Large Population Speaker Identification Using Clean and Telephone Speech

Douglas A. Reynolds

Abstract—**This paper presents text-independent speaker identification results for varying speaker population sizes up to 630 speakers for both clean, wideband speech, and telephone speech. A system based on Gaussian mixture speaker models is used for speaker identification, and experiments are conducted on the TIMIT and NTIMIT databases. The TIMIT results show large popultion performance under near-ideal conditions, and the NTIMIT results show the corresponding accuracy loss due to telephone transmission. These are believed to be the first speaker identification experiments on the complete 630 speaker TIMIT and NTIMIT databases and the largest text -independent speaker identification task reported to date. Identification accuracies of 99.5 and 60.7% were achieved on the TIMIT and NTIMIT databases, respectively.**

I. Introduction

THE task of speaker identification is to associate a speech sample with a speaker from a set of known speakers. In the general case, there is no *a priori* knowledge of the text being spoken; therefore, the system must operate in a text-independent mode. For this task, one of the major factors affecting performance is the size of the speaker population. In a finite feature space, as the number of speakers to be distinguished increases, performance eventually decreases due to speaker distribution overlap. Furthermore, the introduction of degradations such as additive noise and spectral shaping, as imposed by transmission over the telephone network, can further limit the distinguishability of speakers' voices. While, in general, both of these factors have been noted by several researchers, there have been no large scale studies examining both the effects of population size and telephone degradations on speaker identification performance. This letter presents an examination of text-independent speaker identification performance for varying speaker population sizes up to 630 speakers for both clean, wideband speech, and telephone speech.

One of the largest population speaker identification studies published [1] used a 963 speaker population with a text-dependent HMM-based system operating on digit strings spoken over the telephone. Under this vocabulary-constrained situation, a 97.8% accuracy was attained. Several text-independent studies using various recognition techniques have been conducted on the clean, wideband TIMIT database [2]. In [3], a 98.3% accuracy was attained for a 462 speaker population, and in [4], an accuracy of 95.6% using selected phonetic clusters was attained on the complete 630 speaker population. To our knowledge, there have been no published speaker identification experiments conducted on the corresponding telephone version of TIMIT (the NTIMIT database) [5]. On the Switchboard database [6], text-independent identification using a 113 speaker population has produced 82.8% accuracy [7].

In the present study, a system based on Gaussian mixture speaker models [9], [10] is used for speaker identification, and experiments are conducted on the TIMIT and NTIMIT databases. The Gaussian mixture speaker model statistically represents the underlying sounds or vocal tract configurations that characterize a person's voice, and it has proven very effective for several speaker identification and verification tasks [7]. The TIMIT/NTIMIT database pair was selected for this study because it provides both clean, wideband, and telephone speech from a large number (630) of speakers. There are two major aims of this study. The first aim is to establish how well text-independent speaker identification can perform under near-ideal conditions for very large populations. This will provide an indication of the inherent "crowding" of the feature space when the data is free from transmission and intersession variability degradations. The second aim is to gauge the performance loss incurred by transmitting the speech over the telephone network for the same large population experiment.

The remainder of the letter is organized as follows. The next section describes the speaker identification system. This is followed in Section III with a description of the characteristics of the TIMIT and NTIMIT databases. Section IV presents speaker identification performance versus population size for both the TIMIT and NTIMIT databases. Last, discussion and conclusions are given in Section V.

II. Speaker Identification System

A. Speech Analysis

Several processing steps occur in the front-end speech analysis. First, the speech is segmented into frames by a 20-ms window progressing at a 10-ms frame rate. An adaptive, energy-based speech activity detector (SAD) is then used to discard silence/noise frames [10], [11]. The SAD only passes frames with energies that are 5 dB above the estimated noise floor; therefore, low-energy speech frames are discarded as are initial and final pauses. For a NTIMIT utterance, about 26% of the frames are discarded. The SAD is not used on the

Manuscript received September 17, 1994; approved December 13, 1994. This work was sponsored by the Department of the Air Force. The views expressed are those of the author and do not reflect the official policy or positions of the U.S. Government. The associate editor coordinating the review of this letter and approving it for publication was Dr. J. R. Rohlicek.

The author is with the Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02174 USA.

IEEE Log Number 9409859.

1070–9908/95$04.00 © 1995 IEEE

TABLE I
CHARACTERISTICS OF TIMIT AND NTIMIT DATABASES. (PSTN=PUBLIC SWITCHED TELEPHONE NETWORK)

Database	*#speakers*	*#utterance/speaker*	*channel*	*acoustic environment*	*handset*	*intersession interval*
TIMIT	630	10 (3 s/utterance) read sentence	clean	sound booth	fixed wideband	none (same session)
NTIMIT	630	10 (3 s/utterance) read sentence	PSTN local and long distance	sound booth	fixed carbon button	none (same session)

TIMIT data as it was found to slightly decrease performance. Eliminating the noise frames from the NTIMIT data is more important for good performance since the sentence-to-sentence noise variation from transmission over different telephone lines is a source of mismatch between training and testing data.

Next, mel-scale cepstral feature vectors are extracted from the speech frames using a simulated mel-spaced filter bank on the FFT coefficients [10], [11]. For the bandlimited NTIMIT telephone speech, cepstral analysis is performed only over the mel-filters in the telephone passband (300–3452 Hz). All cepstral coefficients except $c[0]$ are retained in the processing. On the TIMIT data, 30 cepstral coefficients per feature vector are used, and 20 cepstral coefficients per feature vector for the NTIMIT data. Differential cepstral parameters are not used. Since there is no variability between recording microphones in both the TIMIT and NTIMIT databases and little variability between the telephone lines in the NTIMIT database,[1] channel equalization via cepstral mean removal is not used. In fact, it was found to decrease performance for both databases.

B. Gaussian Mixture Speaker Model

The basis for the identification systems is the Gaussian mixture model (GMM) used to represent speakers. More specifically, the distribution of feature vectors extracted from a person's speech is modeled by a Gaussian mixture density. For a feature vector denoted as $\vec{x}$, the mixture density for speaker s is defined as $p(\vec{x}|\lambda_s) = \Sigma_{i=1}^{M} p_i^s b_i^s(\vec{x})$. The density is a weighted linear combination of M component unimodal Gaussian densities $b_i^s(\vec{x})$, each parameterized by a mean vector $\vec{\mu}_i^s$ and covariance matrix Σ_i^s. Collectively, the parameters of a speaker's density model are denoted as $\lambda_s = \{p_i^s \vec{\mu}_i^s \Sigma_i^s\}$. In this paper, 32 component mixtures with diagonal covariance matrices are used. Maximum likelihood estimates of the model parameters are obtained using the expectation-maximization (EM) algorithm.

C. Identification Decision

For an utterance $X = \{\vec{x}_1, \cdots, \vec{x}_T\}$ and a reference group of S speakers represented by models $(\lambda_1, \lambda_2, \cdots, \lambda_s)$, identification is performed by using the maximum likelihood classification rule $\hat{s} = \text{argmax}_{1 \leq s \leq S} p(X|\lambda_S) = \text{argmax}_{1 \leq s \leq S} \Sigma_{t=1}^{T} \log p(\vec{x}_t|\lambda_s)$, where the last expression comes from using logarithms and the assumed independence between observations.

III. DATABASES

Some of the characteristics of the TIMIT and NTIMIT databases are shown in Table I.

The TIMIT database allows examination of speaker identification performance under almost ideal conditions. With the 8-kHz bandwidth and lack of intersession variability, acoustic noise, and microphone variability or distortion, recognition errors should almost entirely be a function of nondistinguishable speaker distributions. Furthermore, the speech is read sentences, some of which are designed to have rich phonetic variability. This is a factor that favorably biases TIMIT performance compared with 3-s utterances extracted at random from extemporaneous speech.

The NTIMIT database is the same speech from the TIMIT database played through a carbon-button telephone handset and recorded over local and long-distance telephone loops. This provides the identical TIMIT speech, except that it is degraded through carbon-button transductions and actual telephone line conditions. Performance differences between identical experiments on TIMIT and NTIMIT should arise mainly from the effects of telephone transmission degradations.

In the following experiments, all 630 speakers (438 males and 192 females) are used. Each speaker's model has 32 Gaussians and is trained using his/her two sa sentences, three si sentences, and three sx sentences (approximately 24 s). The remaining two sx sentences (ordered alphabetically) per speaker are individually used as tests (a total of 1260 tests of 3 s each).

IV. EXPERIMENTAL RESULTS

Fig. 1 shows speaker identification accuracy versus population size on the TIMIT and NTIMIT databases. Identification accuracy for a population size S is computed by performing speaker identification tests on 50 sets of S speakers randomly selected from the 630 speakers and averaging the results. This helps average out the bias of a particular population composition. Population sizes of (10, 100, 200, 300, 400, 500, 600, 630) are used.

Under the near-ideal TIMIT conditions, performance is barely affected by increasing population sizes, indicating that the limiting factor in speaker identification performance is not a crowding of the feature space. However, with telephone line degradations, the NTIMIT accuracy steadily decreases as population size increases. With the 630 speaker populations, there is a gap of 39 percentage points between TIMIT and NTIMIT accuracy (TIMIT Pc = 99.5%, NTIMIT Pc

[1] Based on examination of the channel responses derived from the sweep tones supplied with the NTIMIT database.

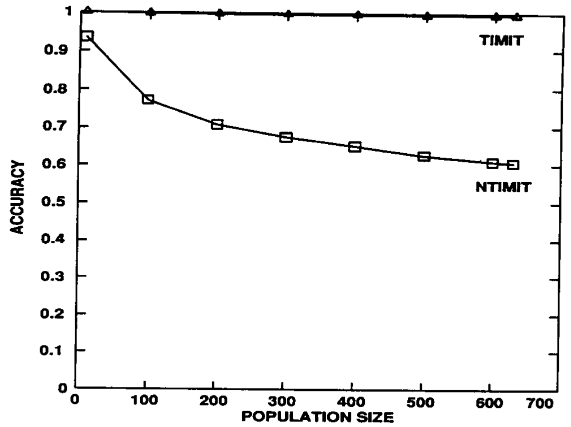

Fig. 1. Speaker identification accuracy as a function of population size on TIMIT and NTIMIT databases.

= 60.7%). A detailed study of performance loss associated with the degradations of bandlimiting, spectral shaping, and noise addition can be found in [8]. The effects of further degradations from intersession and microphone variability could not be assessed with these databases.

On the TIMIT database, there were no cross-sex errors, with male and female accuracies of 99.8 and 99.0%, respectively. On the NTIMIT database, there were four cross-sex errors, with male and female accuracies of 62.5 and 56.5%, respectively.

V. Conclusion

The excellent performance on the TIMIT database indicates that the limiting performance factor in large population speaker identification is not a crowding of the feature space (at least up to 630 speakers). Rather, based on the accuracy decrease in the corresponding NTIMIT experiments, it appears that corruption of the speech signal from transmission effects is a much stronger factor. This corruption affects recognition performance by both the loss of speaker-dependent spectral information through bandiimiting and noise addition and by creating mismatches between training and testing data. Both cases can severely impact the separability of speaker's voices. Other mismatch effects such as intersession, microphone, and phonetic variability could not be examined with these databases, but they are also strong factors in the performance of speaker identification algorithms.

References

[1] J. J. Webb and E. L. Rissanen, "Speaker identification experiments using EMMs," in *Proc. Int. Conf. Acoust, Speech, Signal Processing*, 1993, pp. II-387–II-390.

[2] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Workshop Speech Recognition*, Feb. 1986, pp. 93–99.

[3] L. F. Lamel and J. L. Cauvain, "Cross-lingual experiments with phone recognition," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1993, pp. II-507–II-510.

[4] J. L. Floch, C. Montacie, and M. J. Caraty, "Investigations on speaker characterization from Orphee system technics," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, Apr. 1994, pp. I-149–I-152.

[5] C. Jankowski *et. al.*, "NTIMIT: A phonetically balanced, continuous speech telephone bandwidth speech database," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, Apr. 1990, pp. 109–112.

[6] J. J. Godfrey, E. C. Holliman, and J. MacDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1992, pp. I-517–I-520.

[7] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," in *Proc. ESCA Workshop Automat. Speaker Recognition, Identifcation Verification*, Apr. 1994, pp. 27–30.

[8] ——, "Effects of population size and telephone degradations on speaker identification performance," in *Proc. SPIE Conf. Automat. Syst. Identification Inspection Humans*, July 1994.

[9] R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1990, pp. 293–296.

[10] D. A. Reynolds, "A Gaussian mixture modeling approach to text-independent speaker identification," Ph.D. thesis, Georgia Inst. of Technol., 1992.

[11] D. A. Reynolds, R. C. Rose, and M. J. T. Smith, "PC-based TMS320C30 implementation of the Gaussian mixture model text-independent speaker recognition system," in *Proc. Int. Conf. Signal Processing Applications Technol.*, Nov. 1992, pp. 967–973.