9-2016

Effect of Speech Recognition Errors on Text Understandability for People who are Deaf or Hard of Hearing

Sushant Kafle

Matt Huenerfauth

Effect of Speech Recognition Errors on Text Understandability for People who are Deaf or Hard of Hearing

Sushant Kafle[1], *Matt Huenerfauth*[1]

[1]Rochester Institute of Technology (RIT)
Golisano College of Computing and Information Sciences
20 Lomb Memorial Drive, Rochester, NY 14623
sxk5664@rit.edu, matt.huenerfauth@rit.edu

Abstract

Recent advancements in the accuracy of Automated Speech Recognition (ASR) technologies have made them a potential candidate for the task of captioning. However, the presence of errors in the output may present challenges in their use in a fully automatic system. In this research, we are looking more closely into the impact of different inaccurate transcriptions from the ASR system on the understandability of captions for Deaf or Hard-of-Hearing (DHH) individuals. Through a user study with 30 DHH users, we studied the effect of the presence of an error in a text on its understandability for DHH users. We also investigated different prediction models to capture this relation accurately. Among other models, our random forest based model provided the best mean accuracy of 62.04% on the task. Further, we plan to improve this model with more data and use it to advance our investigation on ASR technologies to improve ASR based captioning for DHH users.

Index Terms: Accessibility for People who are Deaf or Hard-of-Hearing; Captioning System; Speech Recognition; Human Computer Interaction; Computer Linguistics

1. Introduction

Captions provide a way to represent aural information in visual text for people who are Deaf or Hard-of-Hearing (DHH). Today there are more than 360 million people worldwide with hearing loss [1] and they use services such as captioning to get access to information existing in the form of speech such as information from mainstream classes, meetings, and live events. Several methods have been explored in providing such a service; a popular alternative includes the use of captionist to transcribe audio information to text using a keyboard, with the captions displayed on a screen for those in attendance. Captioning services produce a digital textual output which can be processed and represented in various forms easily, or it can be stored as a transcript, making it useful in various scenarios such as classrooms and meetings, where it could be reviewed later.

Over the past few decades, automated speech recognition (ASR) technologies have seen major progress in their accuracy and speed. With its increasing maturity, ASR technologies are now being used commercially for many consumer applications. Due to their cheap and scalable ability (compared to other captioning alternatives) to generate real-time text from live audio or recordings, ASR systems have a potential for the task of captioning. Researchers have begun to investigate the suitability of ASR to automate or semi-automate the process of captioning with the use of ASR systems [2, 3, 4, 5] in various application settings.

Despite the growing use of ASR systems, accurate, large-vocabulary, continuous speech recognition is still considered an unsolved problem; the performance of ASR system is not on par with humans [6], who currently provide most caption text for DHH users. Due to unpredictable ambiguity in human speech and ever existing noise, ASR systems often make errors, and it is likely that this technology will continue to be imperfect in the near future as well. Researchers have also argued that ASR generated errors on captions are more comprehension-demanding than human produced errors [7, 8]. While all users of ASR technology must cope with errors in the output, there is potential that this issue has greater significance when focusing on applications for DHH users. Past research has indicated that the majority of deaf high school graduates in the U.S. have an English literacy level at the fourth grade or below [9], and approximately 20% leave school with a reading level at or below second-grade [10]. This presents a huge challenge for caption acceptance by DHH individuals given the error-probable output from ASR.

For a successful use of an ASR system in captioning, errors that affect comprehension of a caption for DHH users might need to be appropriately reduced or at least sufficiently modulated. It may be the case that some classes of errors from an ASR system are especially problematic for DHH users (perhaps based on their unique English literacy profile), and other classes of errors are less problematic. Understanding this trade-off could make way for designing an adaptive ASR system optimized for the task of captioning, specifically for DHH users.

In this paper, we present a method to study the effect of different ASR-generated errors on the understandability of a text for DHH users. For our task, we formulate a user study with DHH users who are given imperfect English texts (containing ASR errors) and asked to answer some questions based on the information from the text. With the data collected from the user study, we model the relationship between ASR errors and the impact it has on the understandability of a text for DHH users. We also discuss the possible application of this model in designing a custom loss function that could be utilized during the decision making process of the ASR to produce better outputs for captioning for DHH users.

2. Background: N-best list Rescoring Technique

In an ASR system, the function of the decoder is to find the most likely word sequence given the sequence of audio features. Although decoders are designed primarily to find a single solution, in practice, it is relatively simple to generate not just the most

	Transcription	WER loss	Avg. Understandability loss
Reference	The meeting today has been cancelled and is scheduled for next Thursday.	NA	NA
ASR Hypothesis 1	The *meet in* today has been cancelled *an* is scheduled for next Thursday.	25%	8.425%
ASR Hypothesis 2	The meeting today has been *capital* and is *skidoo* for next Thursday.	16.67%	46.425%

Table 1: Example shows how Understandability loss penalizes texts containing different errors as compared to WER loss. Higher loss value indicates worse output for the metric.

likely hypothesis but the n-best set of hypotheses. Therefore, in most ASR systems, along with the most likely word sequence, a list of n-best hypotheses can also be obtained as output. Other compact forms of representation of this n-best hypotheses list are also commonly used such as a word lattice representation [11] or a confusion network [12].

These representations have been popular especially because they provide a reduced search-space (out of all possible word sequence) that can be further decoded, with more flexibility, to improve the ASR output. This post processing technique of "rescoring" or "reranking" candidate hypotheses also allows for general-purpose hypothesis to be tuned in a domain-specific or user specific way without having to design the whole ASR engine to do so [13]. Furthermore, the n-best hypotheses generated as an output from the ASR system can be processed with complete independence from the ASR system; thus, it can be treated as a separate stage in an ASR pipeline.

Researchers [14, 15, 16, 17, 18, 19] have utilized various rescoring techniques to select the best hypothesis from an ASR n-best hypotheses. In [19], Stockle et al. presented an N-best list rescoring algorithm to improve upon the shortcomings of the ASR decoding process to produce more accurate output. A standard Hidden Markov Model (HMM) based ASR system uses Maximum A Posteriori (MAP) technique as a decoding criterion. The problem with the application of the MAP approach to speech recognition is that it is sub-optimal with respect to minimizing the number of word errors in the system output. Instead, it has been shown to minimize sentence error rate which is only loosely linked to the recognition Word Error Rate (WER) [19]. Subsequently, Stolcke et al. [19] proposed a rescoring algorithm that explicitly minimizes expected word error for recognition hypotheses. In [20] researchers provided a Decision Theoretic perspective to the work from [19] as a Bayes decision rule under word error loss, as shown in Equation (1).

$$\delta(X) = \underset{W \in \mathcal{W}}{\operatorname{argmin}} \sum_{W' \in \mathcal{W}} WER(W, \delta(X)) P(W'|X). \quad (1)$$

Goel et al. [20] proposed a modified loss function (as shown in Equation (2)) to be minimized during the modified decoding process by adding additional degree of freedom which can be "tuned" appropriately during training. Additionally, [20] also make simplifying assumptions to compute $P(W|X)$ with joint distribution $P(X, W)$ which are accessible from the n-best lists.

$$l(W, \delta(X)) = [WER(W, \delta(X))]^x. \quad (2)$$

This framework suggested by [20] provides a flexible way to incorporate a custom loss function in the decoding process of ASR, and this approach is how we intend to adapt ASR in our work. This Minimum Bayes Risk (MBR) based decoding has been shown to provide statistically significant improvements in recognition task compared to MAP based decoding as it explicitly incorporates task performance criterion to the decoding process of ASR. Successes of hypotheses scoring systems like ROVER [14] (and its variants) has been credited to MBR based decoding to directly improve WER. Several research groups have investigated this method of decoding in recent years [21, 22, 23].

3. Design & Implementation

The approaches discussed above utilize an n-best list rescoring technique to improve the WER of an ASR system. We propose to compare the efficacy of these rescoring approaches for optimizing ASR for real-time captioning, a task for which there may be better metrics than WER. We propose to learn a custom loss function (based on the analysis of data from experiments with DHH users) to optimize the comprehensibility of ASR output for DHH users. Unlike WER, our loss function may provide a better measure of text understandability for this group of users. Table (1) shows a comparative example of our loss function (based on the data and modeling presented later in Section 3.2 of the paper) against the traditional WER loss. In the example, we can see how our Understandability model prefers Hypothesis 1 over Hypothesis 2 as compared to WER metric which does the opposite.

This paper, in general, is about creating this loss function using a prediction model which captures the relation between different types of error and their impact on the understandability of sentence for DHH users. As a final step (in the future), we will be looking to see if this loss function can be incorporated into the decision-making process of an ASR system, following the framework provided by [20], such that the ASR can produce output that is optimized to be more comprehensive for our user group.

3.1. User Study

We performed a user study with a goal of understanding how ASR errors affect DHH users' performance on a comprehension task, given that a text contains some ASR generated errors. In this study, users were presented with imperfect English text passages (containing artificially inserted errors, based on real ASR errors for that passage) and were asked to answer questions that required understanding the information content of those passages. Based on the answers, we collected Comprehension Scores for the respective questions, which we subsequently used to model the relationship between errors in the text and its comprehensibility.

3.1.1. Error Categories used in Designing Stimuli

To guide our creation of stimuli for the user study, we established a hierarchical classification of various sub-types of ASR errors (based on a time-alignment between the ASR output and the gold-standard). Broadly, ASR errors can be categorized into three types: substitution, deletion and insertion errors. Further, we divided substitution errors into four types: one to one substitution, one to many substitution, many to one substitution and many to many substitution. One to one substitution refers to the errors when one word is substituted by the other. One to many substitution errors are the error due to substitution of one word by many (for e.g., *undistinguished* substituted by *on distinguished*). Similarly, many to one errors are the errors when many words are substituted by a single word. Many to many errors corresponds to a multi-word span of text in the reference transcript with inaccurate recognition such that none of the word boundaries within the span align with those within the corresponding span of ASR output. We further subcategorized one to one substitution errors into three types namely, morphologically similar substitution, phonetically similar substitution and remaining other types of substitution errors. The morphologically similar errors are the errors where the actual word is substituted by another word with an inflectional or derivational morphological relationship to the first (for e.g., *developed* substituted by *develop*). The phonetically similar errors are the errors due to the substitution of a word by another word with similar phoneme representation; for example, the words *table (T EY B AH L)* and *stable (S T EY B AH L)* have a very close ($\geq$ 60% match) phoneme structure so they are considered as a phone neighbor of each other.

These categories of different error types were meant to be a coarse categorization of the errors and was used as a basis for ensuring that the stimuli presented in our user study contained a good mixture of different error types.

3.1.2. Study Resources

For the user study, we created a dataset of 20 passages (average length 117 words), with each passage containing three sentences marked as our Region Of Interest (ROI). For example, the text below shows a sample text passage used in the study with three bold sentences representing the three ROIs in the text.

People who study film music often complain about the lack of recognition their field receives. The study of film music is an interdisciplinary field, falling in between cinema studies and musicology. *This is one of the reasons why it receives so little attention. For example, when film music scholars, who often do not have music-degree credentials on par with the pure musicologists, write about film soundtracks, their articles are often ignored by the musicologists.* ***Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often treat it as the work of amateurs.*** *So with the members of the two fields most closely related to it ignoring it, it is easy to understand why members of the film music field feel a degree of frustration.*

The questions for passages was designed in such a way that each question was based on information from only one of the ROI sentences in that passage. In total, each passage had three text-explicit questions. As described in [9], text-explicit questions measure exact recall from the text without requiring any inferential use of information from the reader's memory.

The text below shows an example of a question asked during the user study. The question is based on the reading text shown above as an example. This question, in particular, is based on information from the first ROI sentence of the reading text.

A. According to the passage, what do film music students often complain about:

☐ *that their field doesn't receive the recognition they deserve.*
☐ *people who study film music are not recognized.*
☐ *film music study is not up-to the par.*
☐ *extra attention that their field receives.*

For each ROI sentence, an average of 8 different variations were generated where each variation was produced by inserting at most one category of ASR error into the ROI sentence. To produce each variation of the ROI, we began with a perfect text and inserted one of those errors. The text below shows an example of an ROI sentence without any errors:

Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often treat it as the work of amateurs.

We produced different variations of this ROI text by adding ASR generated errors into the sentence. ASR generated errors were collected by creating an audio recording of a male English speaker performing each ROI sentence (multiple times) and running it against the ASR system. Since our goal was to obtain output containing a variety of errors, we used the CMU Sphinx system with its distributed trained models [24]. Some variations of the ROI text are shown below:

- *Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often* ***cricket*** *as the work of amateurs.*
- *Conversely, when the* ***working*** *of film music scholars touches on the visual aspects of film, the cinema studies people often treat it as the work of amateurs.*
- *Conversely, when the work of film music scholars touches on the* ***region*** *aspects of film, the cinema studies people often treat it as the work of amateurs.*
- *Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often treat it* ***has worked*** *amateurs.*

This procedure ensured that the artificially created variations of the ROI sentence agreed with the actual imperfect output produced by an ASR system.

3.1.3. Participants

Participants for the study were recruited from among associate degree students at the National Technical Institute for the Deaf (NTID) at Rochester Institute of Technology (RIT). We collected data from 30 DHH participants (age distribution with μ=22.63 and σ=2.63), 12 men and 18 women, where 26 participants self-identified as Deaf and 4 of participants as Hard-of-Hearing.

3.1.4. Procedure

Each participant was given 10 different comprehension passages to read, each containing three multiple choice questions that needed to be answered in a time period of 70 minutes. A pilot test with a DHH member of our research team helped us to determine an appropriate number of question items for the 70-minute experiment. The comprehension passages given to the participants were generated by replacing each ROI sentence by its erroneous counterpart (one of the variations). The number of errors of each category that were displayed to each participant was balanced among all participants in the study to ensure that individual human differences in task performance did not disproportionately affect the scores for any one category of error. Further, each ROI appeared several times throughout the entire study in a form without any errors inserted so that we could obtain baseline measurements for the difficulty of the particular comprehension question, to enable subsequent normalization of the collected scores. Scores of answers from each question were binary with correct answer receiving the Comprehension Score of 1 and incorrect answer receiving the score of 0.

3.2. Model Fitting

The data collected from the user study enabled us to determine whether there is a relation between the presence of an error with specific linguistic characteristics (see Table (2)) in a sentence and its impact on the comprehension of the sentence (whether or not participants answered the question referring to the sentence). However, the relation between the presence of an error and its impact on sentence is not straightforward. A wide variety of complex semantic factors can lead some ASR errors to be more confusing than others for end-users who are reading the text. For our automatic captioning application, we are interested in focusing on a subset of those aspects of a text that could be automatically computed, using modern computational linguistic software.

Table 2: List of features extracted from the error regions in the hypothesized text for analysis.

	Feature	Description	Type
1.	*WordLength*	Average length of the word in the region.	Numeric
2.	*SaliencyIndex*	Average TF-IDF score of the word in the region representing the importance of the word.	Numeric
3.	*POSTag*	Priority order based Part of Speech tag assigned to the region. The order is described in Section(3.2.2).	Categorical
4.	SyllableLength	Average number of syllables of the word in the region.	Numeric
5.	SentimentOrientation	Indicates whether the region alters the original sentiment (broadly, positive or negative) of the reference word(s) or not.	Categorical
6.	ContentOrFunction	Whether the region contains content word or not.	Categorical

3.2.1. Feature Identification

After consulting prior research on reading skills of deaf users [10, 25], we identified a list of 6 features of each error that we would examine as part of our analysis. The features are summarized in Table 2. Some features (for e.g. row 5 in Table (2)) are computationally more expensive than others. Since this model will eventually be used to produce a loss function to optimize a real-time ASR system, using these computationally expensive features may not be efficient. But, we considered these features in our preliminary analysis to understand their significance in the model.

3.2.2. Feature Extraction

Along with the Comprehension Scores for the text in the passages used in the study, we also extracted some linguistic features, summarized in Table (2). These features were obtained from the imperfect ROI texts in the passage which the users referred to when answering the questions provided during the study.

Each variation of ROI text contained at most one type of error which was created by replacing the actual (reference) word(s) from the error-free ROI text with a different (hypothesized) word(s). Thus, the first step of the feature extraction process involved alignment of error-free ROI text with its erroneous variation to identify the reference word(s) and the hypothesized word(s) pair. As the ROI texts were not time-aligned and there were few errors in each ROI text, we could utilize Levenshtein distance based word alignment technique to align the texts. We utilized CELEX2 [26] as our lexical database for syllable information for calculating the SyllableLength feature. A frequency-based Part-of-Speech (POS) tagger, Unigram Tagger [27], was utilized for POS tagging of words. The tagger was modified to output one of 11 different POS tags (in priority order: noun, verb, pronoun, adverb, adjective, preposition, conjunction, interjection, determiner, number and others) to an input word. The ContentOrFunction feature was calculated with the help of POS tag(s) of the word (a word is labeled as a Content word if it is a Noun, Verb, Adverb, or Adjective). The SaliencyIndex feature represented the general importance of the word and was estimated by calculating Term Frequency-Inverse Document Frequency (TF-IDF) score of a word(s). Scikit-learn's [28] TfidfVectorizer was used as our TF-IDF Scorer, and it was trained with a portion of dataset (N=18 books) from Project Gutenberg [29] corpus and Web Text corpus from NLTK [27]. TextBlob [30] library for python was used to compute the SentimentOrientation feature.

For each type of error, the features were extracted from the reference word (the actual word), except for the insertion error type (an insertion error doesn't have a reference word as it is produced due to an insertion of an extra word) whose features were extracted from hypothesized word(s).

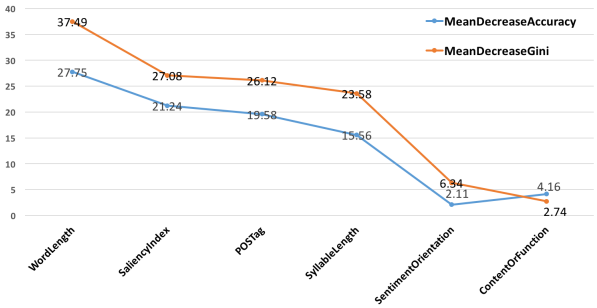

Figure 1: A plot showing the importance of each feature variable in-terms of their contribution to model accuracy and impurity.

3.2.3. Feature Selection

We utilized random forest to rank our 6 features and selected 3 features based on the measure of average accuracy decrease and average impurity decrease in the model without each of these features. As shown in Figure (1), features *WordLength*, *SaliencyIndex* and *POSTag* were among the best contributors to the Gini impurity and the accuracy of the model.

	Evaluation Metrics						
Models	**AUC**	**Cutoff**	**Accuracy**	**F-measure**	**Precision**	**Recall**	**Bal. Accuracy**
Logit (M_l)	0.496	0.539	0.618	0.754	0.618	0.968	0.498
Random Forest (M_{rf})	**0.572**	0.444	**0.620**	0.744	0.631	0.844	**0.533**
SVM (M_s)	0.496	0.605	0.617	0.738	0.625	0.919	0.497

Table 3: Summary of the evaluation of each prediction model on our test dataset. Value on each metric represents the average performance of the model in 5 different train and test partitions of our dataset.

3.2.4. Model Evaluation & Selection

We investigated three models for prediction and evaluated the performance of each model for our task. Table (3) summaries the result our evaluation. For the purpose, we selected 80% of our total observation (N= 862, excluding the baseline measurements) to train the model and used 20% of our remaining observation of test the model. For each model, five-fold cross validation with this 80/20 split was used to build each model, and the performance scores reported in Table (3) are based on the average of the models for each fold. We observed the performance of Random Forest model (M_{rf}) to be slightly better than other models with accuracy of ($\mu = 62.04\%, \sigma = 4.41$).

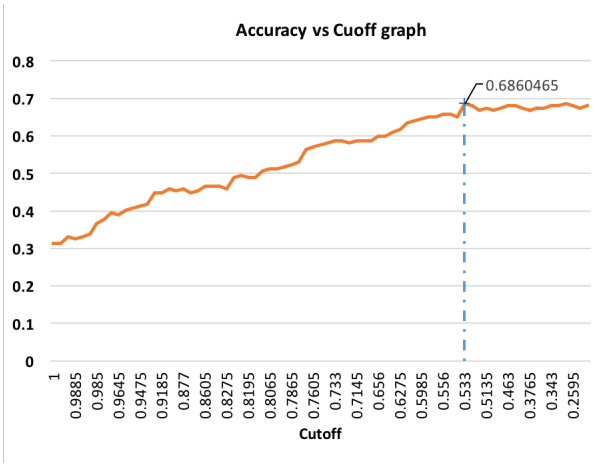

Figure 2: Example of Accuracy vs Cutoff graph for Random Forest Model on a test dataset. The marker represented by the *red-cross* represents the point of maximum accuracy at a cutoff value of 0.31.

During the testing process, the cutoff probability for each classification model, which was used to label output probability to our binary class, was chosen as the mode of the accuracy vs cutoff graph; the graph represented the accuracy of the model considering different cutoff values. Figure (2) shows the accuracy vs cutoff curve of Random Forest model on a test dataset.

4. Discussion

While its performance is above chance, the Random Forest model Accuracy results presented above are modest, but we view these results as preliminary. This study was based on a small amount of data (30 participants on 20 passages), and the set of features explored was relatively small. We view this effort as an initial proof-of-concept of our ability to identify useful features in a loss-function for predicting the comprehensibility of a text for DHH users.

Obviously, if we are to make use of this loss function in real-time captioning system, we would not know which words are errors. Our intention is to use the confidence value of the ASR system as a proxy for this information, and to use our loss-function to guide the hypothesis selection. Specifically, the prediction model (M_{rf}) we built from the user study results will be used in designing our loss function, as shown in Equation (3).

$$\ell(W, \delta(Y)) = -\Big(\sum_{f_i = f(W, \delta(Y))} M_{rf}(f_i) \Big) \tag{3}$$

where $\delta(Y)$ represents our decision rule that maps audio input (Y) to word sequence output ($\hat{W}$). We need a function $f(R, H)$ that returns set of features (listed is Table (2)) for each error type in the hypothesis text (H) when compared to the reference text (R).

This loss function looks to penalize the harsh errors that have significant 'predicted' impact on output comprehension (obtained from M_{rf}) for DHH users.

5. Conclusion & Future Work

The work described in the paper has been concerned with the development of a prediction model that represents the impact of ASR errors present in the text on its comprehension, specifically for DHH users. Beyond our intended application for ASR, we note that research on understanding the relationship between text characteristics and comprehensibility for DHH users may have other applications, such as automatic text readability detection software for these users. Further, we plan to extend our user study and improve our prediction model with more data. As we move on, we will look to investigate the Decision Theoretic framework for n-best list rescoring proposed by [20] to incorporate our custom loss function in to the ASR decoding process.

In addition, we will look to contrast its performance with other discriminative training techniques to optimize ASR components with our loss function. We also intend to do experimental analyses of the effectiveness of the final tool for DHH users.

6. Acknowledgements

This material was based on work supported by the National Technical Institute for the Deaf (NTID). We are grateful to Kellie Menzies, who assisted with data collection for this study, and to our collaborators Larwan Berke, Christopher Caulfield, Micheal Stinson, Lisa Elliot, Donna Easton, and James Mallory.

7. References

[1] W. H. Organization. (2015, March) Deafness and hearing loss. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs300/en/

[2] M. Wald, "Crowdsourcing correction of speech recognition captioning errors," in *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*, ser. W4A '11. New York, NY, USA: ACM, 2011, pp. 22:1–22:2. [Online]. Available: http://doi.acm.org/10.1145/1969289.1969318

[3] I. R. Forman, B. Fletcher, J. Hartley, B. Rippon, and A. Wilson, "Blue herd: Automated captioning for videoconferences," in *In Proc. ASSETS '12*. New York, NY, USA: ACM, 2012, pp. 227–228. [Online]. Available: http://doi.acm.org/10.1145/2384916.2384966

[4] M. Federico and M. Furini, "Enhancing learning accessibility through fully automatic captioning," in *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*, ser. W4A '12. New York, NY, USA: ACM, 2012, pp. 40:1–40:4. [Online]. Available: http://doi.acm.org/10.1145/2207016.2207053

[5] H. Takagi, T. Itoh, and K. Shinkawa, "Evaluation of real-time captioning by machine recognition with human support," in *In Proc. W4A'15*. New York, NY, USA: ACM, 2015, pp. 5:1–5:4. [Online]. Available: http://doi.acm.org/10.1145/2745555.2746648

[6] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Trans Audio Speech*, vol. 21, no. 5, pp. 1060–1089, 2013.

[7] R. S. Kushalnagar, W. S. Lasecki, and J. P. Bigham, "Accessibility evaluation of classroom captions," *ACM Trans. Access. Comput.*, vol. 5, no. 3, pp. 7:1–7:24, Jan. 2014. [Online]. Available: http://doi.acm.org/10.1145/2543578

[8] K. Bain, S. H. Basson, and M. Wald, "Speech recognition in university classrooms: Liberated learning project," in *In Proc. Assets '02*. New York, NY, USA: ACM, 2002, pp. 192–196. [Online]. Available: http://doi.acm.org/10.1145/638249.638284

[9] D. W. Jackson, P. V. Paul, and J. C. Smith, "Prior knowledge and reading comprehension ability of deaf adolescents," *Journal of Deaf Studies and Deaf Education*, pp. 172–184, 1997.

[10] J. L. Luckner and C. M. Handley, "A summary of the reading comprehension research undertaken with students who are deaf or hard of hearing," *American Annals of the Deaf*, vol. 153, no. 1, pp. 6–36, 2008.

[11] F. Richardson, M. Ostendorf, and J. R. Rohlicek, "Lattice-based search strategies for large vocabulary speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1995, pp. 576–579.

[12] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: lattice-based word error minimization." in *Eurospeech*, 1999.

[13] E. K. Ringger and J. F. Allen, "Error correction via a post-processor for continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1, May 1996, pp. 427–430 vol. 1.

[14] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proceedings of the Automatic Speech Recognition and Understanding*, Dec 1997, pp. 347–354.

[15] B. Roark, M. Saraclar, and M. Collins, "Discriminative n-gram language modeling," *Computer Speech & Language*, vol. 21, no. 2, pp. 373–392, 2007.

[16] M. S. Brian Roark and M. Collins, "Corrective language modeling for large vocabulary asr with the perceptron algorithm." in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–749.

[17] T. Oba, T. Hori, and A. Nakamura, "A comparative study on methods of weighted language model training for reranking lvcsr n-best hypotheses," in *In ICASSP'10*. IEEE, 2010, pp. 5126–5129.

[18] J. D. Williams and S. Balakrishnan, "Estimating probability of correctness for asr n-best lists," in *In Proc. SIGDIAL'09*. Association for Computational Linguistics, 2009, pp. 132–135.

[19] A. Stolcke, Y. Konig, and M. Weintraub, "Explicit word error minimization in n-best list rescoring." in *Eurospeech*, vol. 97. Citeseer, 1997, pp. 163–166.

[20] V. Goel, W. Byrne, and S. Khudanpur, "Lvcsr rescoring with modified loss functions: A decision theoretic perspective," in *In Proc. ICASSP'98*, vol. 1. IEEE, 1998, pp. 425–428.

[21] V. Goel, S. Kumar, and W. Byrne, "Segmental minimum bayes-risk asr voting strategies." in *INTERSPEECH*, 2000, pp. 139–142.

[22] V. Goel, S. Kumar, and W. J. Byrne, "Confidence based lattice segmentation and minimum bayes-risk decoding." in *INTERSPEECH*, 2001, pp. 2569–2572.

[23] V. Doumpiotis, S. Tsakalidis, and W. J. Byrne, "Lattice segmentation and minimum bayes risk discriminative training." in *INTERSPEECH*, 2003.

[24] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," Mountain View, CA, USA, Tech. Rep., 2004.

[25] R. R. Kelly, J. A. Albertini, and N. B. Shannon, "Deaf college students' reading comprehension and strategy use," *American annals of the deaf*, vol. 146, no. 5, pp. 385–400, 2001.

[26] R. Baayen, R. Piepenbrock, and L. Gulikers, "Celex2," 1995.

[27] S. Bird, "Nltk: the natural language toolkit," in *In Proc. COLING'06*. Association for Computational Linguistics, 2006, pp. 69–72.

[28] F. Pedregosa, A. Varoquaux, V. Michel, and Thirion, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] M. Hart, *Project gutenberg*. Project Gutenberg, 2004.

[30] S. Loria, "Textblob: simplified text processing," *Secondary TextBlob: Simplified Text Processing*, 2014.