-11-11

ICASSP-92

1992 International Conference on Acoustics, Speech, and Signal Processing

March 23-26, 1992

San Francisco Marriott
San Francisco, California
U.S.A.

IEEE SIGNAL PROCESSING SOCIETY

IEEE

92CH3103-9

VOLUME 2

S_2ANU

Speech Processing 2
Audio
Neural Networks
Underwater Acoustics

Free-Text Speaker Identification Over Long Distance Telephone Channel Using Hypothesized Phonetic Segmentation

Yu-Hung Kao*, P. K. Rajasekaran, John S. Baras**

Texas Instruments Incorporated
P.O. Box 655474
Dallas, TX 75265

Abstract

Experimental investigation of free-text speaker identification method based on long-term statistics is conducted on a widely-used long distance telephone data base [1]. On a 26-speaker subset, an average correct identification of 93.3% is obtained; on the complete 51-speaker set, 67.6% correct identification is obtained. Speaker verification experiments on the data base provided receiver operating characteristics (ROC) comparable to or better than the ones available in open literature.

1. Introduction

This paper presents the results of experimental investigations into several aspects of free-text speaker identification and speaker authentication using a long-distance telephone data base, described in [1]. Specifically, the following speaker identification experiments were carried out to establish the effects of:

1. Using phonetic segments hypothesized by a speaker independent recognizer.
2. Using broad phonetic class segments, obtained by pooling classes from item 1. above.
3. VQ codebook size for the feature vector.
4. Using various parameters as feature vectors: log area ratios, LPC cepstral coefficients, and reflection coefficients.
5. Noise reduction including bandpass filtering.

Speaker verification was carried out using broad phonetic class speaker models along with noise reduction.

2 Algorithms

2.1 Speaker Models

A speaker model consists of one or more vector quantization (VQ) codebooks of a speech parameter vector (for example, LPC cepstral coefficients) derived from the training utterances of the speaker. A single codebook is used when no phonetic hypotheses are to be exploited; all non-speech frames are eliminated by energy thresholding, and the speech frames utilized to build the VQ codebook. Multiple codebooks are used for phonetic hypothesization experiments. The untranscribed training utterances are segmented by a speaker-independent continuous speech recognizer into phonetic categories. VQ codebook for each phonetic category (or broad phonetic category) is built from the corresponding hypothesized segments. A phonotactic grammar was used to improve the performance of the phonetic hypothesization. Phonetic models used in the recognizer were derived from the Voice Across America (VAA) data base described in [5].

2.2 Speaker Identification

For the case of non-phonetic method (single VQ codebook speaker model), each frame of the test utterance is classified as speech or non-speech; the features of the speech frames are compared with the model of each of the speakers in the population to generate the distortion for each candidate speaker according to a pooled speaker discrimination metric. The candidate with the least accumulated distortion is declared as the identified speaker. The pooled speaker discrimination metric is derived from the training data by maximizing the F-ratio to improve separability between speaker classes.

The alternative methods employed a speaker-independent continuous speech recognizer, whose output consists of hypothesized phonetic segments. The frames of speech from a phonetic category is compared with the candidate speaker model for that category as per the pooled speaker discrimination metric. Thus, a distortion for each of the observed phonetic category in the utterance is calculated. These distortions are summed (it is possible to do so in a weighted manner; but this was not done) over all observed phonetic categories to provide the total distortion for each candidate speaker. Again, the candidate with the least distortion is declared as the identified speaker.

2.3 Speaker Verification

VQ speaker models described in Section 2.1 were used along with Euclidean distance as the metric. We chose not to use the pooled speaker discrimination metric because it would have provided an unfair statistical knowledge of the impostors. In this paradigm, half the population (by choosing alternate speakers) was used as impostors, and the target speaker and "normalizing"

* with University of Maryland and Texas Instruments
** with University of Maryland, College Park

0-7803-0532-9/92 $3.00

speakers came from the other half. The total distortion over the input test utterance is computed for the target speaker as well as for each of the "normalizing" speaker. If the total distortion provided by the target speaker is lower than that of the "normalizing" speaker models, the test utterance is verified as belonging to the target speaker; otherwise the test utterance is declared as belonging to an impostor.

3. Data Base

The data base utilized in this study is the digitized subset of speech data collected in 10 sessions from 51 speakers, speaking on several topics (so that the speech is natural) over a long distance telephone line. Of the 51 speakers, 26 were based in San Diego and 25 in Nutley, N.J. The data from Nutley speakers was considerably noisier than that from San Diego. Further, the equipment used for recording had changed from session 6 onwards, establishing a division of the data base into two portions - sessions 1 through 5 (Div1), and sessions 6 through 10 (Div2). The nominal duration of the utterances in each session was about 45 seconds; when non-speech frames were eliminated, the average speech data was about 23 seconds. When the speaker-independent continuous speech recognizer with the phonotactic grammar was used, and non-phonetic categories (silence, background, inhalation, exhalation etc.) eliminated, the average speech data was only about 29 seconds.

Our experiments were conducted for both the 26-speaker (San Diego) subset and the total 51-speaker set. Training and test material were mostly restricted to Div1 or Div2; very limited experiments were performed where the training data was from Div1, and test data was from Div2, or vice versa.

4 Experiments and Results

4.1 Phonetic Segmentation

Previous published research [2] [3] [4] and informal discussions with various speech researchers in this area provided a mixed review of the value of automatic phonetic segmentation. This observation, along with the belief that the act of converting free, unknown text to known, but unfixed text (performed with acoustic consistency) should help, motivated us to investigate this aspect.

Allowing 49 phonetic categories, with a 10-element VQ codebook for each category (20 LPC cepstral coefficients), as the speaker model, an average correct speaker identification of 89.4% was obtained for the 26-speaker (San Diego) set; over all the 51 speakers, it was 63.2%. By collapsing the phonetic categories into 11 broad classes, and retraining the 10-element VQ codebook for each of the 11 classes, the improved average recognition of 93.3% was obtained for the 26-speaker set; average recognition of 67.6% resulted for the 51-speaker data. When no phonetic marking was used with a 110-element VQ codebook, the results were nearly identical to that of broad phonetic class method, and better than that of detailed phonetic marking (49-category).

Session No.		No. of speakers correctly identified		
Training	Test	A	B	C
1, 2, 3	4	25	25	24
1, 2, 3	5	24	24	24
6, 7, 8	9	24	24	23
6, 7, 8	10	24	24	22
Average		93.3%	93.3%	89.4%

A : Non-phonetic
B : Broad Phonetic (11 classes)
C : Detailed Phonetic (49 phones)

Table 1A (26-speaker San Diego data, noise reduced)

Session No.		No. of speakers correctly identified		
Training	Test	A	B	C
1, 2, 3	4	36	36	34
1, 2, 3	5	31	34	31
6, 7, 8	9	37	37	34
6, 7, 8	10	33	31	30
Average		67.2%	67.6%	63.2%

A : Non-phonetic
B : Broad Phonetic (11 classes)
C : Detailed Phonetic (49 phones)

Table 1B (51-speaker complete data, noise reduced)

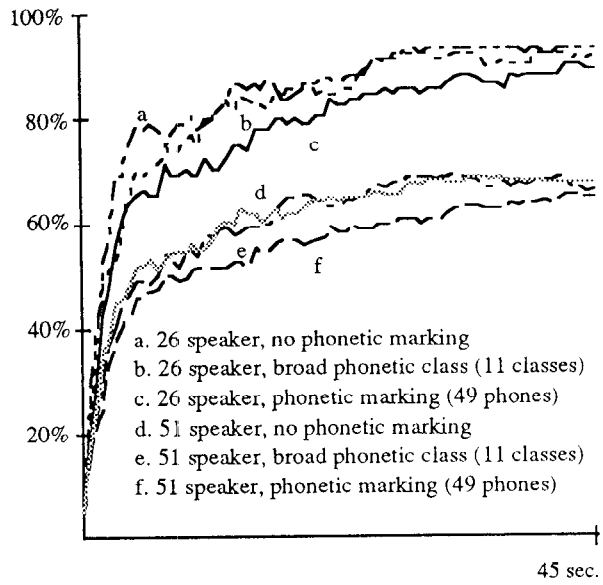

Fig 1 Correct Speaker Identification Rate as a function of input speech durations (including non-speech)

These above results were based on data where noise suppression was used (see Section 4.4). Detailed results are presented in Table 1. Figure 1 presents performance as a function of average utterance duration (see earlier comments on "speech only" duration in section 2).

4.2 VQ Codebook Size

With non-phonetic models, the codebook size was varied between 5 and 110. The performances were virtually identical for codebook sizes of 25, 55 or 110. With a codebook of size 10, it was noticeably worse.

4.3 Features

The effect of using various speech parameters for speaker identification was studied with the detailed phonetic model approach. LPC cepstral coefficients of dimension 20 yielded 89.4%, 10-th order log area ratio coefficients 86.5% and 10-th order reflection coefficients, 83.7% for the 26 San Diego speaker experiments. For the total 51-speaker experiment, the corresponding performances were 63.2%, 61.3%, and 59.3% respectively. All other results presented in this article were obtained with LPC cepstral coefficients.

4.4 Noise Reduction and Filtering

Preliminary listening and spectrographic analyses of the data base clearly showed the noisy nature of the data, especially that of the Nutley speakers. Spectral subtraction [6] method of noise suppression, along with bandpass filtering (300 - 3300 Hz), was used to preprocess the data base. Experimental results indicate that noise suppression increased the speaker identification rate by an additional 10% for both non-phonetic and phonetic methods. What is surprising is that the performance with the relatively cleaner San Diego data (26-speaker set) also showed the improvement. The effects of noise suppression are brought out in the performance results shown in Table 2.

Session No.		No. of speakers correctly identified			
Training	Test	A'	A	B'	B
1, 2, 3	4	25	22	24	20
1, 2, 3	5	24	20	24	18
6, 7, 8	9	24	23	23	24
6, 7, 8	10	24	22	22	21
Average		93.3%	83.7%	89.4%	79.8%

A : Non-phonetic
B : Detailed Phonetic
A' and B' are Noise suppressed versions of A and B

Table 2A (26-speaker San Diego data)

Session No.		No. of speakers correctly identified			
Training	Test	A'	A	B'	B
1, 2, 3	4	36	31	34	28
1, 2, 3	5	31	25	31	29
6, 7, 8	9	37	33	34	30
6, 7, 8	10	33	29	30	27
Average		67.2%	57.8%	63.2%	55.9%

A : Non-phonetic
B : Detailed Phonetic
A' and B' are Noise suppressed versions of A and B

Table 2B (51-speaker complete data)

4.5 Speaker Verification

A set of preliminary speaker verification experiment was carried out on the data base. The following three sets of data were considered: (i) 26-speaker San Diego speakers, (ii) 25-speaker Nutley speakers, and (iii) 51-speaker total population. Speaker models were based on broad phonetic classes, and Euclidean distance was used as the metric. The speaker models were derived from sessions 1, 2 and 3 for Div1 data experiments, and from sessions 6, 7 and 8 for Div2 data experiments. Test data came from sessions 4 and 5 for Div1 experiments and 9 and 10 for Div2 experiments. Figure 2 presents the receiver operating characteristics (ROC) for the three data sets for the case of training on sessions 1-3 and testing on session 4. Figure 3 presents the ROCs with results averaged over all the sessions.

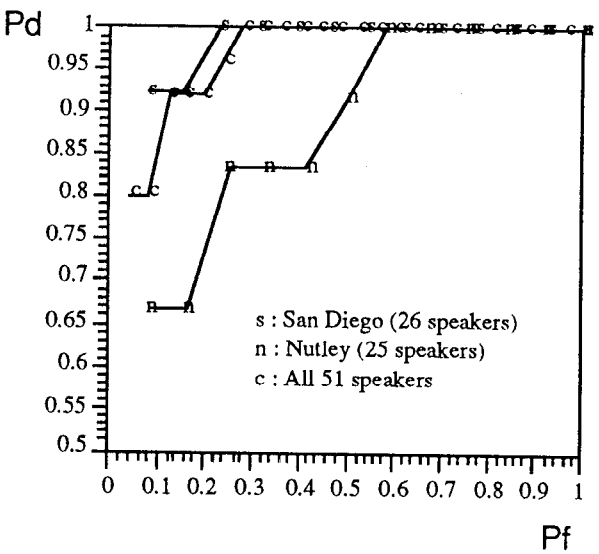

Fig. 2 Speaker Verification Performance : Session 4

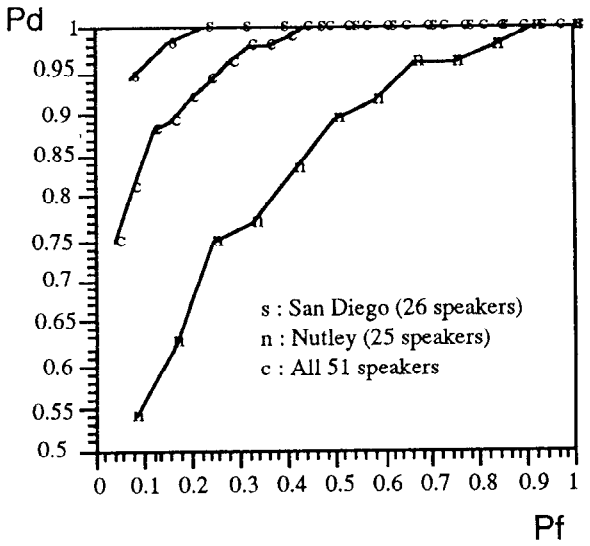

Fig. 3 Speaker Verification Performance : Average of sessions 4, 5, 9, and 10

5 Discussions

Our experiments indicate that detailed phonetic hypothesization for building speaker models and identification did not provide improvement over non-phonetic model approach. We believe that this result is probably due to inadequate training data, and possibly due to some poor phonetic segmentation. A thorough experiment with expert-marked speech data will indeed be very revealing; such data is usually limited (resulting in poor training of speaker models) and hard to obtain. However, if orthographic transcription of the data is available, we could perform supervised recognition [7] to obtain a reasonably good phonetic marking to enable us to determine the value of this approach. A suitable candidate would be the Switchboard data base [8], where the orthographic transcriptions along with word-level segmentation (guided by a pronunciation dictionary) are available.

Broad phonetic class speaker models performed as well as non-phonetic models with equivalent size codebook, and much better than phonetic models. Broad phonetic models also provide computational advantage in the distortion computation portion because of the smaller size codebooks (11 10-element codebooks vs. 110-element codebook); but the computational burden of speaker independent recognition will have to be taken into account in the overall computational requirements.

Noise suppression proved to be very valuable indeed. The average correct identification increased by 10%. We were surprised that it improved the results for even the relatively clean data from San Diego speakers.

Preliminary speaker verification experiments provided very encouraging results compared to other published results [9]. By incorporating a speaker discrimination scheme specific to the verification paradigm, we hope to improve our performance.

6 Acknowledgements

The authors would like to acknowledge the discussions with and comments by Jack Godfrey (TI), Vishu Viswanathan (TI), Curt Boylls (US Government), Herb Gish (BBN) and Al Higgins (ITT).

Reference

[1] H. Gish, "Robust Discrimination in Automatic Speaker Identification," ICASSP 1990, 289 - 292.

[2] S. Furui, "Research on Individuality Features in Speech Waves and Automatic Speaker Recognition Techniques," Speech Communication 5 (1986), 183 - 197.

[3] M. Savic and J. Sorensen, "Text-Independent Speaker Recognition Based on Phonetic Segmentation," Speech Research Symposium X, 1990.

[4] T. Matsui and S. Furui, "A Text-Independent Speaker Recognition Method Robust Against Utterance Variations," ICASSP 1991, 377 - 380.

[5] B. Wheatley and J. Picone, "Voice Across America : Toward Robust Speaker-Independent Speech Recognition for Telecommunications Applications," Digital Signal Processing, Apr. 1991, 45 - 63.

[6] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction,", IEEE Trans. on ASSP, Apr. 1979.

[7] B. Wheatley, et al, "Robust Automatic Time Alignment of Orthographic Transcriptions with Unconstrained Speech," ICASSP 92.

[8] J. Godfrey, et al, "SWITCHBOARD : Telephone Speech Corpus for Research and Development," ICASSP 92.

[9] A. L. Higgins and L. G. Bahler, "Text-Independent Speaker Verification by Discriminator Counting," ICASSP 91, 405 - 408.