THE EFFECTS OF HANDSET VARIABILITY ON SPEAKER RECOGNITION PERFORMANCE: EXPERIMENTS ON THE SWITCHBOARD CORPUS *

Douglas A. Reynolds

Lincoln Laboratory, Massachusetts Institute of Technology
244 Wood Street
Lexington, MA 02173–9108, USA
Voice: (617) 981-4494 Fax: (617) 981-0186
E-mail: DAR@SST.LL.MIT.EDU

ABSTRACT

This paper presents an empirical study of the effects of handset variability on text-independent speaker recognition performance using the Switchboard corpus. Handset variability occurs when training speech is collected using one type of handset, but a different handset is used for collecting test speech. For the Switchboard corpus, the calling telephone number associated with a file is used to imply the handset used. Analysis of experiments designed to focus on handset variability on the SPIDRE database and the May95 NIST speaker recognition evaluation database, show that a performance gap between matched and mismatched handset tests persists even after applying several standard channel compensation techniques. Error rates for the mismatched tests are over 4 times those for the matched tests. Lastly, a new energy dependent cepstral mean subtraction technique is proposed to compensate for nonlinear distortions, but is not found to improve performance on the databases used.

1. INTRODUCTION

In [1], we examined speaker recognition performance losses associated with various telephone transmission degradations using the TIMIT and NTIMIT corpora. One effect which could not be examined with these corpora, however, was that of handset variability[1]. Handset variability occurs when training speech is collected using one type of handset, but a different handset is used for collecting test speech. In many telephone based speaker recognition applications, handset variability is likely to be encountered and can adversely affect system performance by modifying a person's voice in test utterances in ways not represented in the training data. Many successful compensation techniques, such as cepstral mean subtraction [2], RASTA filtering [3] and delta coefficients [4], have been proposed and evaluated for mitigating the performance loss due to handset mismatches [5]. However, most of these compensations remove only first-order linear filter effects and, as shown in the experiments below, even after compensation, a performance gap still exists between the matched and mismatched conditions. This implies that there are other effects, beyond simple linear filtering, induced by the different handsets which are not being adequately removed.

In this paper, we present an empirical study of the effects of handset variability on text-independent speaker recognition performance using the Switchboard corpus [6]. The coded version of the caller's telephone number available for each conversation is used to imply the handset used. Conversations originating from identical telephone numbers can generally be assumed to be over the same telephone handset. Similarly, it may be assumed that conversations originating from different telephone numbers are over different handsets[2]. In the first part of the study, we present an analysis of results from experiments on the SPIDRE database[3] designed to explicitly focus on the matched and mismatched handset conditions between training and testing utterances. The second part of the study is an analysis of results form the May 1995 NIST Switchboard Speaker Evaluation database with respect to handset variability. We apply several standard channel compensation techniques and show that the matched-mismatched performance gap still persists. Lastly, a new channel compensation techniques aimed at addressing potential non-linear effects is described and applied to the May95 evaluation database.

The remainder for the paper is organized as follows. The next section briefly describes the recognition system used. This is followed by a description of the SPIDRE database experiments and analysis of results. Section 4 describes the May95 speaker recognition evaluation database and presents results emphasizing the matched-mismatched conditions. A new energy dependent cepstral mean subtraction compensation technique is then described in Section 5.

*THIS WORK WAS SPONSORED BY THE DEPARTMENT OF THE AIR FORCE. THE VIEWS EXPRESSED ARE THOSE OF THE AUTHORS AND DO NOT REFLECT THE OFFICIAL POLICY OR POSITIONS OF THE U.S. GOVERNMENT.

[1]More precisely, transducer variability.

[2]Neither assumption is strictly true, since callers can use different telephone units with the same telephone number and similar telephone units can be used at different telephone numbers. Furthermore, there are other factors, such as different transmission paths and acoustic environments, which also change with same or different telephone numbers. For this study we will assume that telephone number variability implies handset variability.

[3]SPIDRE is a subset of the Switchboard database available through the LDC.

0-7803-3192-3/96 $5.00©1996 IEEE

2. RECOGNITION SYSTEM

Speaker recognition was performed using Gaussian mixture speaker models [7]. In this system, each speaker's acoustic parameter distribution is represented by a speaker dependent Gaussian mixture model (GMM),

$$p(\vec{x}_t|\lambda_s) = \sum_{i=1}^{M} p_i^s \; b_i^s(\vec{x}),$$

with mixture weights p_i^s and Gaussian densities $b_i^s(\vec{x})$. Maximum likelihood parameters are estimated using the EM algorithm. Mel-cepstra and delta-cepstra features are used as acoustic observations in the experiments.

The average log-likelihood of a model given an utterance $X = \{\vec{x}_1 \ldots \vec{x}_T\}$ is computed as

$$\mathcal{L}(X|\lambda_s) = \frac{1}{T}\sum_{t=1}^{T} \log p(\vec{x}_t|\lambda_s)$$

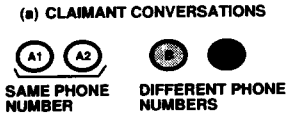
For closed-set identification, the speaker associated with the most likely model for the input utterances is chosen as the recognized speaker; $\hat{s} = \arg\max_s \mathcal{L}(X|\lambda_s)$. For verification, a likelihood ratio score between the claimed speaker's model likelihood score and the average likelihood score of a background set is used [8].

$$\Lambda(X|s) = \mathcal{L}(X|\lambda_s) - \log \sum_b \exp\{\mathcal{L}(X|\lambda_b)\}.$$

In the experiments, a speaker's background set was not selected in any special way; it merely consisted of the other claimant speaker models available for the test. It should be noted, however, that careful selection of background sets can greatly improve verification performance [7]

3. SPIDRE DATABASE EXPERIMENTS

The SPIDRE database is a subset of the Switchboard corpus, created for speaker recognition research. SPIDRE consists of 4 conversation halves each from 45 claimant speakers (27 male, 18 female) and 200 conversations total from 160 imposter speakers (82 male, 78 female). Conversation halves have approximately two minutes of speech total. The 4 conversations from each claimant originate from 3 different phone numbers (handsets) with 2 conversations from the same phone number (see Figure 2 (a)).



(b) DATA SPLITS

	TRAIN	TEST
SPLIT1	A1	A2
SPLIT2	A1	A2

Figure 2. Claimant conversations in SPIDRE database. (a) Association of telephone numbers with conversations. (b) Train/test data splits of conversations.

To examine handset variability between train and test conversations, the claimant conversations were divided into two train/test splits (Figure 2 (b)). In split1, we trained on one conversation and had one matched test (test handset used in train data) and one mismatched test (test handset not used in train data). This represents the most severe handset variability scenario in which we have only a single session of training speech from a single handset. In split2, we trained on two conversations and again had one matched and one mismatched test. The extra training conversation was added to introduce session variability and equal amounts of data were extracted from each conversation for training (the total amount being the same as used from the single conversation in split1).

Speaker models of order 32 were trained using the first 60 seconds of claimant speaker speech found in each training conversation. For split2, the first 30 seconds from each training conversation was used for training speaker models. The first 15 seconds of speech from each test conversation (both claimant and imposter speakers) was used for test utterances. Observation vectors consisted of mel-cepstra features extracted over the entire 4 kHz band, processed through a RASTA channel equalization filter.

Matched, mismatched and combined results for both identification and verification for the two splits are shown in Figure 1. The verification results are obtained by averaging the false rejection rate over all claimant speakers for a constant false acceptance rate. There are two points to make about these results. First, it is clear that, despite using RASTA channel equalization, there persists a large gap in performance between the matched and mismatched tests. The RASTA filtering does greatly improve performance over no channel compensation, but apparently there remains some uncompensated effect. Cursory inspection of the mismatched tests found no obvious degradations in the signal waveform (such as clipping) but noticeable audible differences from the different handsets were apparent. Figure 3 compares long term averaged log mel-filterbank spectra from two utterances by the same speaker over different handsets. In this example there are obvious mismatches in bandwidth and spectral dynamic range between the different handsets.

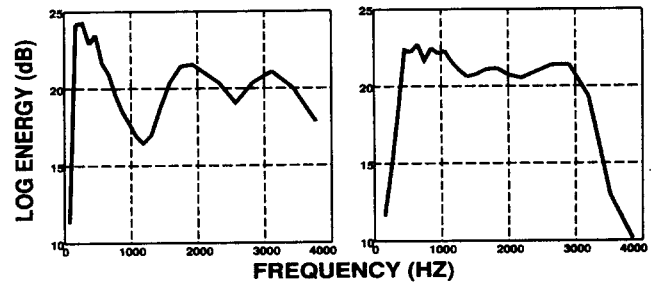


Figure 3. Comparison of long term averaged log mel-filterbank spectra from two utterances by the same speaker over different handsets.

Second, comparing split1 and split2 performance, we see that the addition of the second training utterances, even from a handset not used in the test utterances, improves the mismatched performance. It is not clear how much of this improvement is due to the addition of session variability to the training data as opposed to the additional of handset

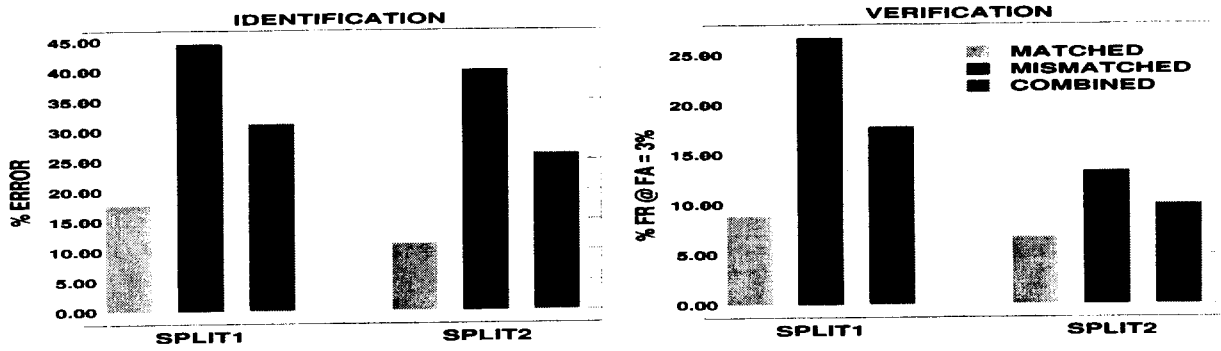


Figure 1. Identification and verification results for two data splits on SPIDRE database. Results are shown for matched, mismatched and combined tests. (FR="False Rejections", FA="False Acceptances")

diversity. The extra session may be acting to "broaden" the speaker model so that it is not too tuned to a particular handset.

4. MAY95 EVALUATION DATABASE EXPERIMENTS

The second database we present results on is from the May 1995 NIST administered speaker recognition evaluation. The evaluation database, derived from the Switchboard corpus, contains 26 claimant speakers (15 male, 11 female) and 80 imposter speakers (33 male, 47 female). There are four conversations designated for training for each claimant and around six conversations per claimant and imposter speaker designated for testing. The claimant's training and testing conversations were selected to introduce a large number of handset (phone number) mismatches. For all but one of the claimant speakers, the four training conversations came from a single handset, so there was almost no handset diversity in the training data. Out of a total of 141 claimant test conversations, 57% (81) were from handsets not used in the speaker's training conversations and 43% (60) were from a matched training handset.

Within each conversation, segments containing speech from a single speaker were designated for specific training and testing conditions. There are three training conditions consisting of 10 seconds of speech drawn from one training conversation (TRAIN:10s), 30 seconds of speech drawn equally from three training conversations (TRAIN:30s), and unlimited amounts of speech drawn from four training conversations (TRAIN:UNL). There are also three testing conditions of nominal durations of 5, 10 and 30 seconds, denoted TEST:5s, TEST:10s, and TEST:30s, respectively. To minimize the confounding effects due to limited train and test utterance durations and focus on the handset effects, we present results using the (TRAIN:UNL, TEST:30s) conditions of the evaluation. There are a total of 155 claimant tests and 428 imposter tests for the TEST:30s condition.

The baseline system used in these experiments used 64 mixtures per claimant speaker trained on all four training conversations with mel-cepstra observations. In Table 1 we show the effect of applying several standard channel compensation techniques to the feature stream in order to boost the mismatched performance. These include RASTA filtering (rasta), cepstral mean subtraction (cms), using bandlimited filterbank outputs (bl,300-3100Hz), appending delta cepstra coefficients (cep+dcep), appending a pitch estimate (cep+dcep+pit), and some of their combinations. As seen, each compensation, with the exception of appending pitch, decreases the error rate, but mostly for the mismatched tests. The matched tests have low error to begin with and are relatively unaffected. Even after applying all the compensations, however, the performance gap still persists. The last line in the table (cep+dcep-bl-cms+rasta-bkg), shows the best performing system, in which we select claimant specific background speakers for likelihood normalization, has all errors occurring in the mismatched tests.

Examining the test likelihood scores shows that claimant speaker models are producing significantly lower likelihood scores for the mismatched tests than for the matched tests. For closed set identification this causes spurious mismatches to other speaker models and for verification it causes false rejections by producing low likelihood ratio scores. The fact that the mismatch models are scoring poorly, indicates the claimant speaker model is being overly tuned to the training handset.

5. ENERGY DEPENDENT CEPSTRAL MEAN SUBTRACTION

Many of the standard channel compensation techniques are based on a linear filter model of the channel, thus they remove only first order linear filtering effects. While removal of these linear filtering effects are quite effective, the fact that a performance gap still remains after compensation, indicates there are other effects imparted by the handset not being removed.

Handset microphones, especially carbon button transducers, are well known to produce nonlinear distortions on speech. A simple, piece-wise linear model of the nonlinear distortion is to assume the speech is passed through an energy dependent linear filter[4]. That is, for different input signal energy levels, the microphone has a different frequency response. This effect was illustrated in [1] figure 3. An approach to removing these energy dependent chan-

[4]This idea was originally proposed to the author by Marc Zissman. Any implementation/experimental errors are sole property of the author.

Table 1. Results of applying standard channel compensation techniques to May95 evaluation database. cep+dcep="appended delta cep", cep+dcep+pit="appended pitch", cms="cepstral mean subtraction", rasta="RASTA filtering", cms+rasta="rasta followed by cms", bl="band limiting", bkg="speaker specific background sets".

FEATURE	Identification (% error)			Verification (% FR @ FA=3%)		
	combined	matched	mismatched	combined	matched	mismatched
cep-rasta	22	5	34	23	5	36
cep-cms	19	5	28	21	5	32
cep-bl-cms	15	5	23	18	5	26
cep+dcep-bl-cms	14	5	21	16	5	24
cep+dcep+pit-bl-cms	15	5	22	19	6	27
cep+dcep-bl-cms+rasta	13	5	18	12	3	17
cep+dcep-bl-cms+rasta-bkg	9	0	12	5	0	9

Table 2. Results of applying energy dependent cepstral mean subtraction to the May95 evaluation database. cms10="10 level cms", cms5db="5dB spacing cms"

FEATURE	Identification (% error)			Verification (% FR @ FA=3%)		
	combined	matched	mismatched	combined	matched	mismatched
cep+dcep-bl-cms	14	5	21	16	5	24
cep+dcep-bl-cms10	15	5	22	17	5	25
cep+dcep-bl-cms5dB	16	5	24	17	5	26

nel responses from the speech is to subtract cepstral means computed over different energy ranges. For an input feature stream $\{\vec{x}_1, \ldots, \vec{x}_T\}$ and corresponding frame energies $\{e_1, \ldots, e_T\}$, we compute L energy dependent mean vectors as

$$\vec{m}_l = \frac{1}{T_l} \sum_{\{t: E_l \le e_t < E_{l+1}\}} \vec{x}_t, \quad l = 1, \ldots, L,$$

where $\{E_l\}$ are the divisions between energy levels. Mean vector $\vec{m}_l$ is then subtracted from the frames corresponding to level l.

The energy dependent cepstral mean subtraction was applied to the May95 evaluation database. Two methods were used to set the energy levels. In the first method, energy ranges were set to correspond to each decile of the cumulative energy histogram for the input utterance, giving 10 ranges for each utterance. In the second method, the energy levels were equally spaced 5 dB apart starting from the maximum frame energy in the input utterance, giving a variable number of ranges per utterance. The results are shown in Table 2 along with the results using standard cepstral mean subtraction. Unfortunately, the energy dependent cms did not help the mismatched test and actually dropped performance slightly. It may be that the nonlinear distortions are more complex than piece-wise filters, such as airflow-transducer interactions or spectral dynamic range compression, as discussed in [9],

6. CONCLUSION

This paper has presented some experiments on subsets of the Switchboard corpus focused on the effects of handset variability on speaker recognition performance. While many of the linear filter compensation techniques do indeed improve performance under mismatched handset conditions, the fact that a performance gap between matched and mismatched handset conditions persists points to other non-compensated effects. Before more effective compensation techniques can be developed, a better understanding of the acoustic characteristics and speech transformations imposed by different handset microphones needs to be obtained. Using the phone numbers provided with Switchboard conversations is adequate for examining the effects of handset mismatches on speaker recognition system performance, but, ideally, a more controlled handset database needs to be used to truly address the effects of handset variability.

REFERENCES

[1] D. A. Reynolds *et al.*, "The effects of telephone transmission degradations on speaker recognition performance," ICASSP, pp. 329–332, May 1995.

[2] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," JASA, vol. 55, pp. 1304–1312, June 1974.

[3] H. Hermansky *et al.*, "RASTA-PLP speech analysis technique," ICASSP, pp. I.121–I.124, March 1992.

[4] F. Soong and A. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," Trans. ASSP, vol. ASSP-36, pp. 871–879, June 1988.

[5] A. Rosenberg, C.-H. Lee, and F. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification," ICSLP, pp. 1835–1838, 1994.

[6] J. J. Godfrey, E. C. Holliman, and J. MacDaniel, "Switchboard: Telephone speech corpus for research and development," ICASSP, pp. I–517–I–520, March 1992.

[7] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, p. 91–108, August 1995.

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

[9] A. Potamianos, L. Lee, and R. Rose, "A feature space transformation for telephone based speech recognition," Eurospeech, pp. 1533–1536, September 1995.