92-18

ICASSP-93

Speech Processing

Volume II of V

1993 IEEE International Conference on Acoustics, Speech, and Signal Processing

April 27-30, 1993
Minneapolis Convention Center
Minneapolis, Minnesota, USA

93CH3252-4

Robustness Study of Free-Text Speaker Identification and Verification

Yu-Hung Kao[*] John S. Baras[†] P. K. Rajasekaran
Texas Instruments Incorporated
Dallas, TX 75265

Abstract

Usable free-text speaker identification and voice verification systems must exhibit robustness under varying operational conditions. We study the degree of robustness provided by various signal processing techniques [1] [2] [3] by experimenting on a widely used long distance telephone data base [4] [5] [6]. This data base consists of data recorded at two different sites, with data from one site much poorer in quality than the other; further, the recording equipment had been inadvertently changed for the later half of the sessions resulting in a significantly changed environment. Our study identifies the combination of techniques that provide consistent and significant improvements; our results surpass other published results [4] [5] [6] on the same task. Specifically, in the task of identifying 16 speakers, with training data from the recording prior to equipment change and testing on data from a set after the change (the most challenging condition), we obtain a correct identification rate of 87.5% with an average rank of 1.12; [4] obtains the hitherto best result of 75% correct identification with an average rank of 1.56; without any robustness processing, the rate was only 12%. Detailed results on exhaustive experimentation are presented along with appropriate discussions.

1 Introduction

This paper presents the results of experimental investigations of several environmental robustness algorithms on the free-text speaker identification and verification tasks. The following algorithms (with slight modification to fit the speaker identification context) were studied and extensive experiments were performed on the "King" database [4] [5] [6]:

1. ISDCN (Interpolated SNR Dependent Cepstral Normalization) [1].
2. Bandpass liftering [2].
3. RASTA-PLP [3].
4. RASTA [3].
5. Bandpass liftering and RASTA.

ISDCN provided only marginal improvement in performance, and involved a high degree of computational complexity. RASTA-PLP, as described in [3], did not yield any gain in our experimentation. However, by using a modified version of RASTA-PLP, called RASTA (operating in the cepstral domain instead of the perceptual spectral domain as in [3]), provided significant performance improvements. Combining bandpass liftering with RASTA resulted in the best overall performance. Experimental results on ISDCN and RASTA-PLP are not presented in this paper for the sake of brevity.

2 Algorithm

The front-end of our system consists of extracting 20 cepstral coefficients from a 14-th order LPC analysis (20 ms frame period, 30 ms window) on speech data sampled at 8KHz. Experiments were performed on cepstral coefficients without (baseline) and with robustness processing, results tabulated and compared. A simple energy threshold was used to discard non-speech. A 30-element codebook was trained for each speaker as the speaker model.

- Speaker identification: Test utterances were compared frame by frame with each speaker model; best codeword match was selected for each model, and then distortions were accumulated to make the final decision.

[*] with University of Maryland and Texas Instruments Incorporated

[†] University of Maryland, College Park; Martin Marietta Chair in Systems Engineering

0-7803-0946-4/93 $3.00 © 1993 IEEE

- Open set speaker verification: Use half of the speakers in the database as registered targets, and the other half as impostors. Test utterances were compared with all the registered target models, and the accumulated distortions were tallied to compute the rank of the claimed identity. If the rank is better than a certain threshold - accept, otherwise - reject. The thresholds were adjusted over a range to generate the ROC plot (detection vs. false alarm).

Because the higher order cepstral coefficients have less discriminating power and the lower order coefficients are more susceptible to environmental variation, we use a window function to de-emphasize both higher and lower order coefficients, called bandpass liftering [2].

$$w(k) = 1 + h\sin(\pi k/L)$$

where $h = L/2$, $k = 1, 2, \ldots, L$ and $w(k) = 0$ for other k, with $L = 20$.

Alternatively, we can filter the cepstral vector, with the following filter:

$$H(z) = \frac{a_0 + a_1 z^{-1} + a_3 z^{-3} + a_4 z^{-4}}{(1 - b_1 z^{-1}) z^{-4}}$$

with the coefficients chosen to approximate a bandpass frequency response. This bandpass operation is supposed to filter out slowly varying components of the cepstral coefficients in order to normalize environmental variations, called RASTA filtering [3].

Further, bandpass liftered cepstral coefficients can be processed by the RASTA filter as well.

3 Data Base

The data base utilized in this study is the narrowband portion of "King", collected in 10 sessions from 51 male speakers, 26 from San Diego and 25 from Nutley. The speakers were asked to talk about several topics, so that the speech is natural and spontaneous. The data were collected over long distance telephone line, and the data for 25 Nutley speakers were much noisier than that of 26 San Diego speakers. The speech material from each session is approximately 45 seconds long; the data were digitized at 8 kHz and 12-bit resolution. Sessions 1 to 5 and sessions 6 to 10 were collected under different environments. This division of data, "the great divide", results in serious degradation of performance as observed in [4] when training on one set and testing on the other. Also the Nutley data are much noisier than San Diego data. We performed our experiments in three contexts: San Diego alone (26 speakers), Nutley alone (25 speakers), and all 51 speakers combined. Further, the experiments were carried out across "the great divide" for the most challenging test condition.

4 Experiments and Results

4.1 Speaker Identification

Table 1 shows the results "within the great divide", table 2 shows the results "across the great divide". As the robustness processing techniques were added, the performance improved significantly.

- Within the great-divide: train on sessions 1, 2, 3, test on sessions 4, 5; train on sessions 6, 7, 8, test on sessions 9, 10.

- Across the great-divide: train on sessions 1, 2, 3, test on sessions 9, 10; train on sessions 6, 7, 8, test on sessions 4, 5.

Table 3 shows the comparison of our results with [4], for 16 San Diego speakers, trained on sessions 1, 2, and 3, tested on sessions 9 and 10.

4.2 Open Set Speaker Verification

Fig 1, 2, and 3 show the ROC plots for San Diego, Nutley and all 51 speakers experiments respectively. There are four curves in each figure:

- b: baseline, within the great divide.

- n: normalized (bandpass liftering & RASTA), within the great divide.

- bx: baseline, across the great divide.

- nx: normalized (bandpass liftering & RASTA), across the great divide.

5 Discussions and Conclusions

Tables 1 and 2 clearly show the independent contributions of bandpass liftering and RASTA to performance improvement. Bandpass liftering deemphasizes the highly variant and noisy cepstral coefficients and is a static correction. RASTA smoothes all of the cepstral coefficients by a bandpass filtering operation thereby attempting to remove the effects of the channel and the transducer. In this sense, the spoken material is "self-normalized," providing robustness. Thus, by combining the static (bandpass liftering) and dynamic (RASTA) techniques, we obtain the benefits of both techniques. Note that improvements are dramatic when the testing is across the great-divide. An interesting observation is that we get these improvements without using any specific noise-removal technique, such as spectral subtraction used in [6]. We have verified the consistency of the results of this paper on an independent data base, called the continuous speech recognition (CSR) data base. CSR consists of simultaneous recording of speech material from subjects using two different types of microphones. Hence CSR provides an excellent means of not only establishing consistency of our results, but also to develop insight into why these techniques work. This is accomplished by examining scattergrams of the various cepstral coefficients for recording with the two different microphones with and without robustness processing. As shown in Fig. 4, the cepstral coefficients of mic-1 vs. mic-2 are more aligned with $x = y$ after RASTA filtering. All of the above discussions hold for speaker verification as well.

Baseline			
	San Diego	Nutley	All
ID-rate	81.73%	35%	58.82%
Average-rank	2.01923	5.58	3.87745

Bandpass Liftering			
	San Diego	Nutley	All
ID-rate	85.58%	47%	66.67%
Average-rank	1.68269	4.11	2.88725

RASTA			
	San Diego	Nutley	All
ID-rate	91.35%	50%	71.08%
Average-rank	1.10577	3.51	2.29902

Bandpass Liftering & RASTA			
	San Diego	Nutley	All
ID-rate	94.23%	61%	77.94%
Average-rank	1.07692	2.72	1.89706

Table 1: Within the great divide

We have identified, by systematic investigation, a combination of techniques that provide a very robust performance. However, additional work is needed to improve the performance across the great divide to be at the level of performance within the great divide. Also, additional investigations planned with experimentation on a highly challenging corpus, the Switchboard [7], will shed further light.

Baseline			
	San Diego	Nutley	All
ID-rate	7.69%	36%	19.61%
Average-rank	10.0385	6.49	11.4608

Bandpass Liftering			
	San Diego	Nutley	All
ID-rate	36.54%	46%	36.76%
Average-rank	4.86538	4.86	6.37745

RASTA			
	San Diego	Nutley	All
ID-rate	42.31%	53%	43.63%
Average-rank	4.56731	3.35	6.23039

Bandpass Liftering & RASTA			
	San Diego	Nutley	All
ID-rate	77.88%	65%	58.82%
Average-rank	1.86538	2.24	3.48529

Table 2: Across the great divide

	[4]	BPL & RASTA
ID-rate	75%	87.5%
Average-rank	1.56	1.12

Table 3: Comparison between [4] and BPL & RASTA

References

[1] Alejandro Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition.* PhD thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, 1990.

[2] Biing-Hwang Juang, Lawrence R. Rabiner, and Jay G. Wilpon. On the use of bandpass liftering in speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-35(7), July 1987.

[3] Hynek Hermansky, Nelson Morgan, Aruna Bayya, and Phil Kohn. Rasta-plp speech analysis technique. In *International Conference on Acoustics, Speech, and Signal Processing*, 1992.

[4] Herbert Gish. Robust discrimination in automatic speaker identification. In *International Conference on Acoustics, Speech, and Signal Processing*, 1990.

[5] A. L. Higgins and L. G. Bahler. Text-independent speaker verification by discriminant count. In *International Conference on Acoustics, Speech, and Signal Processing*, 1991.

[6] Yu-Hung Kao, P. K. Rajasekaran, and John S. Baras. Free-text speaker identification over long distance telephone channel using hypothesized phonetic segmentation. In *International Conference on Acoustics, Speech, and Signal Processing*, 1992.

[7] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *International Conference on Acoustics, Speech, and Signal Processing*, 1992.

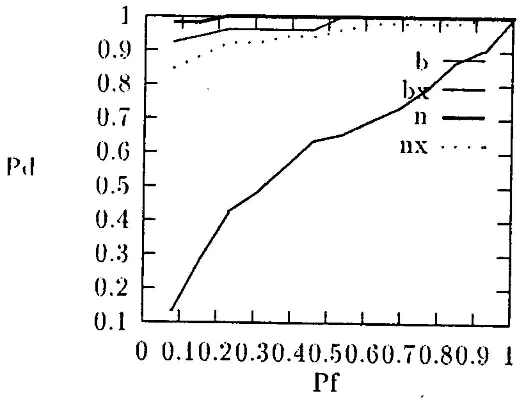

Fig 1: San Diego (26 speakers)

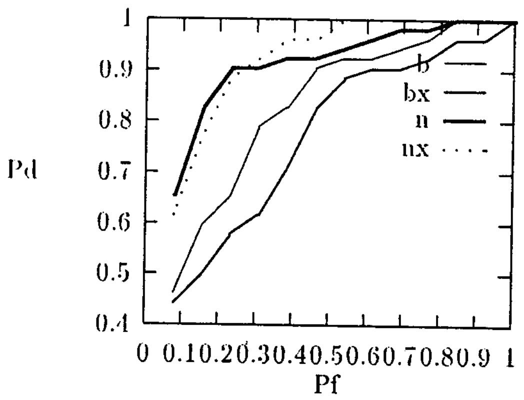

Fig 2: Nutley (25 speakers)

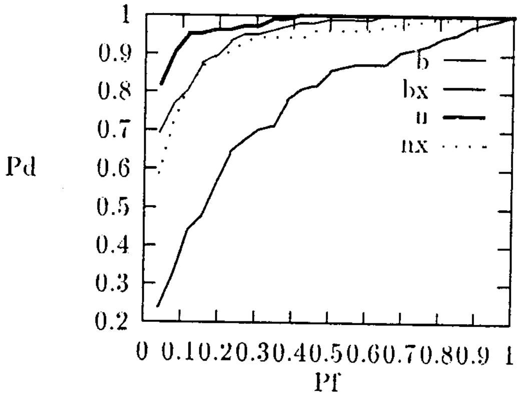

Fig 3: All (51 speakers)

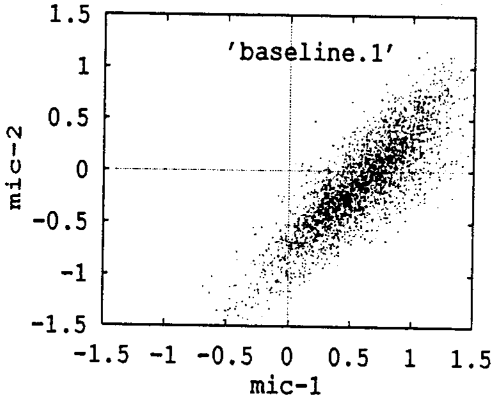

Fig 4A: Scattergram of mic-1 vs. mic-2

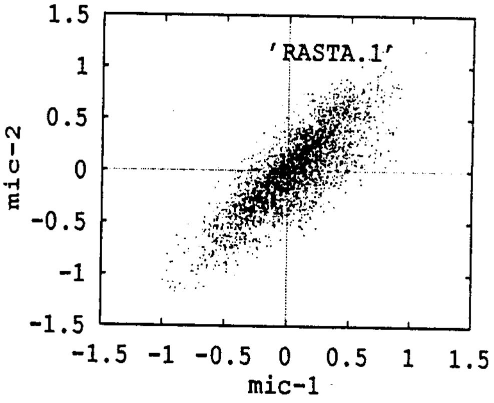

Fig 4B: Scattergram of mic-1 vs. mic-2