HTIMIT AND LLHDB : SPEECH CORPORA FOR THE STUDY OF HANDSET TRANSDUCER EFFECTS

Douglas A. Reynolds

Lincoln Laboratory, Massachusetts Institute of Technology
Lexington, Massachusetts 02173–9185
e-mail: `DAR@SST.LL.MIT.EDU`

ABSTRACT

This paper describes two corpora collected at Lincoln Laboratory for the study of handset transducer effects on the speech signal: the handset TIMIT (HTIMIT) corpus and the Lincoln Laboratory Handset Database (LLHDB). The goal of these corpora are to minimize all confounding factors and produce speech predominately differing only in handset transducer effects. The speech is recorded directly from a telephone unit in a sound-booth using prompted text and extemporaneous photograph descriptions. The two corpora allow comparison of speech collected from a person speaking into a handset (LLHDB) versus speech played through a loudspeaker into a handset (HTIMIT). A comparison of analysis and results between the two corpora will address the realism of artificially creating handset degraded speech by playing recorded speech through handsets. The corpora are designed primarily for speaker recognition experimentation (in terms of amount of speech and level of transcription), but since both speaker and speech recognition systems operate on the same acoustic features affected by the handset, knowledge gleaned is directly transferable to speech recognizers.

Initial speaker identification performance on these corpora are presented. In addition, the application of HTIMIT in developing a handset detector that was successfully used on a Switchboard speaker verification task is described.

1. INTRODUCTION

It is well known that one of the major causes of performance degradation in speech and speaker recognition systems is microphone variability. This is especially detrimental for systems operating over the telephone network in which it is impossible to control the type of telephone handset transducers used. Several sites have reported the performance effects of handset variability[1] on tasks such as speaker verification [1, 2] and continuous digit recognition [3]. For example, it was shown in [1] that speaker verification equal-error rates were 1.8 times greater for training speech collected from a carbon-button transducer and testing speech collected from an electret transducer than for training and testing speech collected both from the carbon-button transducer. Digit recognition error rates also increased by a factor of 3.2 for a similar carbon-electret versus carbon-carbon train/test experiment [3].

Unfortunately, there are very few publicly available speech corpora which contain explicit handset variability as a factor for experimentation and study. While some standard speech corpora, such as Switchboard, do implicitly contain handset variability from a large number of people calling from different telephone numbers, the variability is captured in an uncontrolled, albeit realistic manner, without explicit knowledge of the type of handset transducer used for a particular recording. Using a rough assumption that different telephone numbers imply different handsets, it is possible to conduct experiments on Switchboard examining the performance degradations caused by "mismatched" relative to "matched" handset conditions [4, 2, 5].

The lack of handset type "ground-truth" coupled with the added confounding factors of variable telephone channels, acoustic environments and linguistic content make it difficult to specifically study why and how different handset transducers affect recognition systems. In this paper we describe two speech corpora collected at Lincoln Laboratory designed to explicitly focus on handset effects: *Handset TIMIT* (HTIMIT) and the *Lincoln Laboratory Handset Database* (LLHDB). The goal in collecting these corpora was to minimize all confounding factors and produce speech predominately differing only in handset transducer effects. The speech was recorded directly from a telephone unit (to minimize channel effects) in a sound-booth (to minimize acoustic environment effects) using prompted text (to minimize linguistic effects).

The remainder of the paper is organized as follows. The next two sections describe the HTIMIT and LLHDB corpora, respectively. Section 4 then describes some analysis and experiments conducted using the two corpora. Finally, the last section provides some additional potential applications of the corpora and future plans.

2. HTIMIT CORPUS

The HTIMIT corpus was constructed by playing a subset of the clean TIMIT corpus [6] through various handsets

Paper published in Proceedings of the *International Conference on Acoustics, Speech and Signal Processing*, pp 1535–1538, May 1997 Munich, Germany. This work was sponsored by the Department of the Air Force. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Air Force.

[1]Throughout this paper the terms "handset variability" and "transducer variability" will be used interchangeably.

using the setup shown in Figure 1. We used a gender balanced subset of 384 TIMIT speakers (192 males and 192 females). The aim was to maintain the speaker and linguistic richness of the original TIMIT corpus and impose real handset transducer degradations in a controlled, systematic manner. This is a similar strategy to that used in generating the narrowband TIMIT (NTIMIT) [7] and the cellular TIMIT (CTIMIT) [8] corpora.

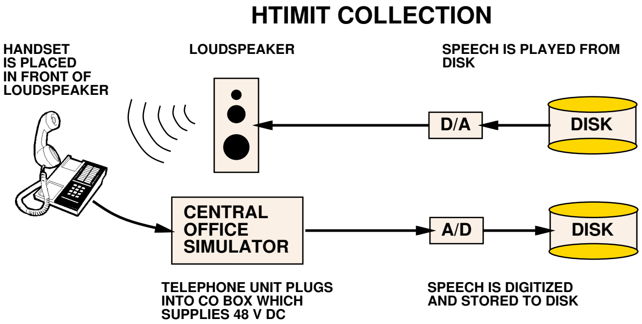

Figure 1: Collection setup for HTIMIT corpora.

For each of the 384 TIMIT speakers, his/her ten sentences are concatenated together and prepended with a half second 1kHz tone. The concatenated signal is then sent through the D/A and played out of a stereo loudspeaker. A handset attached to a telephone unit is fixed approximately 2–2.5 inches in front of the loudspeaker using a test stand, We chose not to use an "artificial mouth" for playing out the speech as was done for NTIMIT, since such devices are designed for single frequency tones, not multi-frequency signals like speech. To avoid any unanticipated degradations from telephone network or PBX transmission, the signal is collected directly by plugging the telephone unit into a "central-office" simulator box which supplies 48V DC power and taps off the signal for digitization. The signal is then digitized at 8kHz and stored back to disk. A correlation detector is finally used to detect the 1kHz prepended tone and align the start of the recorded utterances. The concatenated sentences are segmented back into sentence files using the known durations of the original TIMIT sentences. In addition to the TIMIT speech, a four second, 1kHz/sec sweep tone and five seconds of Gaussian white noise were also played through each handset as test signals.

Nine telephone handsets and one Sennheizer high-quality microphone were used (see Table 1). The Sennheizer was used as a control case to help measure the distortions introduced by the loudspeaker. The handsets consisted of four carbon-buttons, four electret and one cordless telephone handset. Most of the telephone handsets are not new (except el2) and were obtained from the Lincoln Telecom office. Handsets with obvious damage were not used, but in order to obtain some diversity with a limited number of handsets, handsets were selected to have variable sound characteristics, transducer designs or, in the case of electrets, different transducer grill designs. For example, cb1-cb3 have the same handset manufacture name (NT G-type) but the carbon-button transducer is different in each. Handsets cb3 and cb4 were selected because they had particularly poor (although not pathological) sound characteristics.

The HTIMIT collection procedure is obviously not ideal. First, the speech has been played through a loudspeaker which imposes some frequency response on the signal (although this will be a common factor among all recordings in this corpus). Second, the coupling of the transducer to the sound source is not realistic. As discussed in [3] and elsewhere, the direct airflow from a person's mouth, for example during plosive and fricative production, has noticeable effects on the transducer output. To address this second issue we collected the LLHDB using people speaking into the same handsets used in HTIMIT.

3. LLHDB CORPUS

The LLHDB corpus was collected by recording people speaking into the above nine handsets and the Sennheizer high-quality microphone using the setup shown in Figure 2. There were three types of speech recorded for each handset. First, the speaker read the *rainbow passage* [9], a ninety-seven word passage sometimes used in phonetic research. Second, the speaker read 10 sentences extracted from the TIMIT corpus[2]. Finally, the speaker was asked to describe a photograph[3] for approximately 40 seconds. To date we have 53 speakers (24 males and 29 females) in the LLHDB corpus.

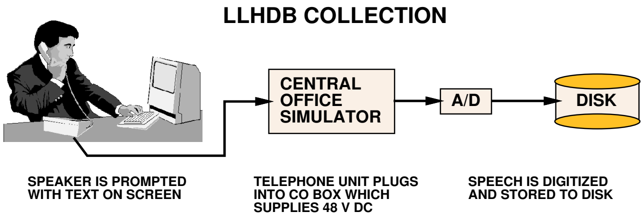

Figure 2: Collection setup for LLHDB corpora.

4. APPLICATION OF CORPORA

This section provides some initial analysis and applications of the above corpora. First, we give results from speaker identification experiments on the two corpora for various same and cross handset conditions showing the relative performance degradations and comparing results for the two different collection procedures. Next, we describe how a handset detector, built using the HTIMIT corpus, was successfully used to greatly improve performance on the Switchboard corpus under mismatched handset conditions.

4.1. Identification Experiments

For these experiments, the Gaussian mixture model (GMM) speaker recognition system was used [4]. The feature vector consisted of mel-scale cepstra, derived over the frequency band 300-3200 Hz, appended with delta cepstra from a 5

[2]Each speaker was was assigned to one of the TIMIT speakers and was prompted to read each of the TIMIT speaker's ten sentences

[3]A different photograph was used for each handset.

Table 1: Brief description of handsets used in corpora.

Handset Name	Description
senh	Sennheizer high-quality, head-mounted microphone
cb1	Northern-Telecom G-type carbon-button (center hole membrane granular container)
cb2	Northern-Telecom G-type carbon-button (6 hole metal granular container)
cb3	Northern-Telecom G-type carbon-button (6 hole membrane granular container)
cb4	ITT carbon-button (6 hole membrane/attached granular container)
el1	Northern-Telecom Unity electret (3-line grill)
el2	Northern-Telecom Unity Noisy-Environment electret (2-line grill)
el3	Unknown manufacture electret (64-hole grill)
el4	Radio Shack Chronophone-255 electret telephone
pt1	Sony portable (cordless) telephone

frame interval. Both cepstral mean subtraction and RASTA filtering were applied to minimize linear filter effects. Silence frames were discarded using an adaptive, energy-based speech detector.

For each speaker, a 32-order GMM was trained using the first (alphabetically) eight TIMIT sentences (approximately 24 seconds for training). The remaining two sentences were used as individual tests (approximately 3 seconds per test). Models were trained with speech from each handset and tested against speech from all handsets.

Results from the male speakers of the LLHDB corpus are given in Table 2. For male speakers, the average identification error rate for same-handset conditions was 6% and for cross-handset conditions was 30%. For female speakers, the average identification error rate for same-handset conditions was 17% and for cross-handset conditions was 45%. It not clear at this time why there is a discrepancy between the male and female results. As seen from the table, the pt1, cb3 and cb4 handsets performed particularly poorly against other handset data. For pt1 this is probably due more to the addition of RF noise from the handset to the base unit than to the transducer. As mentioned earlier, cb3 and cb4 are particularly poor sounding handsets with noticeable granular noise and non-linear distortions.

To compare results between the LLHDB and HTIMIT corpora, HTIMIT experiments were conducted using only 23 males and 28 females randomly selected from the 384 HTIMIT speakers. Speaker identification was run using the same experimental paradigm as was used for the LLHDB corpus. The results from the HTIMIT corpus (23 male speakers only) are given in Table 2. For male speakers, the average identification error rate for same-handset conditions was 22% and for cross-handset conditions was 39%. For female speakers, the average identification error rate for same-handset conditions was 21% and for cross-handset conditions was 37%. The male and female results are more consistent than was seen in the LLHDB experiments, but overall the HTIMIT performance was much lower than the LLHDB results. This additional degradation is probably due to the lower SNR of the HTIMIT speech, a result of the loose source-transducer coupling, and the interposed distortions from the loudspeaker. The same general trends within the handset types, however, seem to hold between the two corpora.

The performance difference between LLHDB and HTIMIT could be due to the particular speakers selected from HTIMIT for the initial experiments. Using the original TIMIT data for these speakers downsampled to 8kHz, we find identification error rates of 7% for the males and a surprising 23% for the females. The high error rate on the female speakers may be due in part to the loss of high frequency information from the 300-3200 Hz bandlimiting.

4.2. Handset Detector

Using the HTIMIT corpus, an automatic handset detector was also derived and applied to the Switchboard corpus. A maximum-likelihood classifier based on Gaussian mixture models was built to discriminate between speech originating from a carbon button handset and speech originating from an electret handset (Figure 3). A 1024-order GMM was trained using the cb3 and cb4 carbon button HTIMIT speech and a 1024-order GMM was trained using the el1 and el2 electret HTIMIT speech. Standard linear filtering compensation (cepstral mean subtraction and RASTA filtering) was applied to the features prior to model training. Since the models were trained with speech from the same speakers and had linear filtering effects removed, differences between the models should mainly be attributable to uncompensated transducer effects.

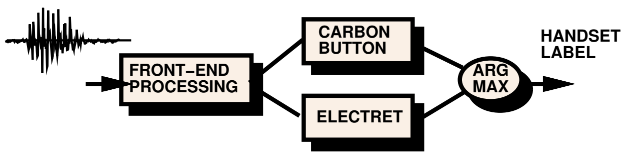

Figure 3: Handset Detector used on Switchboard corpus.

The handset detector was then used to label Switchboard utterances as either carbon-button or electret along with the *a posteriori* match probability. Audible checking of some of the labeling results, did indeed reveal a difference between the utterance marked as different handsets, with those marked as carbon-button having similar quality to the known carbon-button speech in LLHDB and likewise for those marked as electret.

Table 2: Identification error rates (in %) for the LLHDB and HTIMIT corpora (23 male speakers).

	Test																			
	LLHDB										HTIMIT									
Train	sen	cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1	sen	cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1
senh	11	20	28	57	50	33	24	28	39	57	13	22	24	43	33	20	22	26	24	35
cb1	15	2	11	43	26	13	7	13	17	26	22	17	22	46	26	22	20	26	24	30
cb2	22	11	4	41	24	13	9	9	22	43	15	22	15	48	37	22	30	20	20	35
cb3	70	63	74	13	33	65	67	67	65	70	72	76	22	22	28	78	74	70	74	80
cb4	41	30	43	15	7	37	50	39	43	37	70	65	59	33	26	65	65	65	63	65
el1	22	9	4	49	22	4	4	4	13	33	17	17	17	50	24	20	26	24	22	37
el2	22	17	24	54	39	20	4	13	24	35	39	30	37	57	39	35	30	35	33	50
el3	15	7	11	39	22	4	4	7	11	33	26	28	26	50	30	28	24	24	20	37
el4	24	4	22	54	24	13	11	9	2	30	28	26	28	46	24	35	22	33	22	33
pt1	46	37	48	50	46	39	41	39	37	4	50	46	50	70	52	50	52	50	50	26

In addition, the Switchboard handset labels were used to implement a handset dependent normalization technique for a speaker verification task on Switchboard [5]. The compensation technique basically attempts to normalize verification scores based on the handset type of the test utterance. Under the mismatched handset condition, in which test utterances came from conversations originating from different phone numbers than phone numbers used in a speaker model's training speech, this normalization technique reduced the false-alarm rate at a fixed 10% false-reject rate from 8% to less than 2% – a factor of 4 decrease – producing the best verification results in this evaluation. Although there is no ground truth to absolutely verify the veracity of the Switchboard handset labels, based on the audible verification and the dramatic performance improvement using the handset labels, we believe the handset detector is producing meaningful results.

5. SUMMARY

In this paper we have described two corpora collected at Lincoln laboratory for the study of handset effects. The HTIMIT corpora represents an artificial way of generating a large amount of speech passed through (potentially numerous) handsets, at the cost of an unrealistic acoustic coupling between the speaker's mouth and the transducers. The LLHDB corpora represents the traditional way of generating a smaller amount of actual speech passed through handsets, retaining the potentially important mouth-transducer acoustic coupling at the cost of more human labor. Preliminary results presented show that there is some difference in performance between the two corpora, but further analysis is required to determine the true signal differences.

We have also described one application of the HTIMIT corpus for developing a handset detector. This handset detector was applied to the Switchboard corpus and found to greatly improve performance for a speaker verification task under the difficult condition of mismatched handsets between training and testing. There are many other potential ways of using these handset corpora for improving the robustness of speech processing systems on telephone speech and for developing a better understanding of the distortions imposed by different transducers on the speech signal. It is hoped that the release of these corpora through the Linguistic Data Consortium (LDC) will spur on research in this important area.

6. REFERENCES

[1] A. E. Rosenberg, J. DeLong, C. H. Lee, B. H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification," ICSLP, pp. 599–602, November 1992.

[2] D. A. Reynolds, "The effects of handset variability on speaker recognition performance: Experiments on the switchboard corpus," ICASSP, pp. 113–116, May 1996.

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

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

[5] "March 1996 NIST speaker recognition workshop notebook." NIST administered speaker recognition evaluation on the Switchboard corpus, March 27–28 1996.

[6] 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.

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

[8] K. L. Brown and E. B. George, "CTIMIT: A speech corpus for the cellular environment with applications to automatic speech recognition," ICASSP, pp. 105–108, May 1995.

[9] F. Nolan, *The Phonetic Bases of Speaker Recognition.* Cambridge University Press, 1983.