THE EFFECTS OF TELEPHONE TRANSMISSION DEGRADATIONS ON SPEAKER RECOGNITION PERFORMANCE *

D. A. Reynolds *M. A. Zissman* *T. F. Quatieri* *G. C. O'Leary* *B. A. Carlson*

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

The two largest factors affecting automatic speaker identification performance are the size of the population and the degradations introduced by noisy communication channels (e.g., telephone transmission). To examine experimentally these two factors, 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. This is 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% are achieved on the TIMIT and NTIMIT databases, respectively. This paper also presents experiments which examine and attempt to quantify the performance loss associated with various telephone degradations by systematically degrading the TIMIT speech in a manner consistent with measured NTIMIT degradations and measuring the performance loss at each step. It is found that the standard degradations of filtering and additive noise do not account for all of the performance gap between the TIMIT and NTIMIT data. Measurements of nonlinear microphone distortions are also described which *may* explain the additional performance loss.

1. INTRODUCTION

One of the major factors affecting speaker identification performance is the size of the speaker population. In a finite feature space, as the number of speakers to be distinguished increases, performance must eventually decrease due to overlap of speaker distributions. Furthermore, the introduction of degradations 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. The purpose of this paper is to examine text-independent speaker identification performance with varying speaker population sizes up to 630 speakers for both clean, wideband speech and telephone speech.

A system based on Gaussian mixture speaker models [1,2] is used for speaker identification and experiments are conducted on the TIMIT [3] and NTIMIT [4] databases. 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. The NTIMIT database also provides calibration signals for measuring characteristics of the telephone lines used.

There are several 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. The next aim is to gauge the performance loss incurred by transmitting the speech over the telephone network for the same large population experiment. The third aim is to examine the validity of current models of telephone degradations commonly used in developing compensation techniques. The approach is to synthesize the assumed degradations on the TIMIT speech using the NTIMIT calibration signals and determine if they produce the performance loss actually measured when using the NTIMIT speech. A similar type of study for speech recognition was reported in [5]. As shown later, it is found that the standard degradations (filtering and noise addition) do not account for all of the performance loss and the discrepancy *may* be due to some observed microphone nonlinear distortions.

The remainder of the paper is organized as follows. The next section briefly describes the speaker identification system. This is followed in Section 3 with a description of the characteristics of the TIMIT and NTIMIT databases and presentation of speaker identification performance versus population size for both databases. The simulation of telephone degradations and their evaluation is given in Section 4. Measurements and observations of microphone nonlinear effects are then given in Section 5. Last, discussion and conclusions are given in Section 6.

2. SPEAKER IDENTIFICATION SYSTEM

The identification system is a statistical recognition system based on representing each speaker's acoustic parameter distribution 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 $b_i^s(\vec{x})$. The Gaussian mixture speaker model can be viewed as a hybrid between two effective models for speaker recognition: unimodal Gaussian classifiers and vector quantizer codebooks. The GMM has been shown to be an effective speaker representation for both identification and verification tasks [6].

Speech is parameterized as mel-cepstral feature vectors. All cepstral coefficients except $c[0]$ are retained in the processing. For the telephone speech, cepstral analysis is performed only over the mel-filters in the telephone passband (300-3300 Hz) and noise frames are removed using an adaptive, energy-based speech detector. For the clean speech, all mel-filters (24 total) are used and no speech detection is performed.

*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.

0-7803-2431-5/95 $4.00

3. TIMIT AND NTIMIT RESULTS

3.1. Databases

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

The NTIMIT database consists of the TIMIT sentences played through a carbon-button telephone handset[1], transmitted through a local or long-distance central office and looped back for recording. Performance differences between identical experiments on TIMIT and NTIMIT should arise mainly from the effects of the microphone and telephone transmission degradations.

3.2. Results

In the following experiments all 630 speakers (438 males, 192 females) are used. Speaker models with 32 Gaussians are trained using the 2 sa sentences, 3 si sentences and 3 sx sentences (approx. 24 sec). The remaining 2 sx sentences are individually used as tests (a total of 1260 tests of 3 secs each). TIMIT results are from training and testing with TIMIT speech and NTIMIT results are from training and testing with NTIMIT speech.

Figure 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 repeated speaker identification tests on 50 sets of S speakers randomly selected from the 630 pool of available 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.

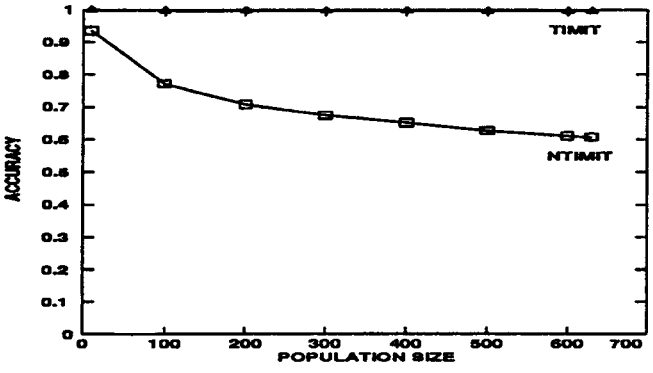

Figure 1. Speaker identification accuracy as a function of population size on TIMIT and NTIMIT databases.

Under the near ideal TIMIT conditions, performance is barely affected by increasing population sizes. This indicates that the limiting factor in speaker identification performance is not a crowding of the feature space (at least for population sizes of 630 speakers). However, with telephone line degradations, the NTIMIT accuracy steadily decreases as population size increases. The largest drop in accuracy occurs as the population size increases to 100. Above 200 speakers the accuracy decrease becomes almost linear. With the full 630 speaker populations, there is a gap of 39 percentage points between TIMIT and NTIMIT accuracy (TIMIT Pc = 99.5% ± 0.2%, NTIMIT Pc = 60.7% ± 1.4%).

[1]The same handset was used for all speech.

The large population TIMIT results are in agreement with results from other sites [7, 8]. To the authors' knowledge, there have been no published results on the NTIMIT database.

4. SIMULATING TELEPHONE DEGRADATIONS

4.1. Telephone Degradations Model

The prevailing model of the telephone transmission path used in speech processing is a linear filter followed by an additive noise source. The microphone and telephone channel are generally combined and modeled as a linear filter. The noise is often assumed to be additive stationary white or colored Gaussian. With this model, the assumed degradations are bandlimiting, spectral shaping, and noise addition. Other degradations such as additive tones, impulsive noise, phase jitter and nonlinear distortions may certainly affect the speech signal but are considered secondary effects in this model.

The system to simulate these degradations on the TIMIT speech using the NTIMIT calibration signals is shown in Figure 2 and described next.

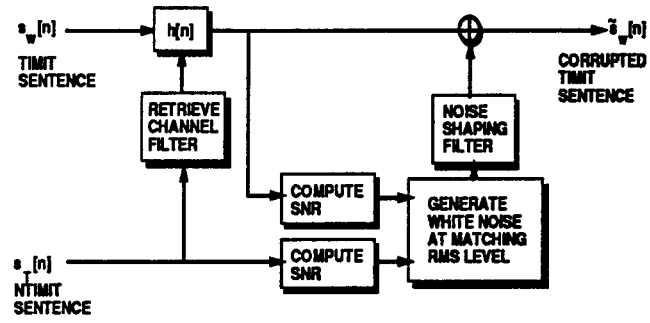

Figure 2. Simulation to corrupt TIMIT sentence to match NTIMIT sentence.

Bandlimiting: Mel-cepstral coefficients were derived from mel-filterbank outputs over the frequency range 300–3400 Hz.

Spectral Shaping (filtering): The sweep tones from each NTIMIT telephone line were used to derive an FIR channel filter by using a MSE FIR filter design to match the sweep-tone's magnitude spectra. Each TIMIT sentence was then filtered by the channel filter from the corresponding NTIMIT sentence[2].

As illustrated in Figure 3, there is little spectral variability among the NTIMIT channels. The dominant spectral shaping appears to be from the carbon-button microphone.

Noise addition: Broadband colored Gaussian noise was added to each TIMIT sentence at a level to match the signal-to-noise (SNR) of the corresponding NTIMIT sentence. In this way the same NTIMIT sentence-to-sentence noise levels are used. The coloring was a high-frequency de-emphasis filter designed to match noise characteristics observed in several NTIMIT sentences. The NTIMIT sentences had a mean SNR of 36.5 dB[3] with a standard deviation of 5.6 dB while the noise corrupted TIMIT sentences had a mean SNR of 36.7 dB with a standard deviation of 5.6 dB. The clean TIMIT had an average SNR of 53 dB.

[2]The NTIMIT database provides a file indicating through which telephone line each NTIMIT sentence was passed.

[3]SNR measured as the ratio of peak-signal energy to mean noise energy.

Table 1. Speaker ID results for TIMIT, NTIMIT and simulated NTIMIT (168 speaker subset)

Database	NTIMIT	TIMIT	TIMIT +bandlimiting	TIMIT +bandlimiting +noise	TIMIT +bandlimiting +filtering	TIMIT +bandlimiting +filtering +noise
Pc (%)	69.0	99.1	95.2	93.8	89.9	85.1

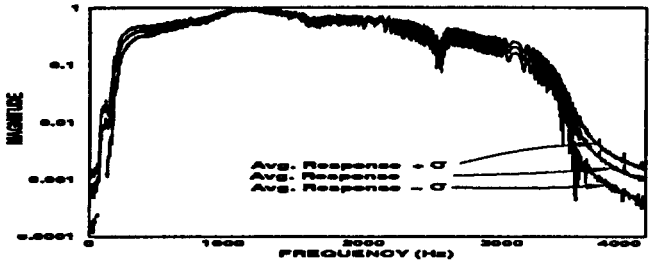

Figure 3. Channel responses derived from NTIMIT sweep tones. Plot shows average response over 220 channels and average ± standard deviation of the channels.

4.2. Evaluation of Simulated Degradations

The above degradations were sequentially imposed on the TIMIT database and speaker identification accuracies calculated. This experiment used a 168 speaker subset (112 males, 56 females; the "test" speakers) of the complete database. Other experimental conditions were set as before, except speech detection was not used. Speaker models were retrained after each degradation was imposed on the TIMIT speech so results are for matched train/test conditions. The results from the experiments are given in Table 1.

Even with all the degradations assumed by the model. the corrupted TIMIT results are still 16 percentage points higher than the NTIMIT results. A further test using speaker models trained on NTIMIT to recognize the corrupted TIMIT speech only produced 39% accuracy. Although not surprising, these results are clear indications that the assumed model is not accounting for all the degradations present.

The largest drop in performance occurred for the filtering; however, blind deconvolution did not help performance. The noise addition caused a smaller performance drop than expected. Adding an extra 5 dB of noise to the estimated noise level did further decrease the accuracy to 76.3%; however, the corrupted TIMIT speech then sounded significantly more noisy than the corresponding NTIMIT speech. While the TIMIT speech *can* be distorted enough to reduce its performance to the NTIMIT level, the aim here is to keep the degradations closely matched to those found in the NTIMIT speech.

5. NONLINEAR MICROPHONE EFFECTS

Nonlinear distortion from the carbon-button microphone and telephone channels is one unaccounted for effect that may help explain some of the performance gap between the corrupted TIMIT and the NTIMIT results. Comparison of TIMIT and NTIMIT speech segments does indeed show evidence of these types of distortions. Simple static nonlinearities (e.g., quadratic and cubic) applied to the speech signal map to convolutions of the original spectrum with itself, introducing "phantom formants" at sums and differences of formant frequencies (i.e., intermodulation distortions). This convolutional view of the nonlinear distortion also indicates resonance bandwidths can be broadened or narrowed depending on the order of nonlinearity. While both of these effects can be found in the speech, it is very difficult to estimate and apply these distortions as was done with the filtering and noise above.

As a first step in analyzing the effects of nonlinear distortion we transmitted and recorded some simple test signals over a telephone handset and channel to characterize the extent of the nonlinearity preset in a typical telephone connection. Two types of test signals were used to probe the handset and channel: (1) sine waves with frequencies of 400, 675 and 1000 Hz, (2) a swept sine wave with a 10–4000 Hz linear sweep over a 2 sec duration. Both signal types were generated at levels of 68, 78, 88, 98, and 108 dB. The signals were digitally generated and stored onto a digital-audio-tape (DAT). The signals were then transmitted under two conditions:

- NOHS: no handset. Test signals were played out of the DAT directly to a telephone line interface, transmitted over a local PBX and the PSTN[4], resampled at the receiving line interface and stored to disk.
- CARB: carbon-button microphone. Test signals were played out of the DAT through an artificial mouth into a handset with a carbon-button microphone. The DAT output was adjusted so the 88 dB sine was played into the handset at 88 dB SPL. The speech was sent over the local PBX and PSTN, resampled at the receiving line interface and stored to disk.

5.1. Telephone Channel Measurements

The NOHS condition provides the opportunity to measure the frequency response and the non-linearities present in the telephone channel without the intervening handset. The left plot in Figure 4 shows the received level of the fundamental and second harmonic for the sine waves. In each case, the labels on the curves indicate the frequency of the input sine wave The three curves for the fundamental are quite linear with a slope of 1.0 and almost coincident, indicating the relative flat frequency response of the channel between 400 and 1000 Hz. The curves for the second harmonic show random behavior when the input signal was below 98 dB, corresponding to a second harmonic output of 50 dB or less. This is because 50 dB is the approximate noise floor for this channel, so the second harmonic was at or below the noise floor for these input signal levels. Although the level of the second harmonic does rise above the noise floor for input signal levels of 98 and 108 dB, the spread between the fundamental and second harmonic levels is still greater than 35 dB. Thus, the telephone channel without a handset appears linear.

The right plot in Figure 4 shows the frequency response of the channel, as measured from the response to the sweep tones, for five input signal levels. The difference between the curves is relatively uniform, again demonstrating the linearity of the channel.

5.2. Carbon-Button Microphone Measurements

Ideally, one would like to measure the characteristics of the carbon microphone in isolation. However, because the channel is relatively linear and given that the frequency response of the channel is almost flat between 400 and 2000 Hz, it is reasonable to ignore the channel for fundamental vs. second harmonic and frequency response measurements. The left

[4]Public Switched Telephone Network

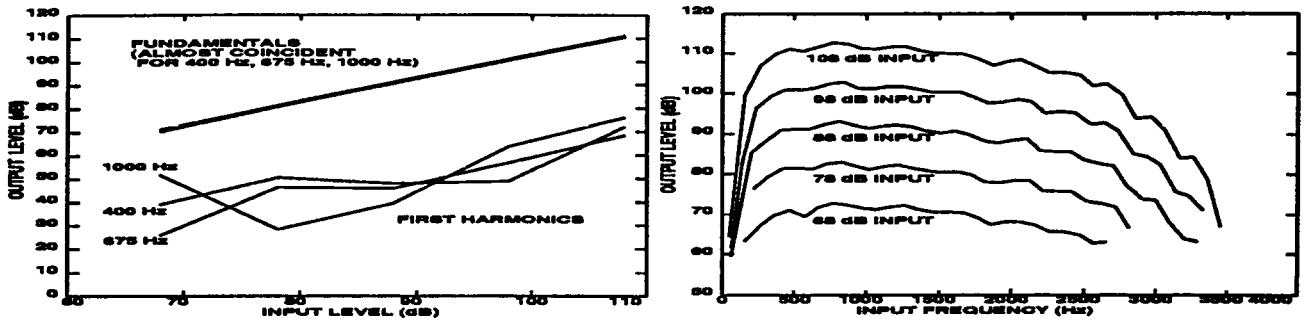

Figure 4. (Left plot) Comparison of measured fundamental and second harmonic signal levels for the NOHS condition. (Right Plot) Frequency response for the NOHS condition

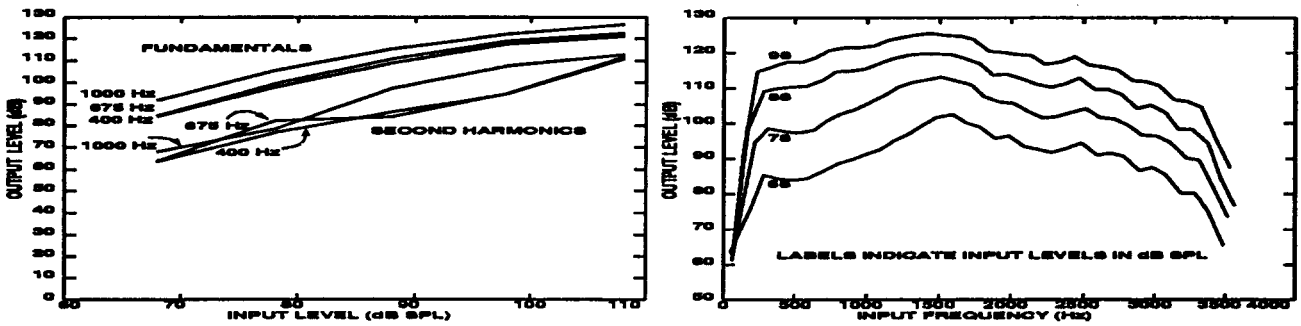

Figure 5. (Left plot) Comparison of measured fundamental and second harmonic signal levels for the CARB condition. (Right plot) Frequency response for the CARB condition as a function of input signal level.

plot in Figure 5 shows the received level of the fundamental and second harmonic for the sine waves. As a result of the non-flat frequency response of the microphone, the fundamental curves are no longer coincident. Furthermore, the fact that the curves are nonlinear is initial evidence of the nonlinear response of the microphone. Additional evidence of microphone nonlinearity is seen in the sharp increase in the output level of the second harmonic as the input signal level increases. For the 1000 Hz tone, the spread between the levels of the fundamental and second harmonic is often as little as 15 dB.

The right plot in Figure 5 shows the frequency response of the microphone, as measured from the response to the sweep tones, for four input signal levels. The extent to which the frequency responses are not merely translations of each other demonstrates the non-linearity in the microphone[5].

6. CONCLUSION

This paper has presented the first speaker identification experiments on the complete 630 speaker TIMIT and NTIMIT databases. The experiments indicate that, under ideal conditions, there is not an inherent crowding of the feature space with increasing population sizes. However, degradations from telephone transmission do indeed diminish the distinguishability of speaker voices, causing considerable accuracy loss with increasing population size.

The simulation of the prevailing telephone degradation model on TIMIT speech failed to produce results which matched the observed performance on NTIMIT, indicating the model is not accounting for some key degradation(s). Evidence of nonlinear distortion was found in the NTIMIT data (via "phantom formants") and measurements were presented clearly showing the nonlinear distortion produced by a carbon-button microphone.

Current effort is focused on developing plausible models and measurement techniques for the nonlinear distortions, so we can determine the extent of performance loss attributable to such distortions.

[5] Presumably the response due to the 108 dB SPL signal would have been even more marked, but severe distortion at that level prevented reliable measurements of the sweep tone.

REFERENCES

[1] R. C. Rose and D. A. Reynolds, "Text-independent speaker identification using automatic acoustic segmentation," in ICASSP90, pp. 293–296, 1990.

[2] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, p. *, January 1995.

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

[4] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech telephone bandwidth speech database," in ICASSP90, pp. 109–112, April 1990.

[5] P. Moreno and R. Stern, "Sources of degradation of speech recognition in the telephone network," in ICASSP94, pp. I–109–I–112, April 1994.

[6] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," in *Proc. of the ESCA Workshop on Automatic Speaker Recognition*, pp. 27–30, April 1994.

[7] L. F. Lamel and J. L. Gauvain, "Cross-lingual experiments with phone recognition," in ICASSP93, pp. II–507–II–510, 1993.

[8] J. L. Floch, C. Montacie, and M. J. Caraty, "Investigations on speaker characterization from Orphee system technics," in ICASSP94, pp. I–149–I–152, April 1994.