvolution.

5. DESCRIPTION OF THE CNORM AND DEMO_RV1 DATABASE

The vocabulary and grammar used for this text dependent task is that of the "combination lock" phrases consisting of two pairs of numbers. A example phrase is 46-79, pronounced "forty six seventy nine." Enrollment samples from the first 50 speakers in the narrowband YOHO database [5][4] are used to train a speaker independent, Gaussian mixture observation HMM. The vocabulary of the test database is further limited to 8 digits "forty", "four", "sixty", "six", "seventy", "seven", "ninety" and "nine".

Table 1: Supervised verification results, in terms of EER (equal error rate) of the ROC curve, for both tasks. Results are shown for baseline, max. likelihood (ML) channel deconvolution and background compensation.

	CNORM	*DEMO_RV1*
baseline	21.7%	4.6%
ML chan decon	3.4%	3.3%
ML bkg comp	7.0%	3.0%

The first test database is the "CNORM" database. It consists of 12 speakers speaking over 4 phones with different channel characteristics. 8 of the speakers had 2 test sessions each on 2 different, randomly selected phones with 2 phrases per test session. These 8 speakers are adjuncts; i.e. they have no enrollment sessions, and are used to generate false alarms only. The other 4 speakers had 2 enrollment and 4 test sessions on each of the 4 phones, with 4 phrases per enrollment session. When testing speaker A from phone X, enrollment sessions for speaker A phone X are not used, but other speakers' phone X enrollments are used. Thus, successful verification requires overcoming channel differences.

The second database, dubbed "DEMO_RV1," consists of 23 speakers calling from different phones. Several speakers in this database knew each other, and were encouraged to disguise their voices. They were given a pay bonus each time they successfully fooled the verifier to accept their voice for one of their co–workers. Each of the 23 speakers had 2 enrollment sessions of 4 phrases each, spoken normally and acquired from the same phone. There is an average of 17 test session per speaker, ranging from a minimum of 1 test session to a maximum of 153 test sessions.

6. EXPERIMENTAL RESULTS

The result of applying the above algorithm to the CNORM and DEMO_RV1 database is shown in Tab. 1. For the CNORM database, which is designed to test channel artifacts, channel deconvolution decreased the

[3] Although we did not perform an exhaustive experimental search for the optimal threshold value, our experiments indicate that not using a threshold (i.e. a threshold of 0) did not work.

[4] The YOHO database consists of 106 males, 32 females, same channel, 4 enrollment sessions per subject with 24 phrases per session, 10 test sessions per subject with 4 phrases per test session.

Table 2: Unsupervised verification results on the CNORM task

baseline	23.3%
blind deconvolution	8.7%
adaptive (Bussgang–like)	4.7%
ML chan decon	3.6%
ML bkg subt	7.8%

EER from 21.7% to 3.4%. Background compensation actually can have a significant effect on channel artifacts. As a result, for the CNORM database, background compensation improved the EER from 21.7% to 7.0%.

In the DEMO_RV1 database, it is not known which of channel and/or noise is a more significant factor. Neither channel nor background compensation provides a good model for non–cooperative speakers. Here, channel deconvolution improved the EER only to 3.3% (from a baseline of 4.6%), while background compensation improved the EER to 3.0%.

Text independent techniques on CNORM are of interest as they are indicative of performance expected in a constrained text independent task. Also, in real application, users may not be careful to repeat their prompted phrases. The relevant text independent techniques include conventional blind deconvolution (i.e. channel/noise estimated via long term spectral averages) and the adaptive techniques described in Sec. 4. Also, all the above text dependent techniques can become text independent if one uses a loose grammar instead of forced decoding to the target phrase. As Tab. 2 shows, repeating the text dependent techniques in the text independent mode results in only a slight performance degradation. This is probably because the recognition rate is high for this constrained grammar task. Also, conventional blind deconvolution provides fairly good performance in this situation (8.7%). Note that the adaptive scheme provided 4.7% performance, which is better than blind deconvolution but not at the 3.4% level achieved using maximum likelihood channel deconvolution. We note, however, that the computational complexity of the adaptive method is similar to that of the blind deconvolution, and is much smaller than that of the maximum likelihood techniques.

7. DISCUSSIONS

The results here show that channel deconvolution for short utterances text dependent speaker verification can be addressed by maximum likelihood estimation of a stationary channel vector using a speaker independent HMM as a statistical model. This result extends to the text independent case. However, this might not work as well in problem domains where the difficulty is not necessarily restricted to the channel. Finally, we also showed that an adaptive technique works better than blind deconvolution in the text independent mode of this task, although performances are not as good as that provided by maximum likelihood. The adaptive technique explored here is analogous to the Bussgang algorithm for blind equalization, and is amenable to real time implementation.

8. REFERENCES

[1] S. Bellini. Bussgang techniques for blind deconvolution and equalization. In S. Haykin, editor, *Blind Deconvolution*, pages 8–59. Prentice Hall, Englewood Cliffs, New Jersey, 1994.

[2] S. J. Cox and J. S. Bridle. Unsupervised speaker adaptation by probabilistic spectrum fitting. *ICASSP 89*, pages 294–7, May 1989.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[4] D. Donoho. On minimum entropy deconvolution. In D. F. Finley, editor, *Applied Time Series Analysis II*, pages 565–608. Academic Press, 1981.

[5] A. Higgins, L. Bahler, and J. Porter. Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1(2):89–106, 1991.

[6] A. L. Higgins, L. G. Bahler, and J. E. Porter. Voice identification using nearest–neighbor distance measure. In *ICASSP 93*, pages II–375–378, 1993.

[7] F.-H. Liu, R. M. Stern, A. Acero, and P. J. Moreno. Environment normalization for robust speech recognition using direct cepstral comparison. In *ICASSP94*, pages II–61–4, 1994.

[8] A. C. Olano. An investigation of spectral match statistics using a phonetically marked data base. In *ICASSP 83*, 1983.

[9] M. G. Rahim and B.-H. Juang. Signal bias removal for robust telephone based speech recognition in adverse environments. In *ICASSP 94*, pages I–445–8, 1994.

[10] R. C. Rose and D. B. Paul. A hidden Markov model based keyword recognition system. In *ICASSP 90*, pages 129–32, 1990.

[11] R. C. Rose and D. A. Reynolds. Text independent speaker identification using automatic acoustic segmentation. In *ICASSP 90*, pages 293–296, 1990.