CORNERSTONE
MINNESOTA STATE UNIVERSITY MANKATO

Journal of Undergraduate Research at Minnesota State University, Mankato

Volume 17 Article 1

2017

Improving Speech Recognition for Interviews with both Clean and Telephone Speech

Sung Woo Choi
Minnesota State University, Mankato, sung.choi-1@mnsu.edu

Improving Speech Recognition for Interviews with both Clean and Telephone Speech

Sung Woo Choi
Integrated Engineering
Minnesota State University, Mankato
Mankato, Minnesota 56001 USA

Faculty Advisor: Dr. Rebecca Bates

Abstract

High quality automatic speech recognition (ASR) depends on the context of the speech. Cleanly recorded speech has better results than speech recorded over telephone lines. In telephone speech, the signal is band-pass filtered which limits frequencies available for computation. Consequently, the transmitted speech signal may be distorted by noise, causing higher word error rates (WER). The main goal of this research project is to examine approaches to improve recognition of telephone speech while maintaining or improving results for clean speech in mixed telephone-clean speech recordings, by reducing mismatches between the test data and the available models. The test data includes recorded interviews where the interviewer was near the hand-held, single-channel recorder and the interviewee was on a speaker phone with the speaker near the recorder. Available resources include the Eesen offline transcriber and two acoustic models based on clean training data or telephone training data. The Eesen offline transcriber is on a virtual machine available through the Speech Recognition Virtual Kitchen and uses an approach based on a deep recurrent neural network acoustic model and a weighted finite state transducer decoder to transcribe audio into text. This project addresses the problem of high WER that comes when telephone speech is tested on cleanly-trained models by 1) replacing the clean model with a telephone model and 2) analyzing and addressing errors through data cleaning, correcting audio segmentation, and adding words to the dictionary. These approaches reduced the overall WER. This paper includes an overview of the transcriber, acoustic models, and the methods used to improve speech recognition, as well as results of transcription performance. These approaches reduced the WER on the telephone speech by 16.3 to 33.2% depending on the talker. Future work includes applying a variety of filters to the speech signal which could reduce both additive and convolutional noise resulting from the telephone channel.

Keywords: Speech Recognition, Noisy Data

1. Introduction

The primary goal of automatic speech recognition (ASR) is to efficiently and accurately transcribe an audio speech signal, turning speech into written text. ASR is especially useful for storing business meetings and medical records in text form. ASR mimics a key function of the human brain which is complex with innumerous and ceaseless tasks being processed simultaneously in parallel. Therefore, developing programs similar to human language processing is intricate and similarly complex. In the last ten years, ASR performance has improved with the use of deep neural networks.[1] Research in designing systems to understand spoken speech correctly has a history of more than 50 years, but there are still many tasks to improve ASR systems.

The goal of this project was to improve the speech recognition of recordings that include both clean and telephone speech. Unlike clean speech, telephone speech is band-pass filtered for ease of telecommunication which in return limits available frequencies, providing less computational data to the system. Telephone speech signals may also be distorted by noise, causing higher word error rates (WER).

Since systems are designed and trained for certain domains in specific environments, if several different types of signals are simultaneously present in one recording, then there exists a mismatch between training data and testing data. If the environment of the recording is different from that of the training data, there is an acoustic mismatch. For example, training data is recorded in a lecture room with clean speech without any noise, while testing data is recorded in a room with an open window, adding the sound of wind and cars passing nearby to the recording. Again, there is acoustic mismatch. In a domain mismatch, test and training data are on different topics. Then the system might not perform well because words are not in the lexicon or dictionary. Additionally, data may differ because of style mismatch, whether conversational, spontaneous, or planned, formal or informal.

Accordingly, if an acoustic model trained with clean, planned speech is used to transcribe spontaneous telephone speech, there are mismatches between the training and testing data. For this reason, various types of acoustic and language models are designed to address environment, domain and style mismatch. For this work, acoustic models trained on TEDLIUM[2] and Switchboard[3] data are used because they are easily available and match some of the test environment. The TEDLIUM model is trained using planned, formal speech recorded in a clean environment, while the Switchboard model is trained with spontaneous, informal speech recorded over telephone lines. The TEDLIUM model is the default model for the freely available Eesen recognition system.

2. Background

In this section, the general process of speech recognition is explained and the ASR system and toolkits that are used to execute speech recognition in this project are introduced.

2.1. Speech Recognition Process

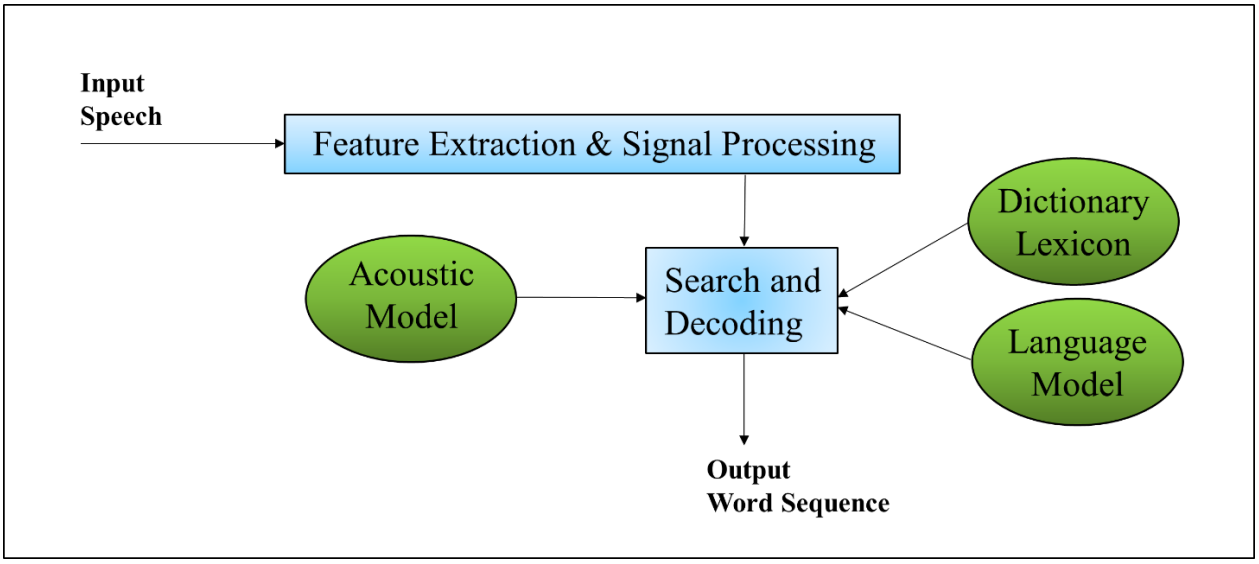

Figure 1: Speech Recognition Process

Figure 1 shows the process of moving from input speech to an output word sequence. Speech recognition requires two major steps: signal processing and feature extraction, followed by the search and decoding process. Signal processing is used to segment the signal of input speech into many utterances for feature extraction.[4] The speech is split at speaker changes and silences. The output of this stage is a digital representation that captures key distinguishing information about speech sounds. To perform the recognition in the search and decoding process, a

dictionary or lexicon (a word list with associated pronunciations) and two models are needed, one representing the acoustics and one representing language through sequences of words. Using the data from signal processing, the acoustic model statistically represents the relationship between words in the audio signal and the phonetic units of the speech, while the language model provides information about the probability of word sequences. After these two basic steps, a hypothesized word sequence of the input speech is produced as an output.

In speech recognition, the acoustic model represents the probability of the spoken sounds given words, *P(a|W)*, and the language model represents the probability of words, particularly given previous words, *P(W)*. The goal is to find the maximum *P(W|a)* for a specific word string w_1, w_2, … w_n and then select that word string. This can be done using Bayes Rule and by selecting the maximum combination of the probabilities represented by the acoustic model and the language model using the following equation:

$$w_1, w_2, \ldots w_n = \underset{w_1, w_2, \ldots w_n}{\operatorname{argmax}} P(W|a) = \underset{w_1, w_2, \ldots w_n}{\operatorname{argmax}} P(a|W)P(W)$$

This can be implemented in multiple ways. More detail can be found in Jurafsky and Martin, 2008.[4]

The best results happen when the models used in ASR are based on data that closely resemble the test data. When there are differences between the test and training data, methods to adapt the models to the test data, or the test data to better match the models can be used to improve recognition results. At initial stages, finding data that best matches different environmental or stylistic components of the test data provides important baselines for the problem. Beyond that, modifying noisy input data to better match model speech is valuable because reducing mismatch is helpful even if both starting points are noisy. Modifying the models requires updating training speech and retraining models, which is resource intensive, so processing the input speech to more closely match the training speech can improve results more easily.

2.2. Speech Recognition Virtual Kitchen (SRVK)

SRVK, a toolkit to support ASR, was created to improve community research and education infrastructure for automatic speech recognition.[5,6,7] It has been developed and evaluated by a team of researchers at Carnegie Mellon University, the Ohio State University and Minnesota State University, Mankato. It includes state of the art Linux-based virtual machines (VMs) with pre-compiled software tools to run various ASR experiments. SRVK provides fully configured VMs with documentation for a range of users, from students to academic researchers in a variety of fields to industry working on natural language processing problems. The SRVK resources can be found online.[8]

2.3. The Eesen ASR System

The Eesen transcriber[1] is an end-to-end speech recognition system on a virtual machine available through the Speech Recognition Virtual Kitchen. Eesen uses an approach based on a deep recurrent neural network acoustic model and a weighted finite state transducer (WFST) decoder to transcribe audio into word sequences. The recurrent neural network (RNN) acoustic models are trained in a single step to reduce the complexity of ASR system development. The WFST-based decoding approach can incorporate lexicons and language models into connectionist temporal classification (CTC) decoding in an effective and efficient way. It has a graphics processing unit (GPU) implementation of long short-term memory (LSTM) model training and CTC learning. Multiple utterances are processed in parallel for faster training. The size of the decoding graph space in Eesen is small, resulting in 263 megabytes and reducing memory needs in computation[1]. Accordingly, the decoding process executes faster. The Eesen character-based system trained results in 7.34% word er ror rate (WER) on the Wall Street Journal (WSJ) corpus, which consists of recordings of Wall Street Journal news articles.

3. Data

In order to perform speech recognition, data is needed to train statistical models and then to test the system. In this section, characteristics of training and testing data are described and problems with our challenging test data are discussed.

3.1 Test Data

There were seven audio recordings of varying lengths used for the test data. They consisted of recorded interviews with SRVK users about their experiences with SRVK and their use in research and education. The audio data was originally collected for a qualitative research experiment so there was no attempt to have uniform recording lengths. The test data was recorded with a hand-held, single-channel recorder (an Olympus DS-50 Digital Voice Recorder). The interviewer was speaking near the recorder providing a clean, direct speech while the interviewee was speaking on a speaker phone providing telephone speech. Since the audio data came from interview recordings, they consisted of spontaneous speech about ASR systems and the SRVK. Table 1 shows the speakers listed by gender (F/M) and native (E) or non-native (N) speaker of American English, recording length, analysis of audio files, and out of vocabulary percentage.

The test set audio files were recorded with a sampling rate of 44.1kHz in WMA format. The ASR system requires other formats, so the audio files were upsampled and converted to mp3 format. For calculating dynamic range of the test data, maximum and minimum amplitude were measured using the frequency analysis toolkit available in the open source software Audacity[9], which plots a graph of amplitude (dB) versus frequency (Hz). Software limitations allow only the first 237.8 seconds of audio files to be analyzed. Analysis options used were a rectangular window ranging from 0 to 22,005 Hz.

With total test data of less than 5 hours, there is not enough to train either acoustic or language models using just this data. This requires using other forms of data to train appropriate models for automatic speech recognition. Model training data should include similar characteristics as this test data and be easily available. However, there are no large corpora that match this test set.

Table 1. Description of Test Data.
Speakers are described as female (F) or male (M) and native speakers of American English (E) or non-native speakers (N).

	Speaker (clean)	Speaker (phone)	Time	Audio Analysis (first 237.8 seconds)			Out of Vocabulary Rates	
				Max Amplitude (dB)	Min Amplitude (dB)	Dynamic Range (dB)	TEDLIUM before / after lexicon update	Switchboard
1	FE	MN	29:13	-32.3	-78.1	45.8	4.02 / 0.52	3.76
2	FE	ME	1:15:55	-31.3	-79.7	48.4	3.95 / 0.41	5.04
3	FE	ME	53:22	-21.7	-71.5	49.8	3.19 / –	3.74
4	FE	FE	30:24	-19.9	-68.9	49.0	2.54 / 0.29	3.67
5	FE	FN	40:04	-18.1	-68.7	50.6	4.65 / –	4.26
6	FE	MN	20:37	-19.1	-71.0	51.9	6.75 / 0.88	4.94
7	FE	ME	21:42	-21.0	-68.8	47.8	5.11 / 0.41	5.93

3.2 Training Data

In this project, two distinct types of acoustic and language models were built from training data to improve recognition of telephone speech while maintaining or improving results for clean speech in mixed telephone-clean speech recordings. They were built from two different types of data: TEDLIUM and Switchboard models.

The TEDLIUM acoustic model was trained with 81 hours of transcribed speech from TED talks, which are generally well prepared, cleanly recorded lectures. TED talks are usually spoken by a single person in a quiet auditorium without any overlapping speech or background noise. They are recorded with a high-quality recorder in calm settings. Examples of transcribed speech are shown in Figure 2. The TEDLIUM language model was trained

Here I've plotted for you the mean household income received by each fifth and top five percent of the population over the last 20 years. In 1993, the differences between the different quintiles of the population, in terms of income, are fairly egregious. It's not difficult to discern that there are differences. But over the last 20 years, that significant difference has become a Grand Canyon of sorts between those at the top and everyone else. In fact, the top 20 percent of our population own close to 90 percent of the total wealth in this country. We're at unprecedented levels of economic inequality. What that means is that wealth is not only becoming increasingly concentrated in the hands of a select group of individuals, but the American dream is becoming increasingly unattainable for an increasing majority of us. And if it's the case, as we've been finding, that the wealthier you are, the more entitled you feel to that wealth, and the more likely you are to prioritize your own interests above the interests of other people, and be willing to do things to serve that self-interest, well, then, there's no reason to think that those patterns will change. In fact, there's every reason to think that they'll only get worse, and that's what it would look like if things just stayed the same, at the same linear rate, over the next 20 years.

Figure 2: TEDLIUM Transcribed Speech Example

using a TED talk based lexicon and dictionary, where the dictionary contains complete words related to the various talk topics.

The Switchboard acoustic model was trained with about 260 hours of transcribed spontaneous conversations held over the phone[3]. Generally, these are dual channel interpersonal conversations between paired strangers with about 70 topics. These conversations were held by 543 speakers from all over the United States, with dialect regions noted. These conversations were not prepared and contain constant spontaneous speech. Since these training data were recorded over the phone, the quality of speech signal is poorer than that of TEDLIUM model and contained limited frequencies and unintended noise. Examples of transcribed speech are shown in Figure 3. The Switchboard language model has a lexicon of words that is used daily for conversation. Unlike the TEDLUM language model, it contains spontaneous words, such as um, um-hum and aha, and partial or incomplete words such as ‘cause and eleva- (instead of elevator). A comparison of the training data features for TEDLIUM and Switchboard is shown in Table 2.

A: Uh, let’s see. How about, uh, let’s see, about ten years ago. Uh, what do you think was different ten years ago from now?
B: Well, I would say as far as social changed go, uh, I think families were more together. They, they did more things together
A: Uh-huh
B: Uh, they ate dinner at the table together. Uh, the parents usually took out time, uh, you know, more time than they do now to come with the children and just spend the day doing a family activity.
…
B: Do you think that the individual has as much time as they did, let's say, ten, twenty years ago?
A: Um. It depends. Uh, it's hard to say because I think people were busy ten twenty years ago too.
B: Uh-huh
A: Uh, I just. Well, how old are you?
B: I’m twenty-eight
A: Twenty-eight

Figure 3: Switchboard Transcribed Speech Example

Table 2: Training Data Features

Training Corpus	Hours of Data	Recording Environment	Sample Rate	Speaking Style
TEDLIUM	81 hours	High-quality recording in a quiet environment without any intruding noise, except occasional applause.	16kHz	A well planned speech spoken by a single person in a conference auditorium to an audience.
Switchboard	~260 hours	Telephone speech recorded on two channels.	8kHz	Interpersonal spontaneous conversations held over the phone between paired strangers.

3.3 Problems with the Test Data

The initial transcription files of the audio recordings were imperfect, resulting in poor initial recognition results when hypotheses were compared to the transcriptions. To improve recognition results, misspelled and mistranscribed words of the reference transcriptions were corrected, time boundaries of utterance and sentences were hand adjusted, out of vocabulary (OOV) words were found and added to the dictionary, and language models were retrained. Since test data were recorded interviews specifically about speech recognition while training data were trained based on either TED talks and spontaneous conversations, there were many OOV words. Before cleaning and after cleaning out of vocabulary rates are in Table 1. The OOV list for the TEDLIUM data is in Table 3, clearly showing the domain mismatch based on key words missing from the TEDLIUM dictionary. All the words with * represent domain specific words and with $^{+}$ represent speech errors or disfluencies.

There were additional problems as well. Due to the spontaneous nature of the discussions, overlapping speech exists. Overlapping speech becomes noise for a given sequence of speech and makes it hard for the system to separate the speakers, especially in a single channel recording. Also, interviewing through the telecommunication line, such as Skype or phone call, with a speaker and microphone close to each other caused echo problems in some files. Since echoes are reflection of sound through the telecommunication line, they were very noisy and hard for the system to recognize. Therefore, echoes were considered noise, rather than the intended speech to be transcribed, and were removed from the reference transcription file and replaced with a noise marker. Reference transcription files were fixed according to the methods described in Section 4.

4. Methodology

Recognition results were obtained by following experimental recipes in the SRVK virtual machines that housed either the TEDLIUM or Switchboard acoustic model using the training data. The initial baseline ASR results were found using both TEDLIUM and Switchboard acoustic models. This highlighted the need for improvements in the process and data cleaning in both time boundaries and vocabulary to improve the results to where they might be useful for transcription and analysis of the interview recordings.

Several issues were at fault. First, automatic segmentation based on speaker changes was not reliable. Utterance boundaries happened when the speaker changed or there was more than a second of silence. Having a brief silence before and after an utterance helped the system match the reference file better so those were added. It is also important to have word labels associated with the correct time point in the audio because ASR output transcripts were compared with these human extracted labels based on time marks to calculate the final word error rate (WER).

To improve reference transcription files of audio recordings, multiple steps were required. After a first pass correcting some time boundaries and word labeling using the free software, Audacity[9], word label files were extracted from the software and used to find OOV words. OOV words and OOV percentages were found using error analysis software provided on the SRVK website by comparing words from the reference transcript and the lexicons

Table 3. Out of Vocabulary word lists for the TEDLIUM lexicon.

Word	Freq	Word	Freq	Word	Freq	Word	Freq
mhmm	795	GMMs*	2	FST*	1	recurnal+	1
okay	447	GPUs*	2	GMM*	1	resynthesis+	1
mmkay	33	hmmm	2	GPU*	1	resynthesizing*	1
cause	20	IRB*	2	handicampt+	1	re-understand+	1
recognizer*	16	pre-read*	2	Hasegawa-Johnson*	1	selve+	1
Kaldi*	14	Riebling	2	HCI*	1	seping+	1
Praat*	13	self-directed*	2	HCLG*	1	shh+	1
festvox*	9	self-paced*	2	imple-+	1	somethi-+	1
ASR*	7	semi-structured*	2	interes-+	1	spectrogram*	1
API*	6	SSH*	2	Interspeech*	1	spectrograms*	1
MATLAB*	6	ACL*	1	k-+	1	speechy+	1
NLTK*	6	almo-+	1	kitch-+	1	SRILM	1
HTK*	5	ASEE*	1	kuh	1	tetnian+	1
TEDLIUM*	5	assonance+	1	LTI*	1	there're	1
APIs*	4	builts+	1	MATLAB's*	1	thi-+	1
buh-bye	4	coarticulation*	1	mmhmmmm	1	tic-tac-toe	1
homeworks+	4	compo-+	1	Nasalized*	1	transf-+	1
laplace*	3	concatenous*	1	NDT*	1	transcriptions*	1
MFCC*	3	CPU*	1	netbooks	1	tuh	1
MFCCs*	3	crosslisted	1	NLP*	1	uhmm	1
pre-VM*	3	decompositioning*	1	otherwi-+	1	USTC*	1
righ-+	3	deprecated	1	parser	1	versioning	1
VMs*	3	discriminative*	1	phonings+	1	vo-coder*	1
Convolution*	2	dynam-+	1	prepper	1	XML*	1
diarization+	2	equivalently	1	pre-read	1	yadir+	1
discriminant*	2	exerc-+	1	pre-use	1		
Florian's	2	FLYTEC*	1	prosodic	1		

* domain specific words + incomplete words, speech errors or disfluencies

for different systems (TEDLIUM or Switchboard). These words were then used to check for any misspelled or mistranscribed words in the reference transcript. After fixing misspellings and mistranscribed words in the reference file, regions with high error rates in the transcript were analyzed. In these regions, there were frequently reference file segmentation errors, where utterance or sentence boundaries were automatically assigned. Based on these regions, time segmentation boundary errors were hand corrected using Audacity. Noisy areas were also checked to make sure that appropriate noise labels, e.g., [noise], were included in the transcriptions using the labeling function in Audacity. After correcting all the transcription errors, the OOV software was re-run to get the maximum possible hypothetical ASR result for both the current TEDLIUM based- and Switchboard based-VMs, with differently trained acoustic models, language models, and dictionaries. After this, OOV words were added to the dictionary and language models were retrained. Finally, ASR was re-run with the new dictionary and language models.

5. Results

ASR results are described in word error rate (WER). To compute WER, the total number of reference words were divided by the sum of numbers of deleted (D), inserted (I), and substituted (S) words as shown below.

$$WER = \frac{number\ of\ (S + I + D)}{Total\ number\ of Refrence\ Words}$$

Table 4: Order of Experiments

Step	Focus Area	Action
1	Baseline	Using TEDLIUM and Switchboard acoustic model and unmodified reference transcripts, get initial baseline ASR result.
2	Reference Word Sequence	Modify time boundaries and word labeling of the trancript, and provide utterance a brief silence before and after a utterance boundary happens.
3	Dictionary	Extract OOV words, calculate OOV percentage, and correct misspelled or mistranscribed words in the reference transcript.
4	Analyze Transcript	Find, analyze, and correct high error rate regions in the transcript. If noisy area is detected, transcribe with appropriate noise labels.
5	Transcript Correction	Fix all transcription errors and get maximum possible hypothetical ASR result and OOV words.
6	Language Model	Add OOV words to the dictionary and language models
7	Final Value	Re-run ASR with new dictionary and language models

Examples of how deleted, inserted, and substituted words were labled are shown in Figure 4. Deletions are when no words are recognized by the system but they exist in the reference transcription. Substitutions are when a different word is recognized by the system. In Figure 4, the system recognized the word "ARE" rather than "WERE" and "AND" rather than "OR." Insertions are words added by the system during recognition even though a speaker did not say them ("WORK").

As shown in Table 5, results ranged from 39.4 – 86.6% WER using the TEDLIUM model and from 39.4 – 92.5% WER with the Switchboard model before adding new vocabulary words. In the table, boldfaced numbers highlight improvement when using the Switchboard model.

Both models generally did better recognition on clean speech than telephone speech, better on female speakers than male speakers, and better on native speakers of American English than non-native English speakers. Even though the Switchboard acoustic model is a better match for spontaneous conversation, noisy speech, and the recording environment, it also had better recognition on clean speech than telephone speech, and did not necessarily show improvement over TEDLIUM models on telephone speech. The system generally did best on female native speakers of American English. Since both TEDLIUM and Switchboard models were primarily trained with data recorded by native American English speakers, they have a known weakness in recognizing speakers whose native language is not American English.

REF:	****	OR	that	if	there's	similar	language	and	you	can	bootstrap
HYP:	WORK	AND	that	if	there's	similar	language	and	you	can	bootstrap
Eval:	I	S									

REF:	WE	WERE	using	THAT	for	our	language	model	training
HYP:	**	ARE	using	THEM	for	our	language	model	training
Eval:	D	S		S					

Figure 4: Word Error Rate Computation Examples.

Table 5. Recognition Results: Word Error Rates and percentage change across acoustic models and updated lexicon. Boldfaced values show improvement of Switchboard model over TEDLIUM model.

	Speaker	TEDLIUM	SWBD	% Change	TEDLIUM + New Words	% Change
1 – clean	FE	39.4	39.4	0	27.0	31.5
phone	MN	77.2	92.5	-19.8	64.0	17.1
2 – clean	FE	51.5	58.2	-13.0	40.2	21.9
phone	ME	62.3	69.4	-11.4	48.3	22.5
3 – clean	FE	44.2	**41.0**	3.2	-	-
phone	ME	63.2	**61.9**	1.3	-	-
4 – clean	FE	64.2	**53.1**	17.3	48.3	24.8
phone	FE	68.3	**51.8**	24.1	56.0	18.0
5 – clean	FE	42.1	**39.6**	2.5	-	-
phone	FN	86.6	88.6	-2.0	-	-
6 – clean	FE	56.2	56.6	-0.7	42.7	24.0
phone	MN	68.2	73.0	-7.0	57.1	16.3
7 – clean	FE	45.5	52.6	-15.6	29.9	32.3
phone	ME	60.3	**56.9**	5.6	40.3	33.2

After adding in new vocabulary words, the TEDLIUM results ranged from 27-64% WER, with improvements of up to 33.2% per speaker. Adding all out of vocabulary words lists listed in Table 3 to the TEDLIUM dictionary and training a new language model to include these words definitely improved the recognition results. Since the original TEDLIUM language model was trained using TED talks, it lacks representation of spontaneous flow of words as well as partial or incomplete words. Adding these words and words that are related to the conversational domain of speech recognition and the SRVK toolkit improved recognition results. Because of resource limitations, final results for conversations 3 and 5 were not available at time of publication.

6. Conclusion

Although a new acoustic model, the Switchboard model based on spontaneous, conversational speech, better matches the speaking style and recording environment, it only gave limited improvements. Results were mixed for the matched telephone environment. Frequently, it performed better than the TEDLIUM model for native speakers of American English. However, the TEDLIUM model with trained new words had dramatic improvement of up to 33.2%. However, until the next step of updating the Switchboard language model and lexicon to reduce out of vocabulary words, we cannot yet say that TEDLIUM provides a better model than the Switchboard model. Replication of the experiments described here is not directly possible because of IRB limitations on the use of the test data. However, because virtual machines were used, all experiments could be re-run exactly with test data of a user's choice.

7. Future Work

As introduced in Figure 1 and discussed in Section 2.1, approaches to improve the results either address the signal processing stage or the search stage of the system. In signal processing, adaptation to reduce noise, both additive and convolutional, in the test data to better match the acoustic model training data is one approach. In the search stage, improving the acoustic and language models as well as the dictionary could improve results, as could combining different search processes in post-processing. Although the TEDLIUM model with OOV words inserted outperformed the Switchboard model with no vocabulary updates, creating acoustic and language models that incorporate both types of training data would be the next step to improve speech recognition beyond the approach tried here. Another form for this option would be to combine two models into a hybrid model where the best results are chosen based on factors like the speaker, the conversation partner, and probability or confidence values for the hypothesized word strings. Having non-native speakers in the test set could require finding other training data or better adapting the dictionary pronunciations and acoustic model to match different accents. The language model could be improved by adding more words to the dictionary and training it with data better matches the test set. Once these tasks are achieved, exploring ways to adapt the incoming acoustic feature vectors to better match either of the available models would also be future work. An example of this is applying a variety of filters to the speech signal which could reduce both additive and convolutional noise resulting from the telephone channel. The very first step, though, would be training OOV words into the Switchboard language model to better check the performance of Switchboard model.

8. Acknowledgements

We gratefully acknowledge the work of Florian Metze and Erik Riebling at Carnegie Mellon University, Eric Fosler-Lussier and Andrew Plummer at the Ohio State University, and Zheijan Wang at Minnesota State University, Mankato.

This material is based on work supported by the National Science Foundation under grant nos. CNS-1305365, CNS-1305319, and CNS-1305215

9. References

[1] Y. Miao, M. Gowayyed, & F. Metze, “Eesen: End-To-End Speech Recognition Using Deep RNN Models and WFST-Based Decoding”, IEEE Workshop on Automatic Speech Recognition and Understanding, Oct. 2015.

[2] A. Rousseau, P. Deleglise, & Y. Esteve, “TED-LIUM: an Automatic Speech Recognition dedicated corpus”, Proc. LREF. Istanbul, May. 2012

[3] J. Godfrey & E. Holliman. Switchboard-1 Release 2 LDC97S62. Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[4] D. Jurafsky & J. H. Martin, Speech and Language Processing (2nd Edition), Prentice-Hall, Inc., Upper Saddle River, NJ, 2008.

[5] F. Metze, E. Riebling, E. Fosler-Lussier, A. Plummer, & R. Bates. "The Speech Recognition Virtual Kitchen Turns One," Proc. INTERSPEECH, Dresden, Germany, Sept. 2015.

[6] F. Metze, E. Fosler-Lussier, & R. Bates, “The Speech Recognition Virtual Kitchen,” Proc. INTERSPEECH, Lyon, France, Aug. 2013.

[7] A. Plummer, E. Riebling, A. Kumar, F. Metze, E. Fosler-Lussier, & R. Bates, “The Speech Recognition Virtual Kitchen: Launch Party,” Proc. INTERSPEECH. Singapore, Sept. 2014.

[8] The Speech Recognition Virtual Kitchen. Web site: http://www.speechkitchen.org. Accessed August 1, 2017.

[9] Audacity(R) software is copyright (c) 1999-2017 Audacity Team. Web site: http://audacityteam.org/. It is free software distributed under the terms of the GNU General Public License. Accessed August 1, 2017.