Which Words Are Hard to Recognize? Prosodic, Lexical, and Disfluency Factors that Increase ASR Error Rates

Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase ASR error rates

Sharon Goldwater, Dan Jurafsky and Christopher D. Manning
Department of Linguistics and Computer Science
Stanford University
{sgwater,jurafsky,manning}@stanford.edu

Abstract

Many factors are thought to increase the chances of misrecognizing a word in ASR, including low frequency, nearby disfluencies, short duration, and being at the start of a turn. However, few of these factors have been formally examined. This paper analyzes a variety of lexical, prosodic, and disfluency factors to determine which are likely to increase ASR error rates. Findings include the following. (1) For disfluencies, effects depend on the type of disfluency: errors *increase* by up to 15% (absolute) for words near fragments, but *decrease* by up to 7.2% (absolute) for words near repetitions. This decrease seems to be due to longer word duration. (2) For prosodic features, there are more errors for words with *extreme* values than words with *typical* values. (3) Although our results are based on output from a system with speaker adaptation, speaker differences are a major factor influencing error rates, and the effects of features such as frequency, pitch, and intensity may vary between speakers.

1 Introduction

In order to improve the performance of automatic speech recognition (ASR) systems on conversational speech, it is important to understand the factors that cause problems in recognizing words. Previous work on recognition of spontaneous monologues and dialogues has shown that infrequent words are more likely to be misrecognized (Fosler-Lussier and Morgan, 1999; Shinozaki and Furui, 2001) and that fast speech increases error rates (Siegler and Stern, 1995; Fosler-Lussier and Morgan, 1999; Shinozaki and Furui, 2001). Siegler and Stern (1995) and Shinozaki and Furui (2001) also found higher error rates in very slow speech. Word length (in phones) has also been found to be a useful predictor of higher error rates (Shinozaki and Furui, 2001). In Hirschberg et al.'s (2004) analysis of two human-computer dialogue systems, misrecognized turns were found to have (on average) higher maximum pitch and energy than correctly recognized turns. Results for speech rate were ambiguous: faster utterances had higher error rates in one corpus, but lower error rates in the other. Finally, Adda-Decker and Lamel (2005) demonstrated that both French and English ASR systems had more trouble with male speakers than female speakers, and found several possible explanations, including higher rates of disfluencies and more reduction.

Many questions are left unanswered by these previous studies. In the word-level analyses of Fosler-Lussier and Morgan (1999) and Shinozaki and Furui (2001), only substitution and deletion errors were considered, so we do not know how including insertions might affect the results. Moreover, these studies primarily analyzed lexical, rather than prosodic, factors. Hirschberg et al.'s (2004) work suggests that prosodic factors can impact error rates, but leaves open the question of which factors are important at the word level and how they influence recognition of natural conversational speech. Adda-Decker and Lamel's (2005) suggestion that higher rates of disfluency are a cause of worse recognition for male speakers presupposes that disfluencies raise error rates. While this assumption seems natural, it has yet to be carefully tested, and in particular we do not

Proceedings of ACL-08: HLT, pages 380–388,
Columbus, Ohio, USA, June 2008.

know whether disfluent words are associated with errors in adjacent words, or are simply more likely to be misrecognized themselves. Other factors that are often thought to affect a word's recognition, such as its status as a content or function word, and whether it starts a turn, also remain unexamined.

The present study is designed to address all of these questions by analyzing the effects of a wide range of lexical and prosodic factors on the accuracy of an English ASR system for conversational telephone speech. In the remainder of this paper, we first describe the data set used in our study and introduce a new measure of error, *individual word error rate* (IWER), that allows us to include insertion errors in our analysis, along with deletions and substitutions. Next, we present the features we collected for each word and the effects of those features individually on IWER. Finally, we develop a joint statistical model to examine the effects of each feature while controlling for possible correlations.

2 Data

For our analysis, we used the output from the SRI/ICSI/UW RT-04 CTS system (Stolcke et al., 2006) on the NIST RT-03 development set. This system's performance was state-of-the-art at the time of the 2004 evaluation. The data set contains 36 telephone conversations (72 speakers, 38477 reference words), half from the Fisher corpus and half from the Switchboard corpus.[1]

The standard measure of error used in ASR is *word error rate* (WER), computed as $100(I + D + S)/R$, where I, D and S are the number of insertions, deletions, and substitutions found by aligning the ASR hypotheses with the reference transcriptions, and R is the number of reference words. Since we wish to know what features of a reference word increase the probability of an error, we need a way to measure the errors attributable to individual words — an *individual word error rate* (IWER). We assume that a substitution or deletion error can be assigned to its corresponding reference word, but for insertion errors, there may be two adjacent reference words that could be responsible. Our solution is to assign any insertion errors to each of the adjacent words. We could then define IWER as $100(n_i + n_d + n_s)/R$, where n_i, n_d, and n_s are the insertion, deletion, and substitution counts for individual words (with $n_d = D$ and $n_s = S$). In general, however, $n_i > I$, so that the IWER for a given data set would be larger than the WER. To facilitate comparisons with standard WER, we therefore discount insertions by a factor α, such that $\alpha n_i = I$. In this study, $\alpha = .617$.

[1] These conversations are not part of the standard Fisher and Switchboard corpora used to train most ASR systems.

	Ins	Del	Sub	Total	% data
Full word	1.6	6.9	10.5	19.0	94.2
Filled pause	0.6	–	16.4	17.0	2.8
Fragment	2.3	–	17.3	19.6	2.0
Backchannel	0.3	30.7	5.0	36.0	0.6
Guess	1.6	–	30.6	32.1	0.4
Total	1.6	6.7	10.9	19.7	100

Table 1: Individual word error rates for different word types, and the proportion of words belonging to each type. Deletions of filled pauses, fragments, and guesses are not counted as errors in the standard scoring method.

3 Analysis of individual features

3.1 Features

The reference transcriptions used in our analysis distinguish between five different types of words: filled pauses (*um*, *uh*), fragments (*wh-*, *redistr-*), backchannels (*uh-huh*, *mm-hm*), guesses (where the transcribers were unsure of the correct words), and full words (everything else). Error rates for each of these types can be found in Table 1. The remainder of our analysis considers only the 36159 invocabulary full words in the reference transcriptions (70 OOV full words are excluded). We collected the following features for these words:

Speaker sex Male or female.

Broad syntactic class Open class (e.g., nouns and verbs), closed class (e.g., prepositions and articles), or discourse marker (e.g., *okay*, *well*). Classes were identified using a POS tagger (Ratnaparkhi, 1996) trained on the tagged Switchboard corpus.

Log probability The unigram log probability of each word, as listed in the system's language model.

Word length The length of each word (in phones), determined using the most frequent pronunciation

BefRep	FirRep	MidRep	LastRep	AfRep	BefFP	AfFP	BefFr	AfFr
yeah	i	i	i	think you	should um	ask for	the ref-	recommendation

Figure 1: Example illustrating disfluency features: words occurring before and after repetitions, filled pauses, and fragments; first, middle, and last words in a repeated sequence.

found for that word in the recognition lattices.

Position near disfluency A collection of features indicating whether a word occurred before or after a filled pause, fragment, or repeated word; or whether the word itself was the first, last, or other word in a sequence of repetitions. Figure 1 illustrates. Only identical repeated words with no intervening words or filled pauses were considered repetitions.

First word of turn Turn boundaries were assigned automatically at the beginning of any utterance following a pause of at least 100 ms during which the other speaker spoke.

Speech rate The average speech rate (in phones per second) was computed for each utterance using the pronunciation dictionary extracted from the lattices and the utterance boundary timestamps in the reference transcriptions.

In addition to the above features, we used Praat (Boersma and Weenink, 2007) to collect the following additional prosodic features on a subset of the data obtained by excluding all contractions:[2]

Pitch The minimum, maximum, mean, and range of pitch for each word.

Intensity The minimum, maximum, mean, and range of intensity for each word.

Duration The duration of each word.

31017 words (85.8% of the full-word data set) remain in the no-contractions data set after removing words for which pitch and/or intensity features could not be extracted.

[2]Contractions were excluded before collecting prosodic features for the following reason. In the reference transcriptions and alignments used for scoring ASR systems, contractions are treated as two separate words. However, aside from speech rate, our prosodic features were collected using word-by-word timestamps from a forced alignment that used a transcription where contractions are treated as single words. Thus, the start and end times for a contraction in the forced alignment correspond to two words in the alignments used for scoring, and it is not clear how to assign prosodic features appropriately to those words.

3.2 Results and discussion

Results of our analysis of individual features can be found in Table 2 (for categorical features) and Figure 2 (for numeric features). Comparing the error rates for the full-word and the no-contractions data sets in Table 2 verifies that removing contractions does not create systematic changes in the patterns of errors, although it does lower error rates (and significance values) slightly overall. (First and middle repetitions are combined as non-final repetitions in the table, because only 52 words were middle repetitions, and their error rates were similar to initial repetitions.)

3.2.1 Disfluency features

Perhaps the most interesting result in Table 2 is that the effects of disfluencies are highly variable depending on the type of disfluency and the position of a word relative to it. Non-final repetitions and words next to fragments have an IWER up to 15% (absolute) *higher* than the average word, while final repetitions and words following repetitions have an IWER up to 7.2% *lower*. Words occurring before repetitions or next to filled pauses do not have significantly different error rates than words not in those positions. Our results for repetitions support Shriberg's (1995) hypothesis that the final word of a repeated sequence is in fact fluent.

3.2.2 Other categorical features

Our results support the common wisdom that open class words have lower error rates than other words (although the effect we find is small), and that words at the start of a turn have higher error rates. Also, like Adda-Decker and Lamel (2005), we find that male speakers have higher error rates than females, though in our data set the difference is more striking (3.6% absolute, compared to their 2.0%).

3.2.3 Word probability and word length

Turning to Figure 2, we find (consistent with previous results) that low-probability words have dramatically higher error rates than high-probability

		Filled Pau.		Fragment		Repetition				Syntactic Class				Sex		
		Bef	Aft	Bef	Aft	Bef	Aft	NonF	Fin	Clos	Open	Disc	1st	M	F	All
(a)	IWER	17.6	16.9	***33.8***	**21.6**	16.7	**13.8**	***26.0***	***11.6***	***19.7***	***18.0***	19.6	***21.2***	***20.6***	***17.0***	18.8
	% wds	1.7	1.7	1.6	1.5	0.7	0.9	1.2	1.1	43.8	50.5	5.8	6.2	52.5	47.5	100
(b)	IWER	17.6	17.2	***32.0***	**21.5**	15.8	14.2	***25.1***	***11.6***	18.8	**17.8**	19.0	**20.3**	***20.0***	***16.4***	18.3
	% wds	1.9	1.8	1.6	1.5	0.8	0.8	1.4	1.1	43.9	49.6	6.6	6.4	52.2	47.8	100

Table 2: IWER by feature and percentage of words exhibiting each feature for (a) the full-word data set and (b) the no-contractions data set. Error rates that are significantly different for words with and without a given feature (computed using 10,000 samples in a Monte Carlo permutation test) are in **bold** ($p < .05$) or ***bold italics*** ($p < .005$). Features shown are whether a word occurs before or after a filled pause, fragment, or repetition; is a non-final or final repetition; is open class, closed class, or a discourse marker; is the first word of a turn; or is spoken by a male or female. *All* is the IWER for the entire data set. (Overall IWER is slightly lower than in Table 1 due to the removal of OOV words.)

words. More surprising is that word length in phones does *not* seem to have a consistent effect on IWER. Further analysis reveals a possible explanation: word length is correlated with duration, but anti-correlated to the same degree with log probability (the Kendall τ statistics are .50 and -.49). Figure 2 shows that words with longer duration have lower IWER. Since words with more phones tend to have longer duration, but lower frequency, there is no overall effect of length.

3.2.4 Prosodic features

Figure 2 shows that means of pitch and intensity have relatively little effect except at extreme values, where more errors occur. In contrast, pitch and intensity range show clear linear trends, with greater range of pitch or intensity leading to lower IWER.[3] As noted above, decreased duration is associated with increased IWER, and (as in previous work), we find that IWER increases dramatically for fast speech. We also see a tendency towards higher IWER for very slow speech, consistent with Shinozaki and Furui (2001) and Siegler and Stern (1995). The effects of pitch minimum and maximum are not shown for reasons of space, but are similar to pitch mean. Also not shown are intensity minimum (with more errors at higher values) and intensity maximum (with more errors at lower values).

For most of our prosodic features, as well as log probability, extreme values seem to be associated with worse recognition than average values. We explore this possibility further in Section 4.

[3]Our decision to use the log transform of pitch range was originally based on the distribution of pitch range values in the data set. Exploratory data analysis also indicated that using the transformed values would likely lead to a better model fit (Section 4) than using the raw values.

4 Analysis using a joint model

In the previous section, we investigated the effects of various individual features on ASR error rates. However, there are many correlations between these features – for example, words with longer duration are likely to have a larger range of pitch and intensity. In this section, we build a single model with all of our features as potential predictors in order to determine the effects of each feature after controlling for the others. We use the no-contractions data set so that we can include prosodic features in our model. Since only 1% of tokens have an IWER > 1, we simplify modeling by predicting only whether each token is responsible for an error or not. That is, our dependent variable is binary, taking on the value 1 if IWER > 0 for a given token and 0 otherwise.

4.1 Model

To model data with a binary dependent variable, a logistic regression model is an appropriate choice. In logistic regression, we model the *log odds* as a linear combination of feature values $x_0 \ldots x_n$:

$$\log \frac{p}{1-p} = \beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

where p is the probability that the outcome occurs (here, that a word is misrecognized) and $\beta_0 \ldots \beta_n$ are coefficients (feature weights) to be estimated. Standard logistic regression models assume that all categorical features are *fixed effects*, meaning that all possible values for these features are known in advance, and each value may have an arbitrarily different effect on the outcome. However, features

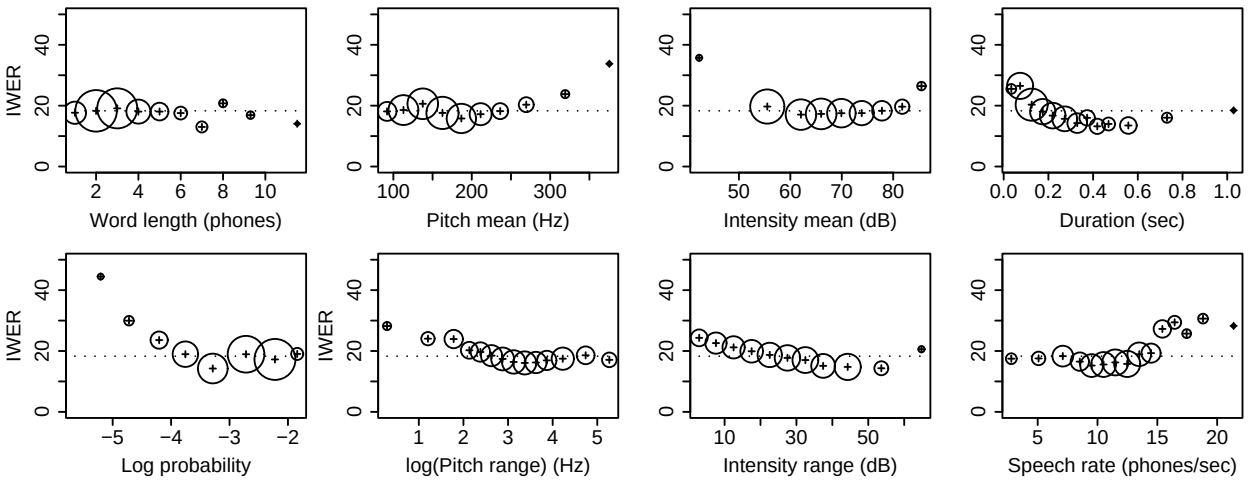

Figure 2: Effects of numeric features on IWER of the SRI system for the no-contractions data set. All feature values were binned, and the average IWER for each bin is plotted, with the area of the surrounding circle proportional to the number of points in the bin. Dotted lines show the average IWER over the entire data set.

such as speaker identity do not fit this pattern. Instead, we control for speaker differences by assuming that speaker identity is a *random effect*, meaning that the speakers observed in the data are a random sample from a larger population. The baseline probability of error for each speaker is therefore assumed to be a normally distributed random variable, with mean equal to the population mean, and variance to be estimated by the model. Stated differently, a random effect allows us to add a factor to the model for speaker identity, without allowing arbitrary variation in error rates between speakers. Models such as ours, with both fixed and random effects, are known as *mixed-effects models*, and are becoming a standard method for analyzing linguistic data (Baayen, 2008). We fit our models using the lme4 package (Bates, 2007) of R (R Development Core Team, 2007).

To analyze the joint effects of all of our features, we initially built as large a model as possible, and used *backwards elimination* to remove features one at a time whose presence did not contribute significantly (at $p \leq .05$) to model fit. All of the features shown in Table 2 were converted to binary variables and included as predictors in our initial model, along with a binary feature controlling for corpus (Fisher or Switchboard), and all numeric features in Figure 2. We did not include minimum and maximum values for pitch and intensity because they are highly correlated with the mean values, making parameter estimation in the combined model difficult. Preliminary investigation indicated that using the mean values would lead to the best overall fit to the data.

In addition to these basic fixed effects, our initial model included quadratic terms for all of the numeric features, as suggested by our analysis in Section 3, as well as random effects for speaker identity and word identity. All numeric features were rescaled to values between 0 and 1 so that coefficients are comparable.

4.2 Results and discussion

Figure 3 shows the estimated coefficients and standard errors for each of the fixed effect categorical features remaining in the reduced model (i.e., after backwards elimination). Since all of the features are binary, a coefficient of β indicates that the corresponding feature, when present, adds a weight of β to the log odds (i.e., multiplies the odds of an error by a factor of e^{β}). Thus, features with positive coefficients *increase* the odds of an error, and features with negative coefficients *decrease* the odds of an error. The magnitude of the coefficient corresponds to the size of the effect.

Interpreting the coefficients for our numeric features is less intuitive, since most of these variables have both linear and quadratic effects. The contribution to the log odds of a particular numeric feature

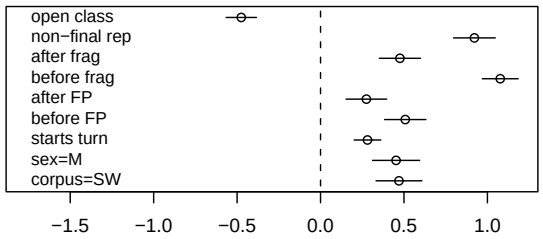

Figure 3: Estimates and standard errors of the coefficients for the categorical predictors in the reduced model.

x_i, with linear and quadratic coefficients a and b, is $ax_i + bx_i^2$. We plot these curves for each numeric feature in Figure 4. Values on the x axes with positive y values indicate increased odds of an error, and negative y values indicate decreased odds of an error. The x axes in these plots reflect the rescaled values of each feature, so that 0 corresponds to the minimum value in the data set, and 1 to the maximum value.

4.2.1 Disfluencies

In our analysis of individual features, we found that different types of disfluencies have different effects: non-final repeated words and words near fragments have higher error rates, while final repetitions and words following repetitions have lower error rates. After controlling for other factors, a different picture emerges. There is no longer an effect for final repetitions or words after repetitions; all other disfluency features increase the odds of an error by a factor of 1.3 to 2.9. These differences from Section 3 can be explained by noting that words near filled pauses and repetitions have longer durations than other words (Bell et al., 2003). Longer duration lowers IWER, so controlling for duration reveals the negative effect of the nearby disfluencies. Our results are also consistent with Shriberg's (1995) findings on fluency in repeated words, since final repetitions have no significant effect in our combined model, while non-final repetitions incur a penalty.

4.2.2 Other categorical features

Without controlling for other lexical or prosodic features, we found that a word is more likely to be misrecognized at the beginning of a turn, and less likely to be misrecognized if it is an open class word. According to our joint model, these effects still hold even after controlling for other features. Similarly, male speakers still have higher error rates than females. This last result sheds some light on the work of Adda-Decker and Lamel (2005), who suggested several factors that could explain males' higher error rates. In particular, they showed that males have higher rates of disfluency, produce words with slightly shorter durations, and use more alternate ("sloppy") pronunciations. Our joint model controls for the first two of these factors, suggesting that the third factor or some other explanation must account for the remaining differences between males and females. One possibility is that female speech is more easily recognized because females tend to have expanded vowel spaces (Diehl et al., 1996), a factor that is associated with greater intelligibility (Bradlow et al., 1996) and is characteristic of genres with lower ASR error rates (Nakamura et al., 2008).

4.2.3 Prosodic features

Examining the effects of pitch and intensity individually, we found that increased range for these features is associated with lower IWER, while higher pitch and extremes of intensity are associated with higher IWER. In the joint model, we see the same effect of pitch mean and an even stronger effect for intensity, with the predicted odds of an error dramatically higher for extreme intensity values. Meanwhile, we no longer see a benefit for increased pitch range and intensity; rather, we see small quadratic effects for both features, i.e. words with average ranges of pitch and intensity are recognized more easily than words with extreme values for these features. As with disfluencies, we hypothesize that the linear trends observed in Section 3 are primarily due to effects of duration, since duration is moderately correlated with both log pitch range ($\tau = .35$) and intensity range ($\tau = .41$).

Our final two prosodic features, duration and speech rate, showed strong linear and weak quadratic trends when analyzed individually. According to our model, both duration and speech rate are still important predictors of error after controlling for other features. However, as with the other prosodic features, predictions of the joint model are dominated by quadratic trends, i.e., predicted error rates are lower for average values of duration and speech rate than for extreme values.

Overall, the results from our joint analysis suggest

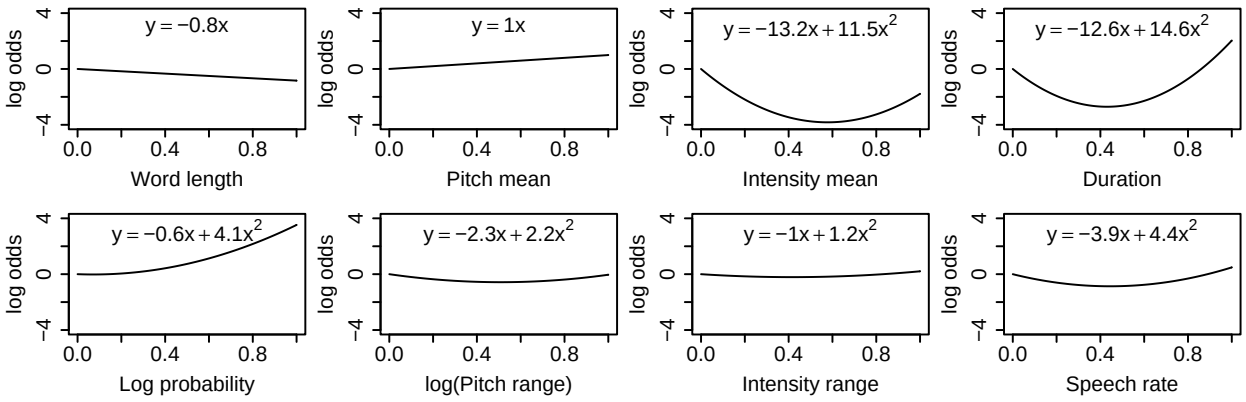

Figure 4: Predicted effect on the log odds of each numeric feature, including linear and (if applicable) quadratic terms.

Model	Neg. log lik.	Diff.	df
Full	12932	0	32
Reduced	12935	3	26
No lexical	13203	271	16
No prosodic	13387	455	20
No speaker	13432	500	31
No word	13267	335	31
Baseline	14691	1759	1

Table 3: Fit to the data of various models. Degrees of freedom (df) for each model is the number of fixed effects plus the number of random effects plus 1 (for the intercept). *Full* model contains all predictors; *Reduced* contains only predictors contributing significantly to fit; *Baseline* contains only intercept. Other models are obtained by removing features from *Full*. *Diff* is the difference in log likelihood between each model and *Full*.

that, after controlling for other factors, *extreme* values for prosodic features are associated with worse recognition than *typical* values.

4.2.4 Differences between lexical items

As discussed above, our model contains a random effect for word identity, to control for the possibility that certain lexical items have higher error rates that are not explained by any of the other factors in the model. It is worth asking whether this random effect is really necessary. To address this question, we compared the fit to the data of two models, each containing all of our fixed effects and a random effect for speaker identity. One model also contained a random effect for word identity. Results are shown in Table 3. The model without a random effect for word identity is significantly worse than the full model; in fact, this single parameter is more important than all of the lexical features combined. To see which lexical items are causing the most difficulty, we examined the items with the highest estimated increases in error. The top 20 items on this list include *yup*, *yep*, *yes*, *buy*, *then*, *than*, and *r.*, all of which are acoustically similar to each other or to other high-frequency words, as well as the words *after*, *since*, *now*, and *though*, which occur in many syntactic contexts, making them difficult to predict based on the language model.

4.2.5 Differences between speakers

We examined the importance of the random effect for speaker identity in a similar fashion to the effect for word identity. As shown in Table 3, speaker identity is a very important factor in determining the probability of error. That is, the lexical and prosodic variables examined here are not sufficient to fully explain the differences in error rates between speakers. In fact, the speaker effect is the single most important factor in the model.

Given that the differences in error rates between speakers are so large (average IWER for different speakers ranges from 5% to 51%), we wondered whether our model is sufficient to capture the kinds of speaker variation that exist. The model assumes that each speaker has a different baseline error rate, but that the effects of each variable are the same for each speaker. Determining the extent to which this assumption is justified is beyond the scope of this paper, however we present some suggestive results in Figure 5. This figure illustrates some of the dif-

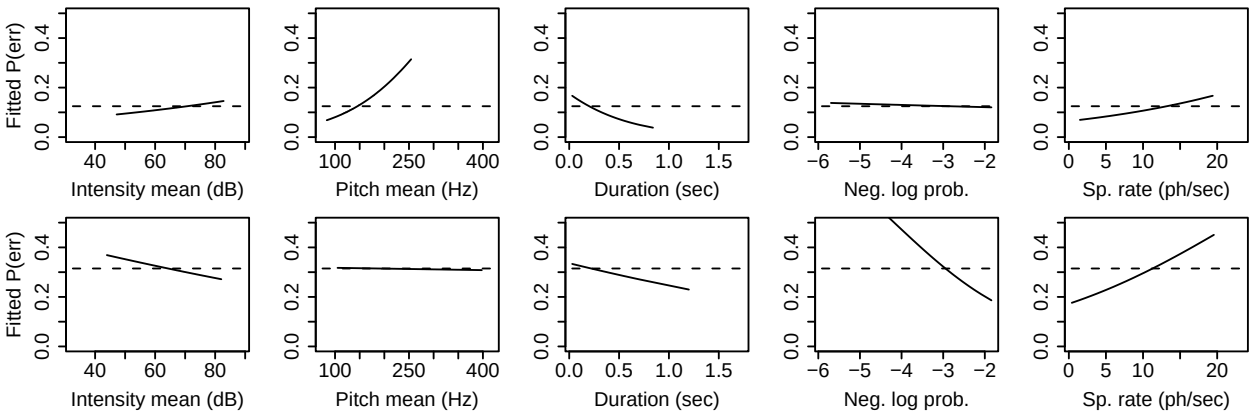

Figure 5: Estimated effects of various features on the error rates of two different speakers (top and bottom). Dashed lines illustrate the baseline probability of error for each speaker. Solid lines were obtained by fitting a logistic regression model to each speaker's data, with the variable labeled on the x-axis as the only predictor.

ferences between two speakers chosen fairly arbitrarily from our data set. Not only are the baseline error rates different for the two speakers, but the effects of various features appear to be very different, in one case even reversed. The rest of our data set exhibits similar kinds of variability for many of the features we examined. These differences in ASR behavior between speakers are particularly interesting considering that the system we investigated here already incorporates speaker adaptation models.

5 Conclusion

In this paper, we introduced the *individual word error rate* (IWER) for measuring ASR performance on individual words, including insertions as well as deletions and substitutions. Using IWER, we analyzed the effects of various word-level lexical and prosodic features, both individually and in a joint model. Our analysis revealed the following effects. (1) Words at the start of a turn have slightly higher IWER than average, and open class (content) words have slightly lower IWER. These effects persist even after controlling for other lexical and prosodic factors. (2) Disfluencies heavily impact error rates: IWER for non-final repetitions and words adjacent to fragments rises by up to 15% absolute, while IWER for final repetitions and words following repetitions decreases by up to 7.2% absolute. Controlling for prosodic features eliminates the latter benefit, and reveals a negative effect of adjacent filled pauses, suggesting that the effects of these disfluencies are normally obscured by the greater duration of nearby words. (3) For most acoustic-prosodic features, words with extreme values have worse recognition than words with average values. This effect becomes much more pronounced after controlling for other factors. (4) After controlling for lexical and prosodic characteristics, the lexical items with the highest error rates are primarily homophones or near-homophones (e.g., *buy* vs. *by*, *then* vs. *than*). (5) Speaker differences account for much of the variance in error rates between words. Moreover, the direction and strength of effects of different prosodic features may vary between speakers.

While we plan to extend our analysis to other ASR systems in order to determine the generality of our findings, we have already gained important insights into a number of factors that increase ASR error rates. In addition, our results suggest a rich area for future research in further analyzing the variability of both lexical and prosodic effects on ASR behavior for different speakers.

Acknowledgments

This work was supported by the Edinburgh-Stanford LINK and ONR MURI award N000140510388. We thank Andreas Stolcke for providing the ASR output, language model, and forced alignments used here, and Raghunandan Kumaran and Katrin Kirchhoff for earlier datasets and additional help.

References

M. Adda-Decker and L. Lamel. 2005. Do speech recognizers prefer female speakers? In *Proceedings of INTERSPEECH*, pages 2205–2208.

R. H. Baayen. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics*. Cambridge University Press. Prepublication version available at http://www.mpi.nl/world/persons/private/baayen/publications.html.

Douglas Bates, 2007. *lme4: Linear mixed-effects models using S4 classes*. R package version 0.99875-8.

A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America*, 113(2):1001–1024.

P. Boersma and D. Weenink. 2007. Praat: doing phonetics by computer (version 4.5.16). http://www.praat.org/.

A. Bradlow, G. Torretta, and D. Pisoni. 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20:255–272.

R. Diehl, B. Lindblom, K. Hoemeke, and R. Fahey. 1996. On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of Phonetics*, 24:187–208.

E. Fosler-Lussier and N. Morgan. 1999. Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, 29:137– 158.

J. Hirschberg, D. Litman, and M. Swerts. 2004. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43:155– 175.

M. Nakamura, K. Iwano, and S. Furui. 2008. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech and Language*, 22:171– 184.

R Development Core Team, 2007. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

A. Ratnaparkhi. 1996. A Maximum Entropy model for part-of-speech tagging. In *Proceedings of the First Conference on Empirical Methods in Natural Language Processing*, pages 133–142.

T. Shinozaki and S. Furui. 2001. Error analysis using decision trees in spontaneous presentation speech recognition. In *Proceedings of ASRU 2001*.

E. Shriberg. 1995. Acoustic properties of disfluent repetitions. In *Proceedings of the International Congress of Phonetic Sciences*, volume 4, pages 384–387.

M. Siegler and R. Stern. 1995. On the effects of speech rate in large vocabulary speech recognition systems. In *Proceedings of ICASSP*.

A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lin, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu. 2006. Recent innovations in speech-to-text transcription at SRI-ICSI-UW. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1729–1744.