# AUTOMATIC TRANSCRIPTION OF GENERAL AUDIO DATA: EFFECT OF ENVIRONMENT SEGMENTATION ON PHONETIC RECOGNITION[1]

*Michelle S. Spina and Victor W. Zue*

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
{spina,zue}@sls.lcs.mit.edu

## ABSTRACT

The task of automatically transcribing general audio data is very different from those usually confronted by current automatic speech recognition systems. The general goal of our work is to determine the optimal training strategy for recognizing such data. Specifically, we have studied the effects of different speaking environments on a phonetic recognition task using data collected from a radio news program. We found that if a single-recognizer is to be used, it is more effective to use a smaller amount of homogeneous, clean data for training. This approach yielded a decrease in phonetic recognition error rate of over 26% over a system trained with an equivalent amount of data which contained a variety of speaking environments. We found that additional gains can be made with a multiple-recognizer system, trained with environment-specific data. Overall, we found that this approach yielded a decrease in error rate of nearly 2%, with some individual speaking environments' error rate decreasing by over 7%.

## 1. INTRODUCTION

For many years, research in automatic speech recognition (ASR) has been driven by our desire to provide a speech-based input modality to computers, whether it be voice dialing (e.g., "Call home"), data entry (e.g., entering a credit card number), or document preparation. More recently, ASR research has broadened its scope to include the transcription of general audio data (GAD), from sources such as radio or television [2]. This shift in research focus is largely brought on by the growing need to shift content-based information retrieval from text to speech [6], so that the computer can satisfy requests such as, "Play me the speech by President Kennedy in which he said, 'Ich bin ein Berliner.'"

GAD pose new challenges to present-day ASR technology because they often contain extemporaneously-generated, and therefore disfluent speech, with words drawn from a very large vocabulary, and they are usually recorded from varying acoustic environments. Also, the voices of multiple speakers often interleave and overlap with one another or with music and other sounds. Since the performance of ASR systems can vary a great deal depending on speaker, microphone, recording conditions and transmission channel, we have argued that the transcription of GAD would benefit from a preprocessing step that first segmented the signal into acoustically homogeneous chunks [9]. Such preprocessing would enable the transcription system to utilize the appropriate acoustic models and perhaps even to limit its active vocabulary. Other researchers have investigated environment-specific techniques for acoustic training with varying results. In [8], Schwartz et al. found that environment-specific training did not prove to be beneficial for an automatic speech recognition task. They determined that gains made from general adaptation techniques applied during both training and testing were significantly larger and resulted in a simpler overall system. However, others [4] have found that environment-specific training did result in significant performance gains on the same task. It is therefore unclear how best to handle data with varying acoustic conditions. The goal of the research reported in this paper is to investigate some of the strategies for training a phonetic recognition system for GAD. The specific questions that we address in this paper are: 1) Can we expect performance gains on a phonetic recognition task when environment-specific information is utilized? 2) What are the trade-offs between recognition performance and the amount and quality of training data? The paper is organized as follows. First, we will describe the corpus that we have created for our experiments. Next, we will describe our experimental set-up and present the results of our various phonetic recognition experiments. Finally, we will conclude with a discussion of our results and an outline of our future plans.

## 2. CORPUS PREPARATION

We have chosen to focus on the *Morning Edition* (ME) news program broadcast by National Public Radio (NPR). NPR-ME is broadcast on weekdays from 6 to 9 a.m. in the US, and it consists of news reports from national and local studio anchors as well as reporters from the field, special interest editorials and musical segments. Since some of the segments are repeated hourly, we have chosen to record approximately 60 minutes of the program on a given day. While data are being collected weekly, the analysis presented in this paper are based on a collection of six hours of recording from November, 1996 to January, 1997.

Data was recorded from an FM tuner onto digital audio tape at 16kHz. A copy of the original recordings was then given to a local transcription agency, who produced orthographic transcriptions of the broadcasts in electronic form.

[1]This research was supported by DARPA under contract N66001-94-C-6040, monitored through Naval Command, Control and Ocean Surveillance Center. M. Spina also receives support from Intel Corporation.

The one-hour shows were transferred to computer disk using a DAT-Link+, and automatically split into manageable sized waveform files at silence breaks. In addition, if any of the resulting waveform files contained multiple sound environments (e.g., a segment of music followed by a segment of speech) they were further split at these boundaries. Therefore, each file was homogeneous with respect to sound environment. Orthographies and phonetic alignments were generated for each of the files using the orthographic transcriptions and a forced Viterbi search [10].

Seven categories were used to characterize the files. These categories were described in our previous work, and are briefly reviewed here: 1) clean speech: wideband (8kHz) speech from anchors and reporters, recorded in the studio, 2) music speech: speech with music in the background, 3) noisy speech: speech with background noise, 4) field speech: telephone bandwidth (4kHz) speech from field reporters, 5) music, 6) silence, and 7) garbage, which accounted for anything that did not fall into one of the other six categories. Figure 1 is a plot of the average spectra for each of the sound environments. Silence and field speech are visually distinct from other classes both in terms of energy and spectral shape. Music differs from speech in its fine harmonic structure. Differences in the average spectra of the other three speech categories are more subtle.

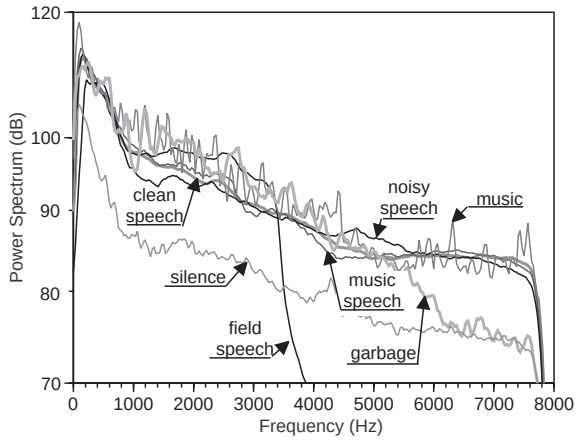


**Figure 1:** Average spectrum for each sound class

In [9], we described some preliminary analyses and experiments that we had conducted concerning the transcription of this data. For the NPR-ME corpus, we were able to achieve better than 80% classification accuracy for these seven sound classes on unseen data, using relatively straightforward acoustic measurements and pattern classification techniques. A speech/non-speech classifier achieved an accuracy of nearly 94%. The level of performance of such a classifier is clearly related to the ways in which it will serve as an intelligent front-end to a speech recognition system. The experiments done for this work attempt to determine if such a preprocessor is necessary, and if so, what level of performance is required for the sound segmentation.

# 3. PHONETIC RECOGNITION EXPERIMENTS

In this section, we describe our phonetic recognition experiments on the NPR-ME speech data under various training techniques. For this purpose, 4.25 hours of the NPR-ME data was used for system training, and the remaining hour was used for system test. Table 1 summarizes the amount of training and testing data (in minutes) available in each environment.

| Environment | Training Data | Testing Data |
|---|---|---|
| **Clean Speech** | 151.3 | 24.9 |
| **Music Speech** | 31.1 | 8.0 |
| **Noisy Speech** | 30.1 | 15.8 |
| **Field Speech** | 42.6 | 3.9 |

**Table 1:** Amount of training and testing data available (in minutes) for each speaking environment.

Acoustic models were built using the TIMIT [3] 61 label set. Results, expressed as phonetic recognition error rates, are collapsed down to the 39 labels typically used by others to report recognition results. The SUMMIT [5] speech recognizer developed by our group was used for these experiments. SUMMIT uses a segment-based framework for its acoustic-phonetic representation. The feature vector for each segment consisted of MFCC and energy averages over segment thirds as well as two derivatives computed at segment boundaries. Segment duration was also included. Mixtures of up to 50 diagonal Gaussians were used to model the phone distributions on the training data. For simplicity, only context-independent models were used. The language model used in all experiments was a phone bigram based on over four hours of training data. This particular configuration of SUMMIT achieved an error rate of 37.1% when trained and tested on TIMIT.

## 3.1. Single Recognizer System

The first question we asked ourselves is: if one is constrained to use only one recognizer for all four different types of speech material present in NPR-ME, how should the recognizer be trained? In this section, we try to determine the trade-offs between using a large amount of data recorded under a variety of speaking environments and a smaller amount of high quality data.

### *3.1.1. Multi-Style Training*

Multi-style training [7] incorporates different environment conditions in the training data, thereby reducing the potential mismatch between training and testing conditions. In addition to having a more diverse training set, we are able to utilize a large amount of data to train our acoustic models. To accomplish this, acoustic models were trained on the entire training set, which amounted to a total of 4.25 hours of speech training data. As shown in the first row of Table 2 the phonetic recognition performance on the test set varied widely across speaking environments, with the lowest error rate arising from clean speech (42.3%) and the highest error rate arising from

field speech (65.1%). The overall phonetic recognition error rate was 48.2%.

To reduce some of the channel differences between the training and testing data, we investigated using cepstral mean normalization [1] as a preprocessing technique. We found that the phonetic error rates were significantly reduced across all of the testing environments, as shown in the second row of Table 2. The resulting overall error rate for this experiment was 45.8%.

#### *3.1.2. Clean Speech Training*

Although multi-style training reduces the mismatch between the training and testing data, this approach may produce models that are too general for some of the testing environments. An alternative is to train models using only the clean, wideband speech material found in the training set. This amounted to a total of 2.5 hours of training data. Again, we found that the phonetic recognition performance on the test set varied across speaking environments. The overall phonetic recognition error rate was 47.7% for this experiment. Again, cepstral mean normalization reduced the error rate across all of the testing environments, resulting in an overall error rate of 45.2%. Since the benefits of cepstral mean normalization are independent of training technique, we continued to use this preprocessing technique for all subsequent experiments. These results are summarized in the third and forth rows of Table 2.

While the resulting error rates for the multi-style and clean speech training approaches were comparable, one must keep in mind that the multi-style approach utilized nearly 1.7 times the amount of data for training the acoustic models. To perform a fair comparison between these two approaches, we trained a multi-style system with an amount of training data equivalent to that of the clean speech system. We found this training approach substantially degraded our results to an overall error rate of 61.5%, an increase of 34%. This result suggests that it is advantageous to use only clean, wideband speech material for acoustic model training when data and computation availability becomes an issue.

| Training Data | Testing Data | | | | |
|---|---|---|---|---|---|
| | Clean Speech | Music Speech | Noisy Speech | Field Speech | Over All |
| **Multi-Style** | 42.3 | 51.8 | 51.8 | 65.1 | 48.2 |
| **With CMN** | 40.4 | 50.2 | 50.0 | 55.5 | 45.8 |
| **Clean Speech** | 39.8 | 58.2 | 50.9 | 63.9 | 47.7 |
| **With CMN** | 38.2 | 53.4 | 49.1 | 57.3 | 45.2 |

**Table 2:** Summary of phonetic recognition error rates for the multi-style and clean speech training systems. The multi-style system uses 1.7 times more data than the clean speech system.

### 3.2. Multiple Recognizer System

While the previous section has shown that segmenting the data to identify the clean, wideband speech is useful for training, we do not yet know if such a step would be useful for testing. In this section we explore the use of a multiple recognizer system for the phonetic recognition of NPR-ME, one for each type of speech material. These results will be compared to the single recognizer systems described in the previous section.

#### *3.2.1. Environment-Specific Baseline*

The environment-specific approach involves training a separate set of models for each speaking environment, and using the appropriate models for testing. In this section, we establish baseline performance by using the manually assigned labels for training and testing. This is equivalent to assuming that the environment segmentation has been done without error. Table 3 details the results in the form of a confusion matrix.

| Training Data | Testing Data | | | | |
|---|---|---|---|---|---|
| | Clean Speech | Music Speech | Noisy Speech | Field Speech | Over All |
| **Clean Speech** | 38.2 | 53.4 | 49.1 | 57.3 | 45.2 |
| **Music Speech** | 53.5 | 50.5 | 61.5 | 67.8 | 56.3 |
| **Noisy Speech** | 54.0 | 62.1 | 59.0 | 66.7 | 57.6 |
| **Field Speech** | 71.0 | 72.8 | 74.3 | 60.9 | 71.5 |

**Table 3:** Summary of phonetic recognition error rates for the environment-specific training system.

We achieve an overall error rate of 47.7%, computed from the diagonal entries of the confusion matrix. This result represents a 5.5% increase over the best single recognizer result. However, upon closer examination of the results, we find that this can largely be attributed to the 20% increase in error rate for noisy speech (from 49.1% to 59.0%). One possible explanation for this result is that many of the noisy speech utterances contain very low levels of background noise and this better fit the clean speech models. This is supported by our early results [9] showing that 53.5% of the noisy speech was misclassified as clean speech, indicating that perhaps these utterances should have been considered clean speech. Also, we can see from Table 1 that the test data contains a disproportional amount of noisy speech data, suggesting that the results may be skewed.

#### *3.2.2. Bandlimited Field Speech Models*

In all of the experiments conducted thus far, the field speech environment has consistently shown the highest phonetic recognition error rates. In an attempt to improve the recognition performance of the field speech, we bandlimited the training data by restricting our analysis to the frequency range of 133Hz to 4kHz. As shown in Table 4, we were able to achieve lower recognition error rates through bandlimiting the training data. Again, the use of clean speech to develop the acoustic models outperformed the use of multi-style data.

#### *3.2.3. Integrated System*

The experiments described thus far have assumed that test utterances have been classified perfectly. Our final experiment used the sound classification system described in [9] as a preprocessor to classify each test utterance as one of

| Training Data | Testing Data Field Speech |
|---|---|
| **Clean Speech** | 53.3 |
| **Music Speech** | 65.2 |
| **Noisy Speech** | 62.4 |
| **Field Speech** | 59.7 |
| **Multi-Style** | 53.7 |

**Table 4:** Summary of field speech phonetic recognition error rates for bandlimited training system.

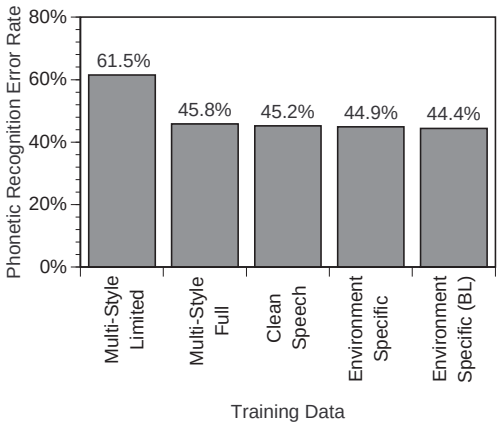


**Figure 2:** Summary of results for different training methods.

the four speech environments. The environment-specific model chosen by the automatic classifier for each utterance was then used to perform the phonetic recognition. This resulted in an overall error rate of 44.9%, which is slightly better than the best single recognizer result. If the bandlimited clean speech model was used for utterances classified as field speech, the overall error rate becomes 44.4%, which is 1.8% better than the best single recognizer result. These results indicate that our sound transcription conventions may need to be refined so as to more accurately label utterances with low-level background music or noise.

## 4. SUMMARY AND FUTURE WORK

This paper described experiments that we have conducted concerning the phonetic recognition of NPR-ME. We found that for all of the training techniques that we investigated the phonetic error rates varied widely across the NPR-ME speaking environments. By systematically exploring different system designs (one recognizer vs. multiple recognizers) and different training techniques, we were able to discover how each technique affected each environment.

We found that cepstral mean normalization was helpful for all training methods, reducing the recognition error rate by an average of 7.3%. Therefore it was included in all of our systems. We investigated the use of single and multiple recognizer systems, and different training techniques associated with each. The results of our experiments are summarized in Figure 2. If a single recognizer system is to be used, we found that training on a smaller amount of homogeneous, clean data was more effective than training with a large amount of data that contains a variety of speaking environments. Further gains can presumably be made with additional clean training data. A multi-style training approach was beneficial for only the music and field speech environments even though this method almost doubled the amount of data available for training the acoustic models.

Overall, a multiple recognizer system slightly outperformed a single recognizer system. We found that by testing with models trained on data with similar environments, we could decrease our phonetic error rate by nearly 2%. We feel that additional gains can be made with this technique with more environment-specific training data. For example, the error rate for music speech decreased by more than 5%, even though the amount of music speech training data available was less than 20% of that of the clean speech system. Bandlimiting the training data proved to be an effective method for significantly improving the recognition results for the field speech environment. We found that automatically selecting the environment-specific models did not degrade our results, and in fact improved them slightly. This indicates that low-levels of background music or noise may be better modeled with clean speech.

In future work, we intend to concentrate on improving the phonetic recognition results from the clean speech environment, and to investigate how the recognition of GAD compares to other automatic speech recognition tasks.

## 5. REFERENCES


[1] Anastasakos, T., Kubala, F., Makhoul, J., Schwartz, R. "Adaptation to new microphones using tied-mixture normalization," In *Proceedings of ICASSP-94*, pages 433-435, April 1994.

[2] Garofolo, J., Fiscus, J.G., Fisher, W.M. "Design and Preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora," In *Proceedings of DARPA Speech Recognition Workshop*, February 1997.

[3] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallet, D., and Dahlgran, N. The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM. NTIS order number PB91-505065, October 1990.

[4] Gauvain, J.L., Lamel, L., Adda-Decker, M. "Acoustic Modelling in the LIMSI Nov96 Hub4 System," In *Proceedings of DARPA Speech Recognition Workshop*, February 1997.

[5] Glass, J., Chang, J., and McCandless, M. "A probabilistic framework for feature-based speech recognition," In *Proceedings of ICSLP-96*, Philadelphia, PA, October 1996.

[6] James, David Anthony. *The Application of Classical Information Retrieval Techniques to Spoken Documents.* PhD thesis, Univ. of Cambridge, February 1995.

[7] Lippmann, R., Martin, E., and Paul, D. "Multi-style training for robust isolated-word speech recognition," In *Proceedings of ICASSP-87*, pages 4.1-4, 1987.

[8] Schwartz, R., Jin, H., Kubala, F., Matsoukas, S. "Modeling those F-conditions - Or not," In *Proceedings of DARPA Speech Recognition Workshop*, February 1997.

[9] Spina, M.S., and Zue, V.W. "Automatic Transcription of General Audio Data: Preliminary Analysis," In *Proceedings of ICSLP-96*, Philadelphia, PA, October 1996.

[10] Viterbi, A. "Error Bounds for Convolutional Codes and an Asymptotic Optimal Decoding Algorithm," In *IEEE Trans. on Information Theory*, 13:260-269, April 1967.