Some Challenges of Developing Fully-Automated Systems for Taking Audio Comprehension Exams

David D. Palmer
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730
palmer@mitre.org

Abstract

Audio comprehension tests are designed to help evaluate a listener's understanding of a spoken passage and are frequently a key component of language competency exams. Just as reading comprehension exams are proving useful in evaluating text-based language processing technology, audio comprehension exams can be used to evaluate spoken language processing systems. In this paper we discuss some of the challenges of developing automated systems for taking audio comprehension exams.

1 Introduction

There is currently interest in using reading comprehension exams to evaluate natural language processing (NLP) systems. Reading comprehension tests are designed to help evaluate a reader's understanding of a written passage and are thus an example of a text-based language processing task. Audio comprehension tests, on the other hand, are designed to help evaluate a *listener's* understanding of a *spoken* passage and are an example of a spoken language processing task. These tests are frequently a key component of language competency exams, such as the Test of English as a Foreign Language (TOEFL) in the United States.

In this paper, we focus on some of the future challenges of developing fully-automated techniques for audio comprehension, in which the system developed processes the exam passages (and possibly questions) from the original audio source. Audio comprehension provides an excellent example of an understanding-based evaluation paradigm for speech systems, in which the emphasis is not solely on "getting all the words right" but rather on using speech recognition technology to automatically accomplish a task with a human benchmark: answering questions about a natural language story. The traditional paradigm for spoken language processing tasks, such as audio comprehension, has consisted largely of applying an existing text-based system to the hypothesis words output by an automatic speech recognition (ASR) system, ignoring the fact that information is lost due to recognition errors when moving from text to speech and the possibility that it can be regained in part via word confidence prediction.

We believe that successful approaches to audio comprehension will tackle the speech problem directly, by avoiding the use of features that are characteristic of written text and by explicitly addressing the problem of speech recognition errors through the use of smoothing techniques and word confidence information. Preliminary research in fully-automated techniques for reading comprehension, such as the Deep Read system developed by Hirschman *et al.* (1999), has included many standard NLP components, such as part-of-speech tagging, coreference/pronoun resolution, proper name finding, and morphological analysis (stemming). While the techniques that are being developed for reading comprehension provide a starting point, these techniques cannot be effectively applied to audio comprehension exams directly, because of the nature of differences between written and spoken language data. In this paper we address three specific challenges in developing audio comprehension system:

- Fundamental differences between text-based data and spoken language data (Section 2)
- Identifying proper names in "noisy" data (Sections 3)
- Dealing with out-of-vocabulary words (Sections 4)

In our discussion we will use examples taken from television and radio broadcast news, a "found" source of audio passages with a virtually unlimited vocabulary and a wide range of opportunities for audio comprehension evaluation. All ASR transcriptions we use will be actual output from a broadcast news ASR system with a word error rate of 30%.

2 Fundamental Data Differences

Beyond the obvious difference between a raw audio signal and a written text, the type of data output by a speech recognizer is fundamentally different from text-based data, even though they both consist primarily of words. On one level, there are im-

portant orthographic differences, since spoken language transcriptions lack many features present in written language. In addition, there is an inherent uncertainty in spoken language transcriptions, which almost always contain word errors. The degree of uncertainty is variable, since the word error rate (WER) of state-of-the-art speech recognizers can range from very low (1-5%) to very high (40-50%), depending on the domain (e.g., digit recognition vs. travel dialog vs. broadcast news vs. telephone conversations).

To illustrate some of the important differences between text-based and spoken language data, Figure 1 shows three versions of a sentence from a 1997 CNN news broadcast. The first version is the sentence as it would typically be written. The second version is the sentence as it would look as output from a "perfect" ASR system in which all spoken words are correctly transcribed (0% WER); we will discuss characteristics of this version in Section 2.1. The third version shows the actual output of a speech recognizer with a word error rate of 30%; in addition to the output hypothesis words, this version also shows word-level confidence scores. We will discuss the problems created by the word errors in Section 2.2.

2.1 Orthographic and Lexical Features

Consider the following (written) questions that may be asked about a spoken passage containing the example sentence in Figure 1, all of which can be easily answered by humans directly from the written passage:

> Who has been seeking Mr. Reineck?
>
> Whom have German authorities been seeking?
>
> How long have German authorities been seeking Mr. Reineck?

In comparing the "clean" transcription in Figure 1 and the written questions above to the perfect (0% WER) transcription, there are several differences that are immediately evident. These differences impact the tokenization of the data as well as the lexical representation of words, which will affect the ability of an audio comprehension system to relate words in a written question[1] to ASR transcriptions.

Lack of capitalization and punctuation: In many languages, including English, capitalization and punctuation, such as periods, commas, and quotation marks, provide important information about sentence/utterance boundaries and the presence of proper names (which we will discuss in Section 3). However, ASR output is usually caseless, such that boundaries and names (e.g., "REINECK") are not as easy to identify as in written language.

Most abbreviations are spelled out: Related to the lack of punctuation, ASR output does not usually contain abbreviations ("MISTER" vs. "Mr.").

Numbers are spelled out: Types of numbers in ASR data are not as easy to recognize ("NINETEEN NINETY TWO" vs. 1992). Tokenization of numbers is very different in ASR output, as a single written token like "$163.75" that is easily recognizable as a monetary amount can result in a large number of ASR output words "ONE HUNDRED SIXTY THREE DOLLARS AND SEVENTY FIVE CENTS," which is not immediately identified as a single quantity.

Presence of Disfluencies: Though not present in this example, spoken language frequently contains disfluencies, such as pause fillers ("UH", "UM"), word fragments, and repetitions, that are not present in written language. For example, the person reading the passage may actually say "MISTER REIN- UH REINECK," making successful processing of the output more difficult.

Effective audio comprehension systems will need to normalize all text-based and spoken language data to address orthographic differences.

2.2 Uncertainty in Speech Transcriptions

One of the primary factors that distinguishes text-based language processing tasks, such as reading comprehension, from spoken-language processing tasks, such as audio comprehension, is the uncertainty inherent in the word sequence output by the speech recognizer. The sequence of output words is rarely the same as the actual spoken word sequence, due to word substitution, insertion, and deletion errors. This uncertainty is clear in the third version of the sentence in Figure 1; of the eleven spoken words, three (27.3%) of the corresponding ASR output words are incorrect (SINKING, IS, ARRIVING). Audio comprehension systems that process this "noisy" third version as if it contained the actual spoken words could not possibly answer any of the sample questions above correctly, since the most important words (MISTER REINECK) are not present in the output.

Hirschman *et al.* (1999) report initial results in developing a reading comprehension system using a "bag of words" approach, in which the sentences in a passage that are deemed most likely to contain the answer are those with the maximum lexical overlap with the question, without regard for word order within the sentence. Recognition word errors would obviously adversely affect such an approach applied to audio comprehension; in cases where the words in the answer to the question were misrecognized, the system would be incapable of answering

[1]Spoken questions would first need to be transcribed by the ASR system and will be addressed in Section 2.2.

correctly. In the case of spoken questions, an additional layer of uncertainty is present since the recognizer may output different hypothesis words for the same word in a question and in a spoken passage; for example, "Reineck" was also misrecognized elsewhere in the same news story as "RIGHT AT, " "RYAN AND," "RYAN EIGHT," "REINER," and "RUNNING AND."

One of the possible ways to address this lexical overlap problem is to expand the set of candidate words: rather than restricting processing to the single best recognizer hypothesis sequence, we can allow the top N hypothesis sequences (known as the "N-best list"). In the example of Figure 1, if (SEEKING, MISTER, and REINECK) are alternative hypotheses for the incorrect (SINKING, IS, and ARRIVING) somewhere in the N-best list, the bag of words approach would at least have a chance of answering the question correctly.

While the bag of words approach is a simple technique providing an initial baseline result, "deeper" understanding of reading (and audio) comprehension passages will require modeling of the sequential nature of the language. Statistical language modeling, an essential component of most state-of-the-art speech recognition systems, seeks to estimate the probability of the sequence of L spoken words, $P(w_1...w_L)$. The language modeling within the ASR system contributes to the output word sequence, but the actual recognizer output is usually not the original sequence $w_1...w_L$, but instead a sequence of M words $h_1...h_M$, where M may not necessarily be the same as L and where $P(h_1...h_M) \neq P(w_1...w_L)$. Systems processing ASR output data must therefore effectively model the difference between the actual sequence $w_1...w_L$ and the hypothesized sequence $h_1...h_M$.

One way to account for word errors in the ASR output sequence $h_1...h_M$ is by integrating word-level confidence scores into the model of the word sequence. This word-level confidence score, which is a number between 0 and 1 produced by many current automatic speech recognition systems, is an estimate of the posterior probability that the word output by an ASR system is actually correct. As such, it provides us with important information about the output transcription that can assist error detection. The third version of the sentence in Figure 1 also includes the word confidence scores that were produced with the output word sequence. In this particular example, the word confidence scores are an excellent indication of the presence of word errors, since the three word errors (SINKING, IS, and ARRIVING) also have the three lowest confidence scores (.14, .09, and .21). Unfortunately, though confidence scores are a good indication of correctness, it is not always this straightforward to distinguish the errors from correctly transcribed words.

3 Robust Name Finding

Extracting entities such as proper names is an important first step in many systems aimed at automatic language understanding, and identifying these types of phrases is useful in many language understanding tasks, such as coreference resolution, sentence chunking and parsing, and summarization/gisting. The targets of proper name finding, names of persons, locations, and organizations, are very often the answers to the common "W-questions" Who? and Where? A common definition of the extended name finding task, known as the "named entity" task, also includes numeric phrases, such as dates, times, monetary amounts, and percents, which are often the answers to other common questions When? and How Much? Identifying named entities in passages should thus help in reading/audio comprehension. In fact, Hirschman *et al.* (1999) report that identifying named entities in reading comprehension passages and questions consistently improves the performance of their system, even when the name recognition has an accuracy as low as 76.5%. We would expect name recognition to also be a very important component of any audio comprehension system.

Figure 2 shows an example of the importance of names in a news story. This example again shows three versions of a sentence from the news. The first version shows the ASR output for a sentence. Due to the word errors "OUR STRAWS YEAR BEHIND IT", an audio comprehension system would be unable to answer most simple questions such as:

> Who is shown on a T-shirt with a sledgehammer?
>
> Where is Jörg Haider from?

However, the second version in Figure 2 shows that if we know that "OUR STRAWS" is a location phrase and that "YEAR BEHIND IT" is a person phrase (albeit incorrectly transcribed), we could at least know *where* in the passage to find the answer to the Who? and Where? questions, since the other words in the sentence are correctly transcribed. This information could be used, for example, to consult other corresponding word sequences in the N-best list or word lattice in which the words "Austria's Jörg Haider" may have been correctly transcribed. In this case "Haider" is an out-of-vocabulary word and would not be present elsewhere in the N-best list; we will discuss this problem in Section 4.

3.1 Name finding techniques

Finding names in text-based sources such as newspaper and newswire documents has been a focus of research for many years, and some systems have

reported performance approaching human performance (96-98%) on the named entity task. Finding names in speech data is a very new topic of research, and most previous work has consisted of the direct application of text-based systems to speech data, with some minor adaptations.

For the range of word error rates common for most large vocabulary ASR systems (< 30%), all the named entity models we will describe in this section produce performance between 70-90%. This is comparable to or better than the accuracy (76.5%) of the named entity system that Hirschman *et al.* (1999) report improves their reading comprehension system. However, there is significant room for improvement of the speech data NE systems. Previous work has found that the absence of capitalization and punctuation information in speech transcriptions results in a 2-3% decrease in name finding performance(Miller et al., 1999), and this degradation is greater in the presence of word errors. The decline in NE performance for text-based systems applied directly to errorful speech data is roughly linear with increase in WER, although the NE performance degrades more slowly than the WER, i.e. each recognition error does not result in an NE error. One of the goals of work directly on speech understanding models should be to improve this linear degradation.

One example of a trainable text-based system that has been applied successfully to speech recognizer output is described by Bikel *et al.*(1999). Each type of entity (person, location, etc.) to be recognized is represented as a separate state in a finite-state machine. A bigram language model is trained for each phrase type (i.e., for each state), and Viterbi-style decoding is then used to produce the most likely sequence of phrase labels in a test utterance. This model incorporates non-overlapping features about the words, such as punctuation and capitalization, in a bigram back-off to handle infrequent or unobserved words. Specifically, each word is deterministically assigned one of 14 non-overlapping features (such as two-digit-number, contains-digit-and-period, capitalized-word, and all-capital-letters), and the back-off distribution depends on the assigned feature. The approach has resulted in high performance on many text-based tasks, including English and Spanish newswire texts. Despite the fact that the original model relied heavily on text-based features such as punctuation and capitalization in the language model back-off, it gives good results on speech data without modifying anything but the training material (Miller et al., 1999).

A closely related statistical approach to named entity tagging specifically targeted at speech data was developed at Sheffield by Gotoh and Renals (2000). In their model, named entity tags are treated as categories associated with words, effectively expanding the vocabulary, e.g. a word that might be both a person and a place name would be represented with two different lexical items. An n-gram language model is trained on these augmented words, using a single model for joint word/tag dependence on the history rather than the two components used in the Bikel model and thus representing the class-to-class transitions implicitly rather than explicitly. A key difference between the approaches is in the back-off mechanism, which resembles a class grammar for the Sheffield system. In addition, the Sheffield approach uses a causal decoding algorithm, unlike the Viterbi algorithm which delays decisions until an entire sentence has been observed, though this is not a restriction of the model. The extended-vocabulary n-gram approach has the advantage that it is well-suited to using directly in the ASR search process.

Palmer, Ostendorf, and Burger (1999; 2000) use a model similar to other probabilistic name finding models, with several important differences in the model topology and the language modeling technique used. A key difference in their approach is that infrequent data is handled using the class-based smoothing technique described in (Iyer and Ostendorf, 1997) that, unlike the orthographic-feature-dependent back-off, allows for ambiguity of word classes. They describe methods for incorporating information from place and name word lists, as well as simple part-of-speech labels, and thus account for the fact that some words can be used in multiple classes. Their results for high error rates (28.2) are slightly better than the simple back-off, suggesting that the POS smoothing technique is more robust to ASR errors. In addition to the robustness provided by the class-based smoothing, they also report initial success in integrating word confidence scores into their model to further improve the robustness of the system to speech recognition errors.

4 Out-of-Vocabulary Words

Historically, the goal of automatic speech recognition (ASR) has been to transcribe the sequence of words contained in an audio stream. State-of-the-art speech recognition systems model this problem using a probabilistic formulation in which the most likely sequence of words is produced given a sequence of acoustic features derived from the raw utterance audio signal. While this approach has been very successful, the model has a serious limitation: it can only produce output hypotheses from a finite list of words that the recognizer explicitly models. This list of possible output words is known as the system *vocabulary*, and any spoken word not contained in the vocabulary is referred to as an out-of-vocabulary (OOV) word. Every OOV word in the input utterance is guaranteed to result in one or more output

errors.

As we discussed in Section 2.2, ASR output word errors, especially from spoken names, will adversely affect audio comprehension performance. However, the methods for dealing with errors that we discussed in previous sections, such as using N-best list output, can only compensate for misrecognitions of *known* words. Since OOV words will never appear in the hypothesized N-best output, other methods are necessary for accounting for their presence in the input audio stream. Figures 1 and 2 both have examples of words that were out-of-vocabulary (Reineck, Haider) for the particular ASR system. Figure 3 shows another example, in which several names are out-of-vocabulary (Brill, Salif, Keta, Nusa, Fateh, Ali-han).

Some examples of questions that might be asked about this passage are:

> Which two musicians did Wally Brill discover?
>
> Where is vocalist Salif Keta from?
>
> Who got turned onto Keta and Ali-han's music?

Clearly, these questions could not be answered directly from the actual output due to the word errors. In fact, identifying likely proper names in the output, as we discussed in Section 3, would also be inadequate, because the output word error "DECATUR" might be mistaken for the answer to a Where? question, and "MISTER FUNG" might be mistaken for the answer to a Who? question. An effective method for dealing with out-of-vocabulary words is thus necessary.

4.1 Increasing ASR Vocabulary

One approach to the OOV problem might be to increase the vocabulary size of the ASR system. Speech recognition systems can have a range of vocabulary sizes, depending on the target domain, the generality required, as well as the availability of computational resources. For example, many research systems designed for constrained environments, such as real-time travel information dialog, use a vocabulary size as small as 1,000-5,000 words. On the other hand, current research systems for unconstrained tasks such as the transcription of broadcast news programs frequently have vocabularies between 25,000 and 64,000 words. Increasing the vocabulary size of a speech recognition system can result in lower error rates, in part by decreasing the percentage of OOV words in the input utterance. However, systems with larger vocabularies require more memory and run slower than those with smaller vocabularies. Since practical ASR systems cannot have unlimited memory and computational requirements, they naturally cannot have unlimited vocabulary sizes.

In addition to increased computational cost, adding words to a vocabulary increases the potential confusability with other vocabulary words. In fact, Rosenfeld (1995) reports that a vocabulary size around 64,000 is nearly optimal for processing read North American Business news, and that increasing the vocabulary size beyond this yields negligible recognition improvement at best. The optimal vocabulary size is also domain dependent: a 64,000 word vocabulary may not be necessary for travel dialog but may be inadequate for directory assistance. Rosenfeld's analysis shows that increasing the system vocabulary size can help recognition rates for many common words while hurting for less common words. Yet the less common words, such as new names introduced as a result of national and international events, usually contain more semantic information about the utterance, and these errors are much more costly for language understanding applications. Since new words are constantly being introduced into common usage, it is impossible to ever have a complete vocabulary of all spoken words, and the treatment of new lexical items is thus an essential element of any system aiming to process natural language.

Hetherington (1995) conducts an extensive empirical study of the out-of-vocabulary problem in his PhD thesis. He presents a demonstration of the magnitude of the OOV problem for a wide range of multilingual natural language corpora and shows that some tasks can require vocabularies larger than 100,000 words to reduce the OOV rate below 1%. He shows that even an OOV rate of 1% results in 15-20% of all utterances containing unknown words. He also produces experimental results of the effect that unknown words have on speech recognition output, showing that, on the average, each OOV input word results in 1.5 actual word errors. Of the errors resulting from OOV words, 20% of these word errors result from in-vocabulary words being misrecognized due to their proximity to an unknown word. This work demonstrates the need for OOV word handling in any speech recognition system.

4.2 Dynamic Vocabularies

The need for unlimited spoken language vocabulary despite a limited ASR vocabulary suggests an alternative view of large-vocabulary spoken language processing, in which rather than trying to include all possible words in the ASR vocabulary we instead develop techniques for dynamically adapting the overall audio comprehension system vocabulary using lexical resources, without requiring a larger ASR vocabulary and the problems this entails.

Geutner *et al.* (1998) describe a multi-pass decoding approach targeted at reducing the out-of-vocabulary rates for heavily inflected language, such as Serbo-Croatian, German, and Turkish. Their

work attempts to dynamically expand the effective vocabulary size by adapting the recognition dictionary to each utterance. In the first recognition pass, an utterance-specific vocabulary list is constructed from the word lattice. They then use a technique they call "Hypothesis Driven Lexical Adaptation" to expand the vocabulary list by adding all words in a full dictionary that are sufficiently similar to those in the utterance list, where "similarity" is determined by the morphology and phonetics of the words. An automatic process then creates a new utterance recognition vocabulary and language model from the expanded vocabulary list , and a second recognition pass is performed using the expanded models. Geutner *et al.* report that the lexical adaptation methods result in a significant decrease of up to 55% in OOV rates for the inflected languages, and that this improvement in OOV rate results in an improvement in the recognition rate of 3-6% (absolute).

Geutner's multi-pass approach requires vocabulary adaptation and re-recognition of each complete utterance. The importance of name finding in audio comprehension suggests an alternative to this approach that would allow more targeted re-recognition of partial utterances. As the example in Figure 2 showed, it is possible to determine the data regions that contain the potential answers, even when the words themselves are misrecognized. For written questions, phonetic information from the question and hypothesis words can be used to help repair key misrecognitions. For example, using phonetic information, it is possible to relate "vocalist Salif Keta" in a question to "VOCALIST SELL THE DECATUR" in the ASR output. This information can be supplemented with external lexical resources, such as word lists for the appropriate type of proper name, to expand the set of possible hypotheses within the region. Large lists of names are available publicly that could be used for this purpose; for example, the U.S. Census publishes a ranked list of the most common surnames and first names in the United States, most of which are OOV words for current ASR systems. Once a region in the ASR output is identified as an OOV person, the Census data could be used to correct the OOV errors. This would then allow the audio comprehension system to answer more Who? questions correctly.

5 Conclusions

Just as research in reading comprehension can help evaluate text-based NLP systems against a human benchmark, audio comprehension can provide a useful task for evaluating speech understanding systems. Audio comprehension provides an understanding-based evaluation paradigm for speech systems that encourages research on a useful spoken language understanding application rather than on "getting all the words right." The techniques developed for audio comprehension promise to be widely useful in many language understanding area.

References

D. Bikel, R. Schwartz, R. Weischedel, "An Algorithm that Learns What's in a Name," *Machine Learning,* 34(1/3):211-231, 1999.

P. Geutner, M. Finke, P. Scheytt, "Adaptive Vocabularies for Transcribing Multilingual Broadcast News," *Proc. International Conference on Acoustic, Speech and Signal Processing,* 1998.

Y. Gotoh, S. Renals, "Information Extraction From Broadcast News," *Philosophical Transactions of the Royal Society*, series A: Mathematical, Physical and Engineering Sciences, vol.358, issue 1769, April 2000.

I.L. Hetherington, "A Characterization of the Problem of New, Out-of-Vocabulary Words in Continuous-Speech Recognition and Understanding," PhD Thesis, Massachusetts Institute of Technology, 1995.

L. Hirschman, M. Light, E. Breck, J. Burger, "Deep Read: A Reading Comprehension System," *Proc. 37th Annual Meeting fo the Association for Computational Linguistics (ACL99)*, pp. 325-332, 1999.

R. Iyer and M. Ostendorf, "Transforming Out-of-Domain Estimates to Improve In-Domain Language Models," *Proc. European Conference on Speech Comm. and Tech.,* Vol. 4, pp. 1975-1978, 1997.

D. Miller, R. Schwartz, R. Weischedel, R. Stone, "Named Entity Extraction from Broadcast News," *Proc. DARPA Broadcast News Workshop,* pp. 37-40, 1999.

D. Palmer, M. Ostendorf, and J. Burger, "Robust Information Extraction from Spoken Language Data," *Proc. European Conference on Speech Comm. and Tech.,* pp. 1035-1038, 1999.

D. Palmer, M. Ostendorf, and J. Burger, "Robust Information Extraction from Automatically Generated Speech Transcriptions," *Speech Communication*, in press, 2000.

P. Robinson, E. Brown, J. Burger, N. Chinchor, A. Douthat, L. Ferro, and L. Hirschman, "Overview: Information extraction from broadcast news," *Proc. DARPA Broadcast News Workshop,* pp. 27-30, 1999.

R. Rosenfeld, "Optimizing Lexical and N-gram Coverage Via Judicious Use of Linguistic Data." *Proc. European Conference on Speech Comm. and Tech.,* volume 2, pages 1763–1766, 1995.

M. Siu and H. Gish, "Evaluation of word confidence for speech recognition systems," *Computer Speech & Language*, Vol. 13, No. 4, Oct 1999, pp. 299-319

German authorities have been seeking Mr. Reineck since 1992.

GERMAN AUTHORITIES HAVE BEEN SEEKING MISTER REINECK
SINCE NINETEEN NINETY TWO

GERMAN(.74) AUTHORITIES(.90) HAVE(.79) BEEN(.82) SINKING(.14) IS(.09) ARRIVING(.21)
SINCE(.60) NINETEEN(.90) NINETY(.95) TWO(.94)

Figure 1: Example of text-based vs. spoken language differences: a written sentence and its ASR transcriptions (WER 0% and 30% with word confidence scores).

THE SHIRTS SHOW OUR STRAWS YEAR BEHIND IT WITH A SLEDGEHAMMER AND A RACIST CAPTION

THE SHIRTS SHOW [location] [person] WITH A SLEDGEHAMMER AND A RACIST CAPTION

The T-shirts showed Austria's Jörg Haider with a sledgehammer and a racist caption

Figure 2: Example showing the importance of names: ASR output (30% WER) for a sentence, the same ASR sentence with locations of proper names labeled, and the correct transcription.

Then a few years ago, Wally Brill got turned onto the music of West African vocalist Salif Keta and the haunting sounds of the late Nusa Fateh Ali-han.

IN A FEW YEARS AGO ALWAYS REAL GOT TURNED ON TO THE MUSIC OF WEST AFRICAN VOCALIST SELL THE DECATUR AND THE HAUNTING SOUNDS OF THE LATE MISTER FUNG'S ALLEY

Figure 3: Example of ASR output showing numerous OOV name errors.