Automatic Transcription of General Audio Data: Preliminary Analyses[1]

Michelle S. Spina and Victor W. Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA

ABSTRACT

The task of automatically transcribing general audio data is very different from the transcription task typically required of current automatic speech recognition systems. The general goal of this work is to quantify the difficult issues posed by such data, thus leading to an understanding of how a speech recognition system may have to be altered to accommodate the added complexities. Specifically, we describe some preliminary analyses and experiments we have conducted on data collected from a radio news program. We found that using relatively straightforward acoustic measurements and classification techniques, we were able to achieve better than 80% classification accuracy for seven salient sound classes present in the data, and nearly 94% classification accuracy for a speech/non-speech decision. In addition, lexical analysis revealed that while the vocabulary size of a single broadcast is moderate, it grows exponentially as more shows are added.

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, television, or movies. This shift in research focus is largely brought on by the growing need to shift content-based information retrieval from text to speech [5], 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, the transcription of GAD can presumably benefit from a preprocessing step, in which the signal is first segmented into homogeneous chunks [1, 4, 6, 7]. This is because accurate sound segmentation will enable us to utilize acoustic models appropriate for each environment. Furthermore, knowing the particular nature of the speech material may help limit the active vocabulary. For example, if one could determine that a news broadcast concerns the traffic report, then one may be able to reduce the recognizer vocabulary only to those words relevant to the subject matter.

The goal of the research reported in this paper is to gain an in-depth understanding of the nature of GAD, in the hope of devising mechanisms that will lead to successful transcription of such data. The specific questions that we address in this paper are: 1) How many types of sounds are there that we can reliably distinguish in GAD? 2) How well can we segment the sound stream into these homogeneous segments? and 3) How does the size of the vocabulary change from one set of data to another? The paper is organized as follows. First, we will describe the data that we have collected, and the transcription and classification procedures that we have adopted. Next, we will present the results of our acoustic analyses and our subsequent sound classification experiments. Finally, we will describe some preliminary findings of our lexical analyses, and outline our future plans.

2. CORPUS PREPARATION

We have chosen to investigate the nature of GAD by focusing 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/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 at the rate of twice per week, the analyses presented in this paper are based on thirteen hours of recording – two consecutive days in July 1995, and 11 days during February and March, 1996.

Data were recorded from an FM tuner onto audio cassette tape and subsequently digitized at 16kHz and separated into 20 s files to ease management of computation. A copy of the original recordings was then given to a local transcription agency, who produced orthographic transcriptions of the broadcasts in electronic form. In addition to the words spoken, the transcripts also included side information about speaker identity, story boundaries, and acoustic environment. The convention for the transcription follows those established by NIST for the ARPA spoken language research community.

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

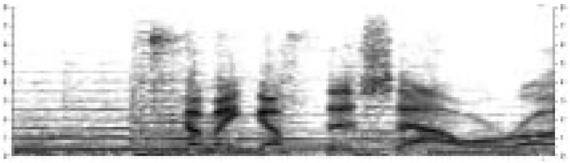

Figure 1: Spectrogram of a segment of music followed by speech superimposed on the background music.

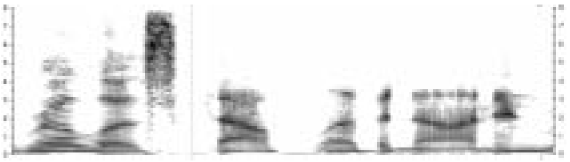

Figure 2: Spectrogram of a segment of clean speech followed by field speech.

Upon listening to some of the NPR-ME data, we reached the preliminary conclusion that there are seven logical categories into which the signal may be classified. These categories are: 1) clean speech (c_s) – wideband (8kHz) speech from anchors and reporters, recorded in the studio, 2) field speech (f_s) – telephone bandwidth (4kHz) speech from field reporters, 3) music speech (m_s) – speech with music in the background, 4) noisy speech (n_s) – speech with background noise, 5) music (m), 6) silence, (sil), and 7) garbage (gar), which accounted for anything that did not fall into one of the other six categories. Figures 1 and 2 show spectrograms illustrating the acoustic characteristics of some of these sounds.

To investigate the feasibility of automatically classifying the signal into these categories, we manually labeled the first two hours of our corpus with one of the seven labels once every 10 ms. The labeling was done through visual examination of spectrograms and through critical listening of the data. In the case of silence, a minimum duration of 150 ms was imposed so as to exclude stop closures. The first hour of the labeled data was used for algorithm development and training, whereas the second hour was set aside for testing. Throughout our investigation, we made heavy use of the Transcription View facility in SAPPHIRE [2], which can simultaneously display the waveform, spectrogram, actual transcription and classification output for each file.

Figure 3 shows the distribution of sound classes in the NPR-ME data. Studio-quality speech constitutes only about half of the entire corpus, and another 20% of the data contain speech superimposed with other sounds. Closer examination of the data revealed that silences occurred not only between speakers and stories, but also within sentences at natural, syntactic boundaries.

Figure 4 is a plot of the average spectra for each of the sound classes. 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, suggesting that confusions may result if these sounds were to be classified using purely spectral features.

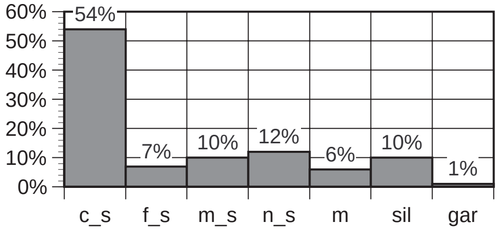

Figure 3: Distribution of sound classes in NPR-ME Data

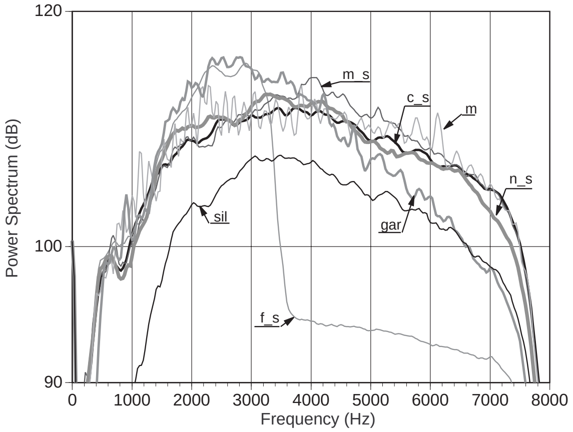

Figure 4: Average Spectrum for each sound class

3. SOUND SEGMENTATION

In this section, we report some preliminary experiments intended to automatically classify the sound stream into seven categories. For this purpose, we used the first hour of our manually-labeled data; approximately 40 minutes was used for training, and the remainder for system development. The audio files were shuffled before being split into the training and development sets to achieve similar distributions for the different classes in each set. The second hour of our labeled data was reserved for system testing.

The maximum *a-posteriori* probability approach was used to classify each frame into one of the seven sound categories. The acoustic models were represented by full Gaussian distributions whose mean vectors and covariance matrices were calculated from the training data. The *a-priori* probability for each sound class was inferred from the corresponding frequency of occurrence in the training set.

For acoustic modeling, fourteen mel-frequency cepstral coefficients (MFCC) were computed every 10 ms using a 20 ms Hamming window. To capture the longer-term spectral characteristics of each class, the feature vector for each frame was formed by averaging the MFCC of adjacent frames centered around the frame of interest. Experiments were performed to determine the optimal segment size. The number of frames included in the analysis segment was varied from 15 (7 frames on each side) to 81 (40 on each side). As shown in Figure 5, the classification accuracy on the development set increased steadily as more context was included in the analysis segment, eventually reaching a peak value of 76.5% (for an analysis segment of 51 frames). The accuracy then began to level off and decrease slightly, as the analysis segment began to include too much

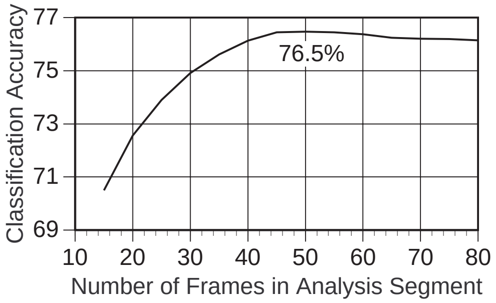

Figure 5: Accuracy as a function of the analysis segment size.

	c_s	f_s	m_s	n_s	m	sil	gar	*% of total*
c_s	92.0	0.1	2.2	4.8	0.2	0.8	0.0	54.4
f_s	0.7	97.7	0.0	0.3	0.1	1.2	0.0	14.9
m_s	38.1	0.0	48.2	9.3	4.2	0.2	0.0	6.7
n_s	53.5	0.4	10.6	30.0	3.0	1.7	0.8	7.6
m	4.1	0.7	6.1	7.5	78.7	1.9	1.0	4.9
sil	25.2	17.9	0.3	1.7	1.1	53.7	0.1	10.5
gar	2.8	3.4	3.6	11.8	69.9	5.6	2.8	1.0

Table 1: Confusion matrix on the sound classification experiments for the test set. The overall classification accuracy is 80.9%.

data from neighboring classes.

Examination of the results from the development set led to further refinement of the feature sets. First, many of the misclassified frames were found to contain small portions of neighboring classes in their analysis segments. To potentially alleviate this problem, MFCC averages across the first and last thirds of the analysis segment were added to the feature vector. Second, examination of the average spectra for each sound class indicated that the average spectral energy in a frame may be a distinguishing feature for the music speech, noisy speech and clean speech sound classes. Referring back to Figure 4, we can see that among the three most confused classes, music speech has the largest average spectral energy, followed by noisy speech and clean speech, respectively. Adding these measurements into the feature vector increased the classification accuracy to 80.0%. Finally, a bigram language model was added to model the sequential constraints of sound classes, resulting in a classification accuracy of 82.5% for the development set.

The classification algorithm we have developed was evaluated using the second hour of the labeled broadcast as an independent test set. Here, the hour-long show was first segmented into individual utterances, varying in length from 0.4 to 62.2 s, using silences of at least 250 ms as delineators. Using the optimal analysis segment size and measurement vector previously determined, the system achieved a classification accuracy of 80.9% on the test set, which is within 2% of the results for the development set. This experiment illustrates the robustness of the sound-class classifier, since comparable results were obtained with testing data recorded on another day.

Table 1 shows the confusion matrix for this experiment. The primary confusions were music speech and noisy speech with clean speech. One possible reason for this confusion is that some of the frames labeled as music speech or noisy speech actually contain very low level music or noise. While those frames may have been erroneously classified, such a mistake may not be very detrimental to the overall goal of providing an accurate transcription, since low-level acoustic disturbances may not affect the recognizer's performance significantly. Nevertheless, better acoustic measurements will clearly be helpful in reducing the errors. It is also possible that the system's performance will improve as more training data are used.

Given the fact that the non-speech sounds are all significantly different from speech sounds, we decided to perform an additional experiment to determine the separability of speech and non-speech frames. The speech class was formed from the union of the clean speech, field speech, music speech and noisy speech classes. The non-speech class consisted of the union of the music, silence and garbage classes. Using the measurements that achieved the best classification performance in the previous experiment, classification accuracies of 95.9% and 93.7% were achieved on the development and test sets, respectively.

4. TRANSCRIPTION ANALYSES

In order to perform content-based information retrieval of GAD, the speech material must first be transcribed using an ASR system. In addition to environmental factors that we have discussed earlier, the ASR system must be capable of handling very large vocabularies. In this section, we will describe some of the text-based analyses that we have begun to conduct, using the orthographic and ancillary transcriptions that accompanied the NPR-ME corpus. These contained the actual words spoken, the identity of the speaker, topic summaries, and story boundaries. Due to space limitations, we will only be able to summarize some of the findings.

Table 2 shows the general characteristics of the NPR-ME show, averaged over the thirteen hours that have been transcribed. There are some 14 music segments, each lasting about 15 s, which usually occur at story boundaries. The number of speakers for an hour-long show ranges from 33 to 65. Since there are about 24 stories in a show, each story typically involves 2-3 speakers. There are over 160 turn-taking events, suggesting that each turn (i.e., a contiguous segment of speech spoken by a given speaker) is just over 20 s. The speaking rate, inferred from the number of word tokens (nearly 10,000) and the fraction of the show containing speech (approximately 83%, or 50 min), is about 200 words per minute.

The working vocabulary of an hour-long show was found to be about 2,500 words, with the frequency of usage of these words highly skewed. The most frequently occurring 20% of the vocabulary words account for over 80% of the ones spoken. However, the least frequently occurring 50% of the vocabulary words are potentially the most important for understanding the content of the utterances (names, cities, etc.), and therefore would be most important to recognize in an automatic transcription system.

At first glance, it may appear that the vocabulary size for transcribing an NPR-ME show is quite manageable. Closer examination of the data, however, reveals otherwise. Figure 6 plots, on a log-log scale, the number of distinct words culled from the data (i.e., the recognizer's vocabulary) as a function of the total number of words encountered, as the number of shows increases from one to thirteen. The upper curve shows the cumulative sum of all the distinct words, and therefore represents the potential vocabulary of the recognizer.

	Average (over 13 shows)
# music segs	14.3
# speakers	45.7
# stories	24.2
# turns	162
# words spoken	9974.2
# vocab words	2515.7

Table 2: Summary of orthographic characteristics of the NPR-ME corpus, averaged over thirteen shows.

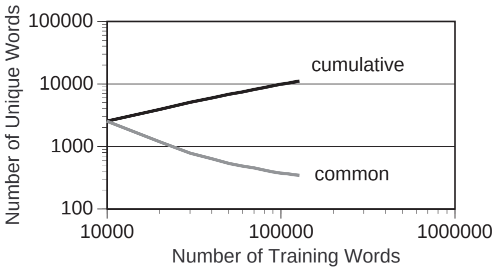

Figure 6: The number of distinct words as a function of the number of shows included in the NPR-ME corpus.

While the actual size of the vocabulary after 13 shows (approximately 10,000 words) is within the capabilities of current-day ASR systems, it is quite alarming that the *growth* of the vocabulary shows no sign of abating. If this trend were to continue, then the vocabulary that an ASR system must contend with will exceed 100,000 words if a whole year's worth of just this one show is to be transcribed and indexed. This trend is similar to, but slightly worse than those of the other large vocabulary corpora such as the Wall Street Journal corpus or the Switchboard corpus [3].

As more shows are included, the size of the *common* vocabulary across the shows will obviously decrease. This is illustrated by the lower curve in Figure 6, which indicates that less than 400 words occur in all of the thirteen shows, most of them being function words and generic words such as "news," "traffic," and "forecast." We have found that the trends revealed in this figure are independent of the order in which the shows are added.

5. SUMMARY AND FUTURE WORK

This paper described some preliminary analyses and experiments that we have conducted concerning the transcription of general audio data. For the NPR-ME corpus, we subjectively identified seven acoustically distinct classes based on visual and aural examination of the data. We were able to achieve better than 80% classification accuracy for these seven classes on unseen data, using relatively straightforward acoustic measurements and pattern classification techniques. A speech/non-speech classifier achieved an accuracy of nearly 94%. Lexical analysis of the transcription reveals that, while the vocabulary size of each show is moderate, it grows exponentially as more shows are added.

The level of performance needed for a sound classifier is clearly related to the ways in which it will serve as an intelligent front-end to a speech recognition system. If a speech/non-speech decision is all that is necessary, measurements that exploit the known regularities of speech should be used as the feature vector for such a system. If, on the other hand, more detailed classification distinguishing among several types of sounds and environments is required, then better acoustic measurements must be discovered and utilized. For example, a measurement that captures the fine harmonic structure exhibited in Figure 4 would be helpful in identifying music and music speech. In fact, it may be worthwhile to explore a hierarchical acoustic modeling scheme, in which the sounds are classified using a decision tree. One may also be able to improve classification performance by increasing the complexity of the acoustic models, e.g., using mixtures of Gaussians. The difference in accuracy between the development and test sets may be reduced by utilizing more training data, thus leading to more robust performance. Finally, the transcription conventions may need to be refined so as to more accurately label music speech or noisy speech.

Our preliminary analyses of the transcription of the NPR-ME corpus reveal some interesting characteristics of GAD. It contains many speakers and stories, with numerous turn takings, most of each lasting about 20 s. In this regard, GAD is very different from the data that the research community has collected. Our analyses of the unique words have serious implications for the transcription of general audio data. If the size of the vocabulary continues to grow unabated, conventional methodology using very large vocabulary speech recognition may prove infeasible, both in terms of computation and accuracy. It may be necessary to construct smaller, topic-specific vocabularies for individual stories within the broadcast, such as for the weather or traffic reports, and tackle the recognition problem separately. One may also need to alter the recognition strategy in some fundamental way – for example, by using syllables as a recognition unit – in order to solve this difficult, but increasingly important problem.

6. REFERENCES

1. Gopalakrishnan, P.S. et al. "Transcription of radio broadcast news with the IBM large vocabulary speech recognition system," In *Proc. DARPA Speech Recognition Workshop*, Feb., 1996.
2. Hetherington, I.L. and McCandless, M. "SAPPHIRE: An extensible speech analysis and recognition tool based on Tcl/Tk," *These proceedings.*
3. Hetherington, I.L. and Zue, V.W. "New words: Implications for continuous speech recognition," In *3rd European Conf. on Speech Comm. and Tech.*, Berlin, Germany, Sept., 1993.
4. Jain, U. et al. "Recognition of continuous broadcast news with multiple unknown speakers and environments," In *Proc. DARPA Speech Recognition Workshop*, Feb., 1996.
5. James, David Anthony. *The Application of Classical Information Retrieval Techniques to Spoken Documents."* PhD thesis, Univ. of Cambridge, Feb., 1995.
6. Kubala, F. et al. "Toward automatic recognition of broadcast news," In *Proc. DARPA Speech Recognition Workshop*, Feb., 1996.
7. Wegmann, S. et al. "Marketplace recognition using Dragon's continuous speech recognition system," In *Proc. DARPA Speech Recognition Workshop*, Feb., 1996.