HAL
open science

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

Sharon Goldwater, Dan Jurafsky, Christopher D. Manning

To cite this version:

Sharon Goldwater, Dan Jurafsky, Christopher D. Manning. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. Speech Communication, 2010, 52 (3), pp.181. 10.1016/j.specom.2009.10.001 . hal-00608401

HAL Id: hal-00608401

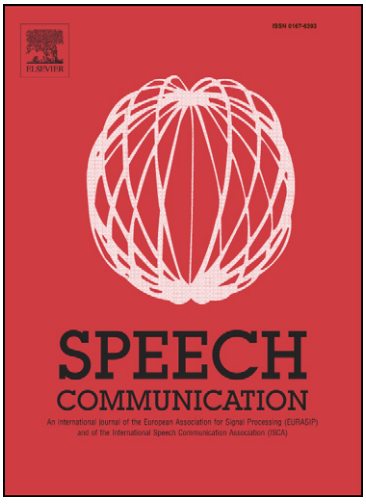
https://hal.science/hal-00608401

Submitted on 13 Jul 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Accepted Manuscript



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

Sharon Goldwater, Dan Jurafsky, Christopher D. Manning

PII: S0167-6393(09)00159-9
DOI: 10.1016/j.specom.2009.10.001
Reference: SPECOM 1837

To appear in: *Speech Communication*

Received Date: 12 July 2009
Revised Date: 6 October 2009
Accepted Date: 9 October 2009

Please cite this article as: Goldwater, S., Jurafsky, D., Manning, C.D., Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates, *Speech Communication* (2009), doi: 10.1016/j.specom.2009.10.001

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

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

Sharon Goldwater *

School of Informatics, University of Edinburgh, 10 Crichton St., Edinburgh, EH8 9AB, United Kingdom

Dan Jurafsky

Department of Linguistics, Stanford University, Margaret Jacks Hall, Bldg. 460, Stanford, CA 94305, USA

Christopher D. Manning

Department of Computer Science, Stanford University, 353 Serra Mall, Stanford, CA 94305, USA

Abstract

Despite years of speech recognition research, little is known about which words tend to be misrecognized and why. Previous work has shown that errors increase for infrequent words, short words, and very loud or fast speech, but many other presumed causes of error (e.g., nearby disfluencies, turn-initial words, phonetic neighborhood density) have never been carefully tested. The reasons for the huge differences found in error rates between speakers also remain largely mysterious.

Using a mixed-effects regression model, we investigate these and other factors by analyzing the errors of two state-of-the-art recognizers on conversational speech. Words with higher error rates include those with extreme prosodic characteristics, those occurring turn-initially or as discourse markers, and *doubly confusable pairs*: acoustically similar words that also have similar language model probabilities. Words preceding disfluent interruption points (first repetition tokens and words before fragments) also have higher error rates. Finally, even after accounting for other factors, speaker differences cause enormous variance in error rates, suggesting that speaker error rate variance is not fully explained by differences in word choice, fluency, or prosodic characteristics. We also propose that doubly confusable pairs, rather than high neighborhood density, may better explain phonetic neighborhood errors in human speech processing.

Key words: speech recognition, conversational, error analysis, individual differences, mixed-effects model

ACCEPTED MANUSCRIPT

1 Introduction

Conversational speech is one of the most difficult genres for automatic speech recognition (ASR) systems to recognize, due to high levels of disfluency, non-canonical pronunciations, and acoustic and prosodic variability. In order to improve ASR performance, it is important to understand which of these factors is most problematic for recognition. 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 is associated with higher error rates (Siegler and Stern, 1995; Fosler-Lussier and Morgan, 1999; Shinozaki and Furui, 2001). In some studies, very slow speech has also been found to correlate with higher error rates (Siegler and Stern, 1995; Shinozaki and Furui, 2001). In Shinozaki and Furui's (2001) analysis of a Japanese ASR system, word length (in phones) was found to be a useful predictor of error rates, with more errors on shorter words. 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 suggested several possible explanations, including higher rates of disfluencies and more reduction.

In parallel to these studies within the speech-recognition community, a body of work has accumulated in the psycholinguistics literature examining factors that affect the speed and accuracy of spoken word recognition in humans. Experiments are typically carried out using isolated words as stimuli, and controlling for numerous factors such as word frequency, duration, and length. Like ASR systems, humans are better (faster and more accurate) at recognizing frequent words than infrequent words (Howes, 1954; Marslen-Wilson, 1987; Dahan et al., 2001). In addition, it is widely accepted that recognition is worse for words that are phonetically similar to many other words than for highly distinctive words (Luce and Pisoni, 1998). Rather than using a graded notion of phonetic similarity, psycholinguistic experiments typically make the simplifying assumption that two words are "similar" if they differ by a single phone (insertion, substitution, or deletion). Such pairs are referred to as *neighbors*. Early on, it was shown that both the number of neighbors of a word and the frequency of those neighbors are significant predictors of recognition performance; it is now common to see those two factors combined into a single predictor known as *frequency-weighted neighborhood density* (Luce

* Corresponding author. Tel: +44 131 651 5609, Fax: +44 131 650 6626.
Email addresses: sgwater@inf.ed.ac.uk (Sharon Goldwater), jurafsky@stanford.edu (Dan Jurafsky), manning@stanford.edu (Christopher D. Manning).

ACCEPTED MANUSCRIPT

and Pisoni, 1998; Vitevitch and Luce, 1999), which we discuss in more detail in Section 3.1.

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, and it is unclear whether including insertions would have led to different results. Moreover, these studies primarily analyzed lexical, rather than prosodic, factors. Hirschberg et al.'s (2004) work suggests that utterance-level prosodic factors can impact error rates in human-computer dialogues, 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 was never carefully tested, and in particular neither Ada-Decker and Lamel nor any of the other papers cited investigated whether disfluent words are associated with errors in adjacent words, or are simply more likely to be misrecognized themselves. Many factors that are often thought to influence error rates, such as a word's status as a content or function word, and whether it starts a turn, also remained unexamined. Next, the neighborhood-related factors found to be important in human word recognition have, to our knowledge, never even been proposed as possible explanatory variables in ASR, much less formally analyzed. Additionally, many of these factors are known to be correlated. Disfluent speech, for example, is linked to changes in both prosody and rate of speech, and low-frequency words tend to have longer duration. Since previous work has generally examined each factor independently, it is not clear which factors would still be linked to word error after accounting for these correlations.

A final issue not addressed by these previous studies is that of speaker differences. While ASR error rates are known to vary enormously between speakers (Doddington and Schalk, 1981; Nusbaum and Pisoni, 1987; Nusbaum et al., 1995), most previous analyses have averaged over speakers rather than examining speaker differences explicitly, and the causes of such differences are not well understood. Several early hypotheses regarding the causes of these differences, such as the user's motivation to use the system or the variability of the user's speech with respect to user-specific training data (Nusbaum and Pisoni, 1987), can be ruled out for recognition of conversational speech because the user is speaking to another human and there is no user-specific training data. However, we still do not know the extent to which differences in error rates between speakers can be accounted for by the lexical, prosodic, and disfluency factors discussed above, or whether additional factors are at work.

The present study is designed to address the questions raised above by analyzing the effects of a wide range of lexical and prosodic factors on the accuracy of two English ASR systems for conversational telephone speech. We introduce a new

ACCEPTED MANUSCRIPT

measure of error, *individual word error rate* (IWER), that allows us to include insertion errors in our analysis, along with deletions and substitutions. Using this measure, we examine the effects of each factor on the recognition performance of two different state-of-the-art conversational telephone speech recognizers – the SRI/ICSI/UW RT-04 system (Stolcke et al., 2006) and the 2004 CU-HTK system (Evermann et al., 2004b, 2005). In the remainder of the paper, we first describe the data used in our study and the individual word error rate measure. 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 accounting for possible correlations, and to determine the relative importance of speaker differences other than those reflected in the features we collected.

2 Data

Our analysis is based on the output from two state-of-the-art speech recognition systems on the conversational telephone speech evaluation data from the National Institute of Standards and Technology (NIST) 2003 Rich Transcription exercise (RT-03). [1] The two recognizers are the SRI/ICSI/UW RT-04 system (Stolcke et al., 2006) and the 2004 CU-HTK system for the DARPA/NIST RT-04 evaluation (Evermann et al., 2004b, 2005). [2] Our goal in choosing these two systems, which we will refer to henceforth as the SRI system and the Cambridge system, was to select for state of the art peformance on conversational speech; these were two of the four best performing single research systems in the world as of the NIST evaluation (Fiscus et al., 2004).

The two systems use the same architecture that is standard in modern state-of-the-art conversational speech recognition systems. Both systems extract Mel frequency cepstral coefficients (MFCCs) with standard normalization and adaptation techniques: cepstral vocal tract length normalization (VTLN), heteroscedastic linear discriminant analysis (HLDA), cepstral mean and variance normalization, and maximum likelihood linear regression (MLLR). Both systems have gender-dependent acoustic models trained discriminatively using variants of minimum phone error (MPE) training with maximum mutual information (MMI) priors (Povey and Woodland, 2002). Both train their acoustic models on approximately 2400 hours of conversational telephone speech from the Switchboard, CallHome and Fisher corpora,

[1] We describe the NIST RT-03 data set briefly below; full details, including annotation guidelines, can be found at http://www.itl.nist.gov/iad/mig/tests/rt/2003-fall/index.html.

[2] The SRI/ICSI/UW system was developed by researchers at SRI International, the International Computer Science Institute, and the University of Washington. For more detailed descriptions of previous CU-HTK (Cambridge University Hidden Markov Model Toolkit) systems, see Hain et al. (2005) and Evermann et al. (2004a).

ACCEPTED MANUSCRIPT

consisting of 360 hours of speech used in the 2003 evaluation plus 1820 hours of noisy "quick transcriptions" from the Fisher corpus, although with different segmentation and filtering. Both use 4-gram models trained using in-domain telephone speech data as well as data harvested from the web (Bulyko et al., 2003). Both use many passes or tiers of decoding, each pass producing lattices that are passed on to the next pass for further processing.

The systems differ in a number of ways. The SRI system (Stolcke et al., 2006) uses perceptual linear predictive (PLP) features in addition to MFCC features, uses novel discriminative phone posterior features estimated by multilayer perceptrons, and uses a variant of MPE called minimum phone frame error (MPFE). The acoustic model includes a three-way system combination (MFCC non-cross-word triphones, MFCC cross-word triphones, and PLP cross-word triphones). Lattices are generated using a bigram language model, and rescored with duration models, a pause language model (Vergyri et al., 2002), and a syntactically rich SuperARV 'almost-parsing' language model (Wang and Harper, 2002), as well as the 4-gram models mentioned above. The word error rate of the SRI system on the NIST RT-03 evaluation data is 18.8%.

The Cambridge system (Evermann et al., 2004b, 2005) makes use of both single-pronunciation lexicons and and multiple-pronunciation lexicons using pronunciation probabilities. The acoustic model also includes a three-way system combination (multiple pronunciations with HLDA, multiple pronunciations without HLDA, and single pronunciations with HLDA). Each system uses cross-word triphones in a preliminary pass, then rescores with cross-word quinphone models. Whereas the SRI system uses a bigram language model in the first pass, then generates lattices with a trigram and rescores with a 4-gram and other language models, the Cambridge system uses a trigram language model in the first pass, then generates lattices with a 4-gram. The 4-gram language model includes a weighted combination of component models, some with Kneser-Ney and some with Good-Turing smoothing, and includes the interpolated 4-gram model used in the 2003 CU-HTK system (Evermann et al., 2004a). The word error rate of the Cambridge system on the NIST RT-03 evaluation data is 17.0%.

We examine the output of each system on the NIST RT-03 evaluation data. (Note that the developers of both the SRI and Cambridge systems had access to the evaluation data, and so the results for both systems will be slightly biased.) The data set contains 72 telephone conversations with 144 speakers and 76155 reference words. Half of the conversations are from the Fisher corpus and half from the Switchboard corpus (none from the standard portions of these corpora used to train most ASR systems). Utterance breakpoint timestamps (which determine the speech sequences sent to the recognizers) were assigned by the NIST annotators. The annotation guidelines state that breakpoints must be placed at turn boundaries (speaker changes), and may also be placed within turns. For within-turn breakpoints, the guidelines encourage annotators to place these during pauses (either disfluent or at

ACCEPTED MANUSCRIPT

```
REF:  but THERE are you know it is like *** other stuff
HYP:  but THEY  are you know ** is like THE other stuff
Eval:     S                  D          I
```

Fig. 1. An example alignment between the reference transcription (REF) and recognizer output (HYP), with substitutions (S), deletions (D), and insertions (I) shown. WER for this utterance is 30%.

phrasal boundaries), but also permit long sequences of fluent speech to be broken up into smaller units for easier transcription. Thus, in this corpus, the "utterances" being analyzed may comprise part or all of a turn, but do not in all cases correspond to natural breath groupings or phrasal units.

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 total number of insertions, deletions, and substitutions found by aligning the ASR hypotheses with the reference transcriptions, and R is the total number of reference words (see Figure 1).[3] However, WER can be computed only over full utterances or corpora. 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 (given a particular alignment), but for insertion errors, there may be two adjacent reference words that could be responsible. Since we have no way to know which word is responsible, we simply assign equal partial responsibility for any insertion errors to both of the adjacent words. That is, we define IWER for the ith reference word as

$$\text{IWER}(w_i) = del_i + sub_i + \alpha \cdot ins_i \tag{1}$$

where del_i and sub_i are binary variables indicating whether w_i is deleted or substituted, and ins_i counts the number of insertions adjacent to w_i. The discount factor α is chosen so that $\alpha \sum_{w_i} ins_i = I$ for the full corpus (i.e., the total penalty for insertion errors is the same as when computing WER). We then define IWER for a set of words as the average IWER for the individual words in the set:

$$\text{IWER}(w_1 \ldots w_n) = \frac{1}{n} \sum_{i=1}^{n} \text{IWER}(w_i) \tag{2}$$

We will sometimes refer to the IWER for a set of words as the average IWER (if necessary to distinguish from IWER for single words), and, as is standard with WER, we will present it as a percentage (e.g., as 18.2 rather than .182). Note that, due to the choice of the discount factor α, IWER = WER when computed over the entire data set, facilitating direct comparisons with other studies that use WER. In

[3] Our alignments and error rates were computed using the standard NIST speech recognition evaluation script `sclite`, along with the normalization (.glm) file used in the RT-03 evaluation, kindly provided by Phil Woodland.

ACCEPTED MANUSCRIPT

	SRI system				Cambridge system				
	Ins	Del	Sub	Total	Ins	Del	Sub	Total	% of data
Full word	1.5	6.5	10.4	18.4	1.5	6.2	9.1	16.8	94.2
Filled pause	0.6	–	15.4	16.1	0.9	–	15.1	16.0	2.9
Fragment	2.2	–	18.8	21.1	2.0	–	18.0	20.0	1.8
Backchannel	0.1	31.3	3.1	34.5	0.5	25.2	2.1	27.9	0.7
Guess	2.0	–	25.3	27.3	2.5	–	26.7	29.2	0.4
Total	1.4	6.4	10.7	18.5	1.5	6.0	9.5	17.0	100

Table 1
Individual word error rates for different word types in the two systems. Final column gives the proportion of words in the data belonging to each type. Deletions of filled pauses, fragments, and guesses are not counted as errors in the standard scoring method. The total error rate for the SRI system is slightly lower than the 18.8 WER from the NIST evaluation due to the removal of the 229 insertions mentioned in Footnote 4.

Reference	Forced alignment
`(%hesitation) in what way`	`um in what way`
`o. k.`	`okay`
`(%hesitation) i think it is because`	`uh i think it's because`

Table 2
Examples of differences in normalization between the reference transcriptions used for scoring and the transcriptions used to create a forced alignment with the speech signal.

this study, $\alpha = .584$ for the SRI system and .672 for the Cambridge system. [4]

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). Using our IWER measure, we computed error rates for each of these types of words, as shown in Table 1. Because many of the features we wish to analyze can be extracted only for full words, and because these words constitute the vast majority of the data, the remainder of this paper deals only with the 71579 in-vocabulary full words in the reference transcriptions (145 OOV full words are excluded). Nevertheless, we note the high rate of deletions for backchannel words; the high rate of substitutions for fragments and guesses is less surprising.

[4] Before computing α or doing any analysis, we first removed some recognized utterances consisting entirely of insertions. These utterances all came from a single conversation (sw_46732) in which one speaker's turns are (barely) audible on the other speaker's channel, and some of these turns were recognized by the systems. A total of 225 insertions were removed from the SRI output, 29 from the Cambridge output.

ACCEPTED MANUSCRIPT

A final point worth noting about our data set is that the reference transcriptions ordinarily used to compute WER (and here, IWER) are normalized in several ways, including treating all filled pauses as identical tokens and splitting contractions such as *it's* and *can't* into individual words (*it is*, *can not*). Unless otherwise specified in Section 3.1, all features we analyzed were extracted using the reference transcriptions. A few features were extracted with the help of a forced alignment (performed using the SRI recognizer, and kindly provided by Andreas Stolcke) between the speech signal and a slightly different set of transcriptions that more accurately reflects the speakers' true utterances. Examples of the differences between the reference transcriptions and the transcriptions used in the forced alignment are shown in Table 2. We describe below how this mismatch was handled for each relevant feature.

3 Analysis of individual features

In this section, we first describe all of the features we collected for each word and how the features were extracted. We then provide results detailing the association between each individual feature and recognition error rates.

3.1 Features

3.1.1 Disfluency features

Disfluencies are widely believed to increase ASR error rates, but there is little published evidence to support this belief. In order to examine this hypothesis, and determine whether different kinds of disfluencies have different effects on recognition, we collected several binary features indicating whether each word in the data occurred before, after, or as part of a disfluency. These features are listed below and illustrated in Figure 2.

Before/after filled pause. These features are present for words that appear immediately preceding or following a filled pause in the reference transcription.

Before/after fragment. These features are present for words that appear immediately preceding or following a fragment in the reference transcription.

Before/after repetition. These features are present for words that appear immediately preceding or following a sequence of repeated words in the reference transcription. Only identical repeated words with no intervening words or filled pauses were considered repetitions. While not all repeated words are necessarily disfluencies, we did not distinguish between disfluent and intentional repetitions.

ACCEPTED MANUSCRIPT

`yeah`	Before repetition
`i`	First repetition
`i`	Middle repetition
`i`	Last repetition
`think`	After repetition
`you`	
`should`	Before filled pause
`um`	
`ask`	After filled pause
`for`	
`the`	Before fragment
`ref-`	
`recommendation`	After fragment

Fig. 2. Example illustrating disfluency features: words occurring before and after repetitions, filled pauses, and fragments; first, middle, and last words in a repeated sequence.

Position in repeated sequence. These additional binary features indicate whether a word is itself the first, middle, or last word in a sequence of repetitions (see Figure 2).

3.1.2 Other categorical features

Of the remaining categorical (non-numeric) features we collected, we are aware of published results only for speaker sex (Adda-Decker and Lamel, 2005). However, anecdotal evidence suggests that the other features may be important in determining error rates. These features are:

Broad syntactic class. We divided words into three classes: open class (e.g., nouns and verbs), closed class (e.g., prepositions and articles), or discourse marker (e.g., *okay, well*). To obtain the feature value for each word, we first tagged our data set automatically and then collapsed the POS tags down to the three classes used for this feature. We used a freely available tagger (Ratnaparkhi, 1996) and trained it on the parsed portion of the Switchboard corpus in the Penn Treebank-3 release (Marcus et al., 1999). [5]

[5] We used the parsed files rather than the tagged files because we found the tags to be more accurate in the parsed version. Before training the tagger, we removed punctuation and normalized hesitations. Words tagged as foreign words, numbers, adjectives, adverbs,

ACCEPTED MANUSCRIPT

First word of turn. To compute this feature, 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. Preliminary analysis indicated that utterance-initial words behave similarly to turn-initial words; nevertheless, due to the possibility of within-turn utterance breakpoint annotations occurring during fluent speech (as described in Section 2), we did not include utterance-based features.

Speaker sex. This feature was extracted from the corpus meta-data.

3.1.3 Probability features

Previous work has shown that word frequencies and/or language model probabilities are an important predictor of word error rates (Fosler-Lussier and Morgan, 1999; Shinozaki and Furui, 2001). We used the n-gram language model from the SRI system in computing our probability features (see Section 2. Bulyko et al. (2003) provides details). Because the language model was trained on transcriptions whose normalization is closer to that of the forced alignment than to that of the reference transcriptions, we computed the probability of each reference word by heuristically aligning the forced alignment transcriptions to the reference transcriptions. For contractions listed as one word in the forced alignment and two words in the reference transcriptions (e.g., *it's* versus *it is*), both reference words were aligned to the same forced alignment word.[6] The probability of each reference word was then determined by looking up the probability of the corresponding forced alignment word in the language model. We used two different probability features, listed below.

Unigram log probability. This feature is based on simple word frequency, rather than context.

Trigram log probability. This feature corresponds more closely to the log probabilities assigned by language models in the two systems. It was computed from the language model files using Katz backoff smoothing.

3.1.4 Pronunciation-based features

The previous set of features allows us to examine the relationship between language model probabilities and word error rates; in this section we describe a set of features designed to focus on factors that might be related to acoustic con-

verbs, nouns, and symbols were assumed to be content words; others were assumed to be function words. In a hand-checked sample of 71 utterances, 782 out of 795 full words (98.4%) were labeled with the correct broad syntactic class.

[6] Due to slight differences between the two sets of transcriptions that could not be accounted for by normalization or other obvious changes, 446 full reference words (0.6%) could not be aligned, and were excluded from further analysis.

ACCEPTED MANUSCRIPT

fusability. Of the features we collect here, only word length has been examined previously in the ASR literature, to our knowledge (Shinozaki and Furui, 2001). Most of our pronunciation-based features are inspired by work in psycholinguistics demonstrating that human subjects have more difficulty recognizing spoken words that are in dense phonetic neighborhoods, i.e., when there are many other words that differ from the target word by only a single phone (Luce and Pisoni, 1998). In human word recognition studies, the effect of the number of neighbors of a word has been found to be moderated by the total frequency of those neighbors, with high-frequency neighbors leading to slower and less accurate recognition (Luce and Pisoni, 1998; Vitevitch and Luce, 1999). The two factors (number of neighbors and frequency of neighbors) are often combined into a single measure, *frequency-weighted neighborhood density*, which is generally thought to be a better predictor of recognition speed and accuracy than the raw number of neighbors. Frequency-weighted neighborhood density is computed as the sum of the (log or raw) frequencies of a word's neighbors, with frequencies computed from a large corpus. [7] It is worth noting that, unlike the words we examine here, the stimuli used in studies of human word recognition are generally controlled for many potential confounds such as word length, syllable shape and number (e.g., only monosyllabic CVC words are used), intensity, and speaker. In addition, stimuli are nearly always words presented in isolation. Thus, acoustically confusable words cannot be disambiguated based on context. It is an open question whether the standard neighborhood-based psycholinguistic measures are important predictors of error in ASR, where words are recognized in context.

Since we did not have access to the pronunciation dictionaries used by the two systems in our study, we computed our pronunciation-based features using the CMU Pronouncing Dictionary. [8] This dictionary differs from those ordinarily used in ASR systems in distinguishing between several levels of stress, distinguishing multiple unstressed vowels (as opposed to only two, ARPABet `ax` and `ix`), and including multiple pronunciations for a large proportion of words. In order to bring the dictionary more in line with standard ASR dictionaries, the following preprocessing steps were performed, as illustrated in Figure 3. First, where two pronunciations differed only by one containing a schwa where the other contained a different unstressed short vowel (non-diphthong), the pronunciation with the schwa was removed. Second, the unstressed central vowel `AH0` was converted to `AX`, and all other stress marks were removed. After preprocessing of the dictionary, the follow-

[7] The literature is inconsistent on the precise calculation of frequency-weighted neighborhood density, with the same authors using raw frequencies in some cases (Luce and Pisoni, 1998) and log frequencies in others (Luce et al., 2000; Vitevitch and Luce, 1999). Since these studies generally group stimuli into only two groups (low vs. high FWND), there is no way to determine whether log or raw frequencies are a better predictor. We will use log frequencies in this paper.

[8] The CMU Pronouncing Dictionary is available from http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

ACCEPTED MANUSCRIPT

Word	Original pronunciation	Final pronunciation
A	AH0	AX
A(2)	EY1	EY
ABDOMEN	AE0 B D OW1 M AH0 N	AE B D OW M AX N
ABDOMEN(2)	AE1 B D AH0 M AH0 N	AE B D AX M AX N
ABDOMINAL	AE0 B D AA1 M AH0 N AH0 L	AE B D AA M AX N AX L
ABDOMINAL(2)	AH0 B D AA1 M AH0 N AH0 L	[removed]
BARGAIN	B AA1 R G AH0 N	[removed]
BARGAIN(2)	B AA1 R G IH0 N	B AA R G IH N
THE	DH AH0	DH AX
THE(2)	DH AH1	DH AH
THE(3)	DH IY0	DH IY

Table 3
Example illustrating the preprocessing of the CMU Pronouncing Dictionary done before computing homophone and neighbor features. Numbers appended to phones in the original pronunciations indicate stress levels (0=none, 1=primary, 2= secondary). Stress marks are removed after deleting extra pronunciations differing only in unstressed non-diphthong vowels and converting AH0 to AX.

ing features were extracted.

Word length. Each word's length in phones was determined from its dictionary entry. If multiple pronunciations were listed for a single word, the number of phones in the first (longest) pronunciation was used. (Frequencies of the different pronunciations are not provided in the dictionary.)

Number of pronunciations. We extracted the number of different pronunciations for each word from the dictionary. Note that this number is not the same as the number of pronunciations used in the ASR systems' dictionaries. For all but very frequent words, ASR systems typically include only a single pronunciation; this feature may provide a better estimate of the actual pronunciation variability of different words.

Number of homophones. We defined a homophone of the target word to be any word in the dictionary with a pronunciation identical to any of the pronunciations of the target word, and counted the total number of these.

Number of neighbors. We computed the number of neighbors of each word by counting the number of distinct orthographic representations (other than the target word or its homophones) whose pronunciations were neighbors of any of the pronunciations of the target word. For example, neighbors of the word *aunt* include *auntie, ain't, and* (based on the first pronunciation, ae n t), as well as *want, on*,

ACCEPTED MANUSCRIPT

and *daunt* (based on the second pronunciation, `ao n t`).

Frequency-weighted homophones/neighbors. Although only frequency-weighted neighborhood density is a standard measure in psycholinguistics, we also computed frequency-weighted homophone density for completeness. Neighbors and homophones were determined as above; we estimated log frequencies using the unigram log probabilities in the SRI language model, subtracting the smallest log probability from all values to obtain non-negative log frequency values for all words. The feature values were then computed as the sum of the log frequencies of each homophone or neighbor.

3.1.5 Prosodic features

Of the prosodic features we collected, only speech rate has been analyzed extensively as a factor influencing word error rates in spontaneous speech (Siegler and Stern, 1995; Shinozaki and Furui, 2001; Fosler-Lussier and Morgan, 1999). We also extracted features based on three other standard acoustic-prosodic factors which could be expected to have some effect on recognition accuracy: pitch, intensity, and duration. The final prosodic feature we extracted was jitter, which is a correlate of creaky voice. Creaky voice is becoming widespread among younger Americans, especially females (Pennock-Speck, 2005; Ingle et al., 2005), and thus could be important to consider as a factor in recognition accuracy for these speakers.

To extract prosodic features, the transcriptions used for the forced alignment were first aligned with the reference transcriptions as described in the section on probability features. The start and end time of each word in the reference transcriptions could then be determined from the timestamps in the forced alignment. For contractions listed as two reference words but one forced alignment word, any word-level prosodic features will be identical for both words. The prosodic features we extracted are as follows.

Pitch. The minimum, maximum, mean, and log range of pitch for each word were extracted using Praat (Boersma and Weenink, 2007). Minimum, maximum, and mean values were then normalized by subtracting the average of the mean pitch values for speakers of the appropriate sex, i.e., these feature values are relative to gender norms.[9] We used the log transform of the pitch range due to the highly skewed distribution of this feature; the transformation produced a symmetric distribution.

Intensity. The minimum, maximum, mean, and range of intensity for each word were extracted using Praat.

[9] Preliminary analysis revealed that the normalized pitch values show a more systematic relationship with error rates than the unnormalized values. In addition, normalizing by gender average removes the correlation between sex and pitch features, which is useful when building the combined model in Section 4.

ACCEPTED MANUSCRIPT

Speech rate. The average speech rate (in phones per second) was computed for each utterance and assigned to all words in that utterance. The number of phones was calculated by summing the word length feature of each word, and utterance duration was calculated using the start and end times of the first and last words in the utterance, as given by the forced alignment. We used the automatically generated utterance timestamps of the forced alignment because the hand-annotated utterance boundary timestamps in the reference transcriptions include variable amounts of silence and non-speech noises at the beginnings and endings of utterances and we found the forced alignment boundaries to be more accurate.

Duration. The duration of each word was extracted using Praat.

Log jitter. The jitter of each word was extracted using Praat. Like pitch range, the distribution of jitter values is highly skewed; taking the log transform yields a symmetric distribution.

Note that not all prosodic features could be extracted from all words in the data set. In what follows, we discuss results using three different subsets of our data: the full-word set (all 71579 full words in the data), the prosodic set (the 67302 full words with no missing feature values; 94.0% of the full-word data set), and the no-contractions set (the 60618-word subset of the prosodic set obtained by excluding all words that appear as contractions in the forced alignment transcriptions and as two separate words in the reference transcriptions; 84.7% of the full-word data set).

3.2 Results and discussion

Error rates for categorical features can be found in Table 4, and error rates for numeric features are illustrated in Figures 3 and 4. (First and middle repetitions are combined as non-final repetitions in the table, because only 92 words were middle repetitions, and their error rates were similar to initial repetitions.) Despite differences in the overall error rates between the two systems we examined, the patterns of errors display a striking degree of similarity. We discuss results for individual features in more detail below, after describing the methodology used to produce the figures and significance values shown.

The error rates shown in Table 4 are based on the full-word data set, with significance levels computed using a Monte Carlo permutation test. [10] For each feature, we generated 10,000 random permutations of the words in the data, and assigned the first n words in the permuted set to one group, and the remaining words to a second group (with n equal to the number of words exhibiting the given feature). The significance level of a given feature's effect on error rates can be estimated as the proportion of these samples for which the difference in IWER between the two groups is at least as large as the actual difference between words that do or

[10] The permutation test is a standard nonparametric test that can be used with data like ours that may not conform to any particular known distributional form (Good, 2004).

ACCEPTED MANUSCRIPT

	SRI system		Cambridge system		
Feature	IWER	MC test	IWER	MC test	% of data
Before FP	16.7	.1146	15.9	.4099	1.9
After FP	16.8	.2418	15.3	.1884	1.8
Before frag	32.2**	.0000	29.2**	.0000	1.4
After frag	22.0**	.0008	18.5	.0836	1.4
Before rep	19.6	.4508	17.0	.8666	0.7
After rep	15.3*	.0486	13.5*	.0222	0.9
Non-final rep	28.4**	.0000	28.6**	.0000	1.2
Final rep	12.8**	.0001	11.8**	.0003	1.1
Open class	17.3**	.0000	16.0**	.0000	50.3
Closed class	19.3**	.0000	17.2**	.0002	43.7
Discourse marker	18.1	.8393	18.2*	.0066	6.0
Starts turn	21.0**	.0000	19.5**	.0000	6.2
Male	19.8**	.0000	18.1**	.0000	49.6
Female	16.7**	.0000	15.3**	.0000	50.4
All words	18.2		17.0		100

Table 4
IWER by feature for the two systems on the full-word data set. *MC test* gives the proportion of samples (out of 10,000) in a Monte Carlo permutation test for which the difference between groups was at least as large as that observed between words with and without the given feature. Features with a significant effect on error rates according to the Monte Carlo test are indicated with * ($p < .05$) or ** ($p < .005$). *% of data* is the percentage of words in the data set having the given feature. *All words* is the IWER for the entire data set. (Overall IWER is slightly lower than in Table 1 due to the removal of OOV words.)

do not exhibit the given feature. Although not shown, we computed error rates and significance levels on the prosodic and no-contractions data sets as well. Overall error rates are somewhat lower for these data sets (SRI: 18.2 full, 17.5 prosodic, 17.4 no-contractions; Cambridge: 16.7 full, 16.0 prosodic, 15.8 no-contractions), but the pattern of errors is similar. For nearly all features, p-values for the smaller data sets are equal to or larger than those for the full data set; i.e., the smaller data sets provide a more conservative estimate of significance. Consequently, we feel justified in basing the remaining analyses in this paper on the smallest (no-contractions) data set, which provides the most accurate feature values for all words.

Figures 3 and 4 were produced using the no-contractions data set. Figure 3 includes the pronunciation-based and probability features, which (with the exception of trigram probability) are all *lexical*, in the sense that every instance of a particular

ACCEPTED MANUSCRIPT

lexical item takes on the same feature value. Figure 4 includes the prosodic features, which vary across different instances of each lexical item. To estimate the effects of each of these numeric features and determine whether they have significant predictive value for error rates, we used logistic regression (as implemented by the `lrm` package in R). In logistic regression, the *log odds* of a binary outcome variable is modeled 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 and $\beta_0 \ldots \beta_n$ are coefficients (feature weights) to be estimated. If IWER were a binary variable, we could estimate the effect of our features by building a separate regression model for each feature based on a single predictor variable – the value of that feature. However, IWER can take on values greater than 1, so we cannot use this method. Instead, we build a model that predicts the probability of an error (i.e., the probability that IWER is greater than zero). This model will be slightly different than a model that predicts IWER itself, but for our data sets, the difference should be minor: the number of words for which IWER is greater than one is very small (less than 1% of words in either system), so the difference between the average IWER and the probability of an error is minimal (SRI: average IWER = 17.4, P(error) = 17.4%; Cambridge: average IWER = 15.8, P(error) = 15.6%). These differences are negligible compared to the sizes of the effects of many of the features illustrated in Figures 3 and 4.

While many of our numeric features exhibit a primarily linear relationship with the log odds of an error, several appear to have more complex patterns. To allow for this possibility, we used restricted cubic splines to create smooth functions of the input features. [11] It is then these functions that are assumed to have a linear relationship with the log odds. We limited the possible functions to those with at most one inflection point (i.e., quadratic-like functions) and built a regression model for each feature to predict the probability of an error based on the value of that feature alone. The predicted values are plotted in Figures 3 and 4 on top of the observed IWER. (Note that, although feature values were binned in order to plot the average observed IWER for each bin, the regression models were built using the raw data.) For each feature, we determined whether that feature is a significant predictor of errors by performing a likelihood-ratio test comparing the model fitted using that feature as its sole predictor to a baseline model that simply fits the overall error probability in the data. All features were found to be significant predictors; the slopes of the fitted probabilities in Figures 3 and 4 give a sense of the relative importance of different features in predicting errors. [12]

[11] Restricted cubic splines were fit using the `rcs` function in the Design package (Harrell Jr., 2007) of R (R Development Core Team, 2007).

[12] Note that, when considering only *linear* relationships between feature values and the log odds of an error, the number of neighbors and mean intensity (for both systems) and

ACCEPTED MANUSCRIPT

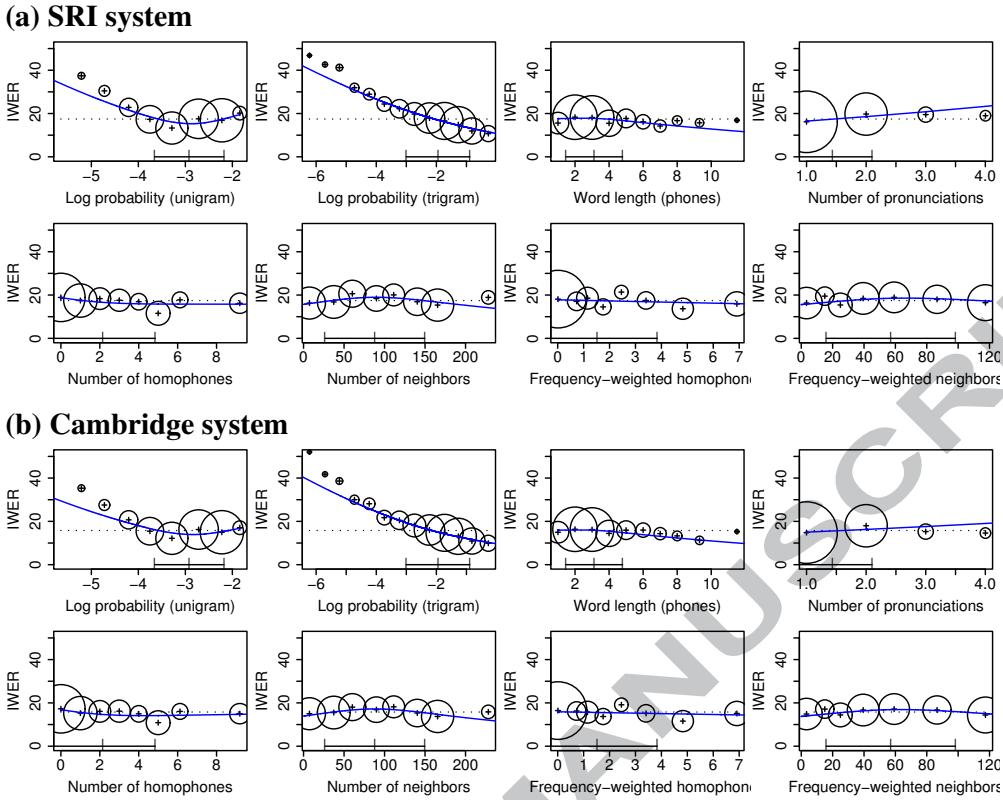


Fig. 3. Effects of lexical features and trigram probability on IWER for (a) the SRI system and (b) the Cambridge system on the no-contractions data set. All feature values were binned, and the IWER for each bin is plotted, with the area of the surrounding circle proportional to the number of points in the bin. The mean value and standard deviation of each feature is provided along the bottom of each plot. Dotted lines show the IWER over the entire data set. Solid lines show the predicted probability of an error using a logistic regression model fit to the data using the given feature as the only predictor (see text).

3.2.1 Disfluency features

Perhaps the most interesting result in Table 4 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 preceding fragments have an IWER 10.2–14 points *higher* than the average word (e.g., words preceding fragments in the SRI system have a 32.2% IWER, 14 points above the 18.2% average), while final repetitions and words following repetitions have an IWER 2.9–5.4 points *lower* (note, however, that the results for words after repetitions are less robust – they just barely reach the .05 significance level for the full-word SRI data set, and do not reach significance in some of the other data subsets). Words following fragments show a smaller increase in errors in the SRI data set, and a non-significant increase in the

frequency-weighted neighbors (for the Cambridge system) are not significant predictors of errors.

ACCEPTED MANUSCRIPT

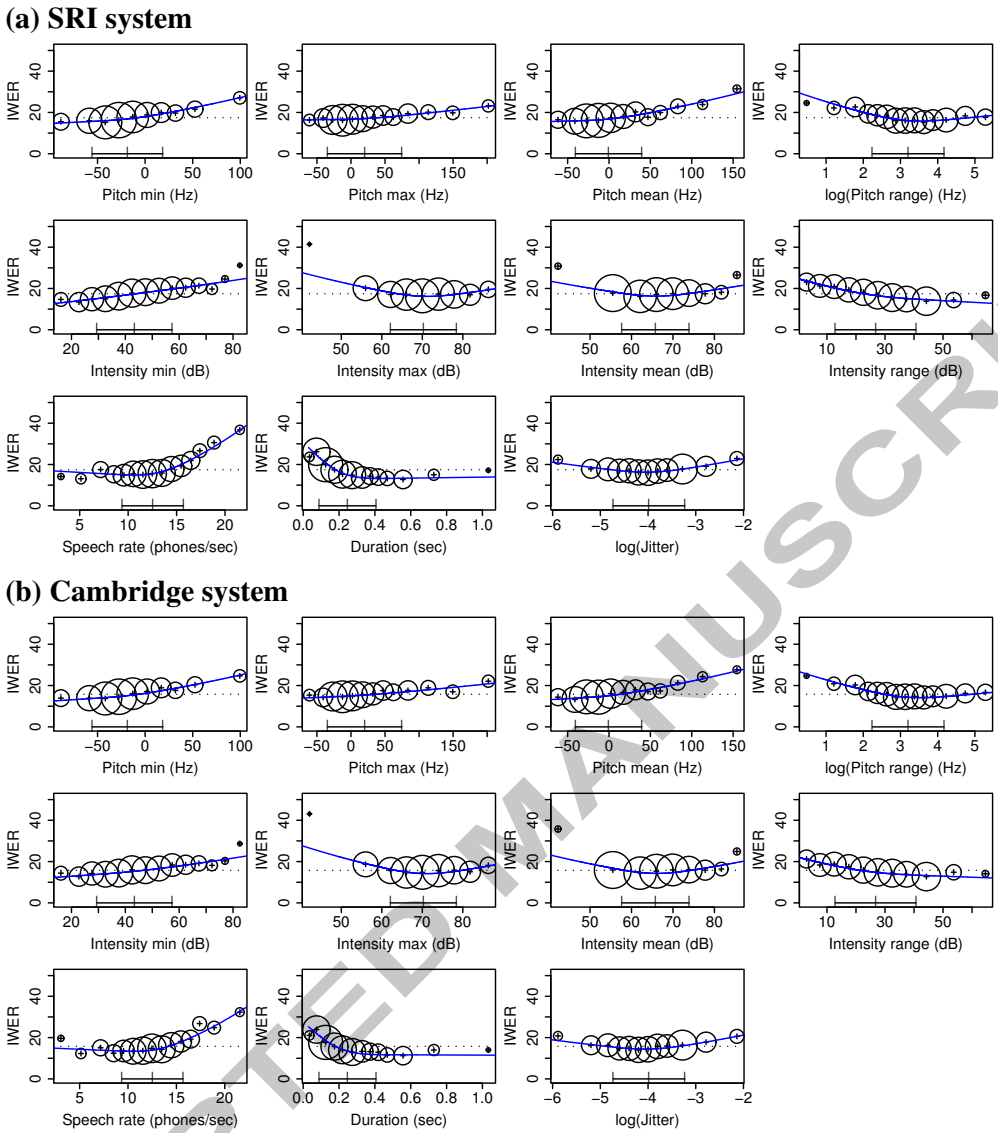


Fig. 4. Effects of prosodic features on IWER for (a) the SRI system and (b) the Cambridge system on the no-contractions data set. Details of the plots are as in Figure 3.

Cambridge data set. Words occurring before repetitions or next to filled pauses do not have significantly different error rates than words in other positions. Our results with respect to repetitions are consistent with the work of Shriberg (1995), which suggested that only non-final repetitions are disfluencies, while the final word of a repeated sequence is fluent.

ACCEPTED MANUSCRIPT

3.2.2 Other categorical features

Consistent with common wisdom, we find that open class words have lower error rates than other words and that words at the start of a turn have higher error rates. In addition, we find worse recognition for males than for females. Although some of these effects are small, they are all statistically robust and present in both systems. The difference in recognition accuracy of 2.8–3.1% between males and females is larger than the 2% found by Adda-Decker and Lamel (2005), although less than the 3.6% we found in our own preliminary work in this area (Goldwater et al., 2008), which analyzed only the SRI system and used a smaller data set.

3.2.3 Word probability

Turning to Figure 3, we find that low-probability words have dramatically higher error rates than high-probability words, consistent with several previous studies (Shinozaki and Furui, 2001; Fosler-Lussier and Morgan, 1999; Goldwater et al., 2008). Comparing the effects of unigram and trigram probabilities, we see that trigram probability shows a far more linear relationship with IWER. This is not surprising: words that have lower language model probabilities can be expected to have worse recognition. Unigram probabilitity, while correlated with trigram probability, is a less direct measure of the language model score, and therefore has a more complex relationship with error.

3.2.4 Pronunciation features

While all of the pronunciation features we examined do have a significant effect on error rates, the sizes of the effects are in most cases fairly small. Not surprisingly, words with more possible pronunciations have higher error rates than those with fewer, and longer words have slightly lower error rates than shorter words. The small size of the word length effect may be explained by the fact that word length is correlated with duration, but anti-correlated with probability. (Table 5 shows the correlations between various features in our model.) Longer words have longer duration, which tends to decrease errors (Figure 4), but also lower probability, which tends to increase errors (Figure 3).

In contrast to the primarily linear effects of length and number of pronunciations, we find that words with *intermediate* numbers of neighbors (or frequency-weighted neighbors) are the most difficult to recognize. This finding seems to contradict those of psycholinguistic studies, but it is important to remember that those studies controlled for word frequency and length, while the results in this section do not control for other variables. Also, the psycholinguistic results pertain to recognition of isolated words, while our results are based on recognition in context.

Finally, we see that words with more homophones (or frequency-weighted homo-

ACCEPTED MANUSCRIPT

phones) have significantly lower error rates than other words. Although the effect is very small, it is nevertheless surprising, and is not obviously due to correlations with other features we examined – the number of homophones is strongly correlated only with the other homophone and neighbor features (Table 5). There are moderate correlations with word duration and word length, but both of these are in the wrong direction, i.e., they would predict that words with more homophones have a greater chance of misrecognition because they tend to be shorter.

3.2.5 Prosodic features

In contrast to the pronunciation-based features, Figure 4 shows that most of the prosodic features we examined are strongly predictive of error rates. Decreased duration is associated with increased IWER, and (as in previous work), we find that IWER increases dramatically for fast speech. Mean pitch also has a large effect, with higher error rates for words with higher pitch relative to gender averages. Minimum and maximum pitch, which are highly correlated with mean pitch (Table 5), show similar trends, but to a slightly lesser degree. Words with smaller ranges of pitch or intensity are more likely to be misrecognized, as are words with higher minimum intensity (a feature that is highly anti-correlated with intensity range). The final three prosodic features – jitter and intensity maximum and mean – show little to no linear effect on errors. Instead, these features are associated with higher error rates at extreme values than at average values. The same pattern, but to a lesser degree, can be observed for several of the other prosodic features. This kind of pattern has been noted before by several authors in the case of speech rate (Shinozaki and Furui, 2001; Siegler and Stern, 1995; Goldwater et al., 2008), but was first discussed for other prosodic features only in the preliminary version of this work (Goldwater et al., 2008).

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 for each system's output with all of our features as potential predictors in order to determine the effects of each feature after accounting for possible correlations. We use the no-contractions data set, and simplify modeling (as above) by predicting only whether each token is responsible for an error or not. That is, we use a binary dependent variable for each token, which takes on the value 1 if the IWER for that token is greater than zero, and 0 otherwise.

ACCEPTED MANUSCRIPT

Feature pair	τ statistic
Duration, Min intensity	-0.31
Unigram prob, # neighbors	0.32
Duration, Log pitch range	0.33
Unigram prob, Trigram prob	0.33
# neighbors, Duration	-0.36
Trigram prob, Length	-0.37
Duration, Intensity range	0.40
Unigram prob, Freq-wtd neighbors	0.40
Length, # homophones	-0.42
Unigram prob, Duration	-0.42
Max pitch, Log pitch range	0.43
Freq-wtd neighbors, Duration	-0.44
Length, Freq-wtd homophones	-0.48
Unigram prob, Length	-0.48
Length, Duration	0.50
Max pitch, Min pitch	0.52
# homophones, Freq-wtd neighbors	0.52
Freq-wtd homophones, Freq-wtd neighbors	0.54
# neighbors, Freq-wtd homophones	0.56
Length, # neighbors	-0.61
Min intensity, Intensity range	-0.63
# homophones, # neighbors	0.64
Mean pitch, Min pitch	0.71
Length, Freq-wtd neighbors	-0.72
# homophones, Freq-wtd homophones	0.75
Mean pitch, Max pitch	0.77
# neighbors, Freq-wtd neighbors	0.78
Mean intensity, Max intensity	0.85

Table 5
Correlations between the numeric features examined here, measured using Kendall's τ statistic, a nonparametric method. Possible values of τ range from -1 (perfect anti-correlation) to 1 (perfect correlation). Only absolute values above 0.3 are shown.

ACCEPTED MANUSCRIPT

4.1 Model

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 such as speaker identity do not fit this pattern. Instead, we account 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. The predictors in our initial model are summarized in Table 6. They included all of the features described above, with the exception of number of neighbors, frequency-weighted homophones, pitch minimum and maximum, and intensity minimum and maximum. These features were excluded because of high correlations with other features in the model, which makes parameter estimation in the combined model difficult. All categorical features (those in Table 4) were converted to binary variables, and additional binary features were added to account for corpus (Fisher or Switchboard) and telephone line type (standard, cellular, cordless, land). All numeric features (those in Figures 3 and 4) were rescaled to values between 0 and 1 in order to make the model coefficients for different features comparable, [13] and then centered to ensure a mean value of 0.

To account for the possibility that some of the numeric features in our model have non-linear effects (as suggested by our analysis in Section 3), our initial model included functions of these features with at most one inflection point, again modeled using the restricted cubic splines (`rcs`) function in R. (The backwards elimination process can be used to eliminate the extra parameters associated with the non-linear components of each predictor as necessary.) In addition, we included random effects for speaker identity and word identity. Thus, the initial model includes 44 degrees of freedom: 43 for the features shown in Table 6, and one for the intercept.

[13] Before rescaling, 39 data points with outlying feature values were removed: two words with speech rate above 27 phones per second, 13 words with duration above 1.25 seconds, and 24 words with log jitter below -7.

ACCEPTED MANUSCRIPT

Feature	F/R	Type	d.f.	Feature	F/R	Type	d.f.
BEFORE-FP	F	B	1	UNIGRAM-PROB	F	N	2
AFTER-FP	F	B	1	TRIGRAM-PROB	F	N	2
BEFORE-FRAG	F	B	1	WORD-LENGTH	F	N	2
AFTER-FRAG	F	B	1	NUM-PRONUNCIATIONS	F	N	1
BEFORE-REP	F	B	1	NUM-HOMOPHONES	F	N	2
AFTER-REP	F	B	1	FREQ-WTD-NEIGHBORS	F	N	2
FINAL-REP	F	B	1	PITCH-MEAN	F	N	2
NONFINAL-REP	F	B	1	LOG-PITCH-RANGE	F	N	2
OPEN-CLASS	F	B	1	INTENSITY-MEAN	F	N	2
DISCOURSE-CLASS	F	B	1	INTENSITY-RANGE	F	N	2
STARTS-TURN	F	B	1	SPEECH-RATE	F	N	2
SEX	F	B	1	DURATION	F	N	2
CORPUS	F	B	1	JITTER	F	N	2
CELLULAR-LINE	F	B	1	SPEAKER-ID	R	C	1
LAND-LINE	F	B	1	WORD-ID	R	C	1
CORDLESS-LINE	F	B	1				

Table 6
Summary of features used in the unreduced joint model, showing whether each feature is a F(ixed) or R(andom) effect, whether it is B(inary), N(umeric), or C(ategorical), and the associated degrees of freedom (d.f.). Numeric features were fit using restricted cubic splines with two degrees of freedom, except for NUM-PRONUNCIATIONS, which does not take on enough different values to fit a non-linear spline.

4.2 Results and discussion

Figure 5 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.

The coefficients for our numeric features are not directly interpretable in most cases, since they are computed in terms of the orthogonal basis functions of the restricted cubic splines used to fit the non-linear components of the model. How-

ACCEPTED MANUSCRIPT

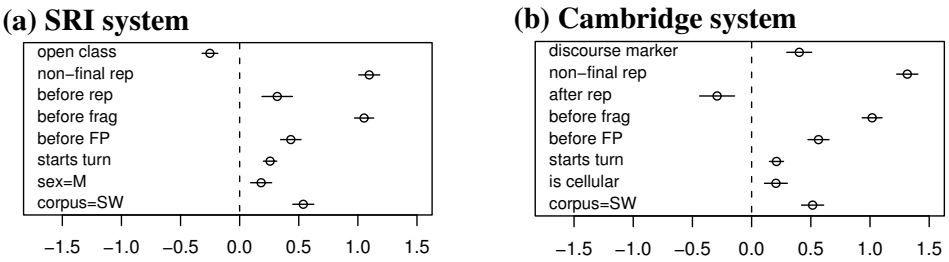


Fig. 5. Estimates and standard errors of the coefficients for the categorical features found to be significant predictors in the reduced model for each system.

ever, the coefficients can be used to plot the predicted effect of each feature on the log odds of an error, yielding the visualization in Figure 6. 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 and centered values of each feature, so that all x axes are one unit long, with the mean observed value of each feature always equal to zero.

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 before fragments have higher error rates, while final repetitions and words following repetitions have lower error rates. After accounting for correlations between factors, a slightly different picture emerges. Non-final repeated words and words before fragments still have the greatest chance of an error, but there is no longer a beneficial effect for final repetitions, and the effect for words after repetitions is only found in the Cambridge system. Both systems now show increased chances of error for words before filled pauses, and words before repetitions are also associated with more errors in the SRI system. Overall, disfluencies tend to have a negative effect on recognition, increasing the odds of an error by as much as a factor of 3.7.

Many of the differences in disfluency patterns from Section 3 (specifically, the reduction or elimination of the apparent beneficial effect of final repetitions and words following repetitions, and the appearance of a negative effect before filled pauses) may be explained as follows. Words near filled pauses and repetitions have longer duration than other words (Bell et al., 2003), and longer duration lowers IWER. Taking duration into account therefore reduces any apparent positive effects of disfluencies, and reveals previously obscured negative effects. Also, according to Shriberg (1995), the vast majority of repetitions are so-called "retrospective" repetitions (Heike, 1981), in which the initial word(s) are disfluent, but the final word resumes fluent speech. Our results fit nicely with this hypothesis, since final repetitions have no significant effect in our combined model, while non-final repetitions incur a penalty.

ACCEPTED MANUSCRIPT

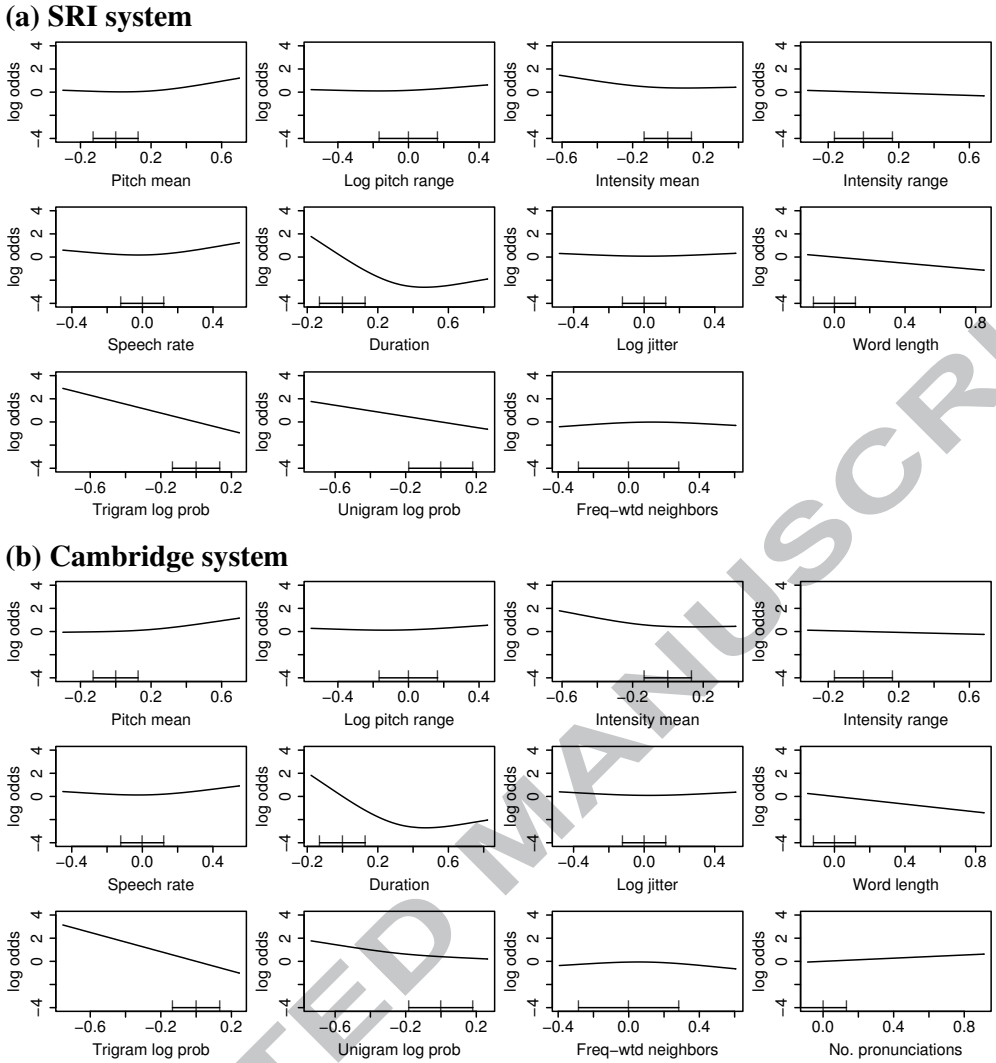


Fig. 6. Predicted effect of each numeric feature on the log odds of an error. Only features found to be significant predictors in the reduced model for each system are shown. The mean value and standard deviation of each feature (after rescaling and centering) is provided along the bottom of each plot. Due to rescaling, all x axes are one unit long; the range of values shown reflects the range of values observed in the data.

4.2.2 Other categorical features

Without including in the model 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, the start-of-turn effect still holds even after accounting for the effects of other features. This suggests that contextual (i.e., language modeling) factors are not the explanation for the start-of-turn effect; articulatory strengthening is a possible alternative (Fougeron and Keating, 1997; Keating et al., 2003). The open-class effect appears

ACCEPTED MANUSCRIPT

in our joint model for the SRI system, although in the Cambridge system open-class words do not seem to have a beneficial effect; instead, discourse markers are found to have a negative effect. As in the individual model, all of these effects are fairly small.

As for speaker sex, we find that male speakers no longer have significantly higher error rates than females in the Cambridge system. Males do have significantly higher error rates than females in the SRI system, but the difference is small (a factor of 1.2 in the odds), and the significance level is now only .04, as compared to below .0001 in the individual analysis. These results shed 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 incorporates the first two of these factors, and by doing so greatly reduces the difference in error rates between males and females. This suggests that disfluency and duration indeed account for much of the difference in recognition accuracy. The remaining small differences may be accounted for by males' increased use of alternate pronunciations, as suggested by Adda-Decker and Lamel (2005). Another 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 Word probability

Not surprisingly, we find that even after accounting for the effects of correlated features, word probability still has a very strong effect on recognition performance, with higher error rates for low-probability words. Interestingly, both unigram and trigram probabilities have strong independent effects, with the trigram language model probability being the more influential. It is also worth noting that the non-linear trends appearing in the individual analysis were not found to be significant in the combined model, except for a small but significant effect ($p < 0.025$) in the Cambridge unigram probability. Thus, our modeling results suggest that a word's frequency and its language model probability are both independently related to the chance of its being recognized correctly in a near linear way.

4.2.4 Pronunciation features

Our combined model considered four pronunciation-based features: word length, number of pronunciations, number of homophones, and frequency-weighted neighbors. Only two of these were found to be significant predictors in both systems: word length (with longer words having lower error rates) and frequency-weighted neighbors (with intermediate values having higher error rates). The effect of word

ACCEPTED MANUSCRIPT

length is greater than in the individual analysis, which supports our hypothesis that correlations with duration and probability obscured the word length effect in that analysis. We have no explanation at this time for the non-linear effect of frequency-weighted neighbors, which persists despite our model's incorporation of other factors such as word frequency and length.

Number of pronunciations was found to be a significant predictor only in the Cambridge system, where words with more pronunciations had higher error rates.

4.2.5 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 now see that means of pitch and intensity are actually stable over a wide range of the most common values, but errors increase for extreme values of pitch (on the high end) and intensity (on the low end). A greater range of intensity is still associated with lower error rates despite accounting for the effects of duration, which one might expect to be the cause of this trend in the individual analysis. However, the benefit of greater pitch range is no longer seen; instead, extreme values of pitch range on either end seem to be problematic (although the effect is small).

In the individual analysis, both speech rate and duration were strongly tied to error rates. While both of these features are still important in the combined model, duration is shown to have by far the greater effect of the two. Unlike most of the other prosodic features we examined, including speech rate, average values of duration do not have the lowest error rates. Rather, above-average duration is associated with the lowest error rates. For words with extremely long duration, recognition begins to degrade again. Note that, although one might expect speech rate and duration to be highly correlated, we found a relatively low correlation of $\tau = -0.15$. Only a small part of the variability in duration can be explained by speech rate; the rest is due to variations in word length and other factors.

For our final prosodic feature, log jitter, we found in the individual analysis that extreme values were associated with higher error rates. This finding was replicated in the combined model.

Overall, the results from our joint analysis suggest that, other things being equal, recognition performance is best for words with typical values of most prosodic features (duration being the notable exception), and degrades as feature values become more extreme.

ACCEPTED MANUSCRIPT

	SRI			Cambridge		
Model	Neg. log lik.	Diff.	df	Neg. log lik.	Diff.	d.f.
Full	24305	0	44	22644	0	44
Reduced	24316	11	27	22651	7	29
Baseline	28006	3701	1	26195	3551	1
No categorical	24475	159	32	22836	185	32
No probability	24981	664	40	23367	716	40
No pronunciation	24347	31	37	22689	38	37
No prosodic	25150	834	30	23449	797	30
No speaker	25069	753	43	23379	727	43
No word	24627	322	43	22901	257	43

Table 7
Fit to the data of various models and their degrees of freedom (d.f.). *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*: all categorical features (disfluencies, sex, syntactic class, start-of-turn), all probability features (unigram and trigram probabilities), all pronunciation features (length, number of homophones, frequency-weighted neighbors, number of pronunciations), all prosodic features (pitch, intensity, rate, duration, jitter), the random effect for speaker identity, or the random effect for word identity. *Diff* is the difference in log likelihood between each model and *Full*. Fits are significantly different for all pairwise comparisons except *Full* vs. *Reduced*, as computed using a likelihood ratio test.

4.2.6 Differences between lexical items

As discussed above, our models contain a random effect for word identity, to account 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 each system's data of two different models: our initial full model containing all of our fixed effects (including both linear and non-linear terms) and random effects for both speaker identity and word identity, and a similar model containing all the same features except for word identity. Results are shown in Table 7. For both systems, the fit of the model without a random effect for word identity is significantly worse than that of the full model; in fact, this single parameter is more important than all of the categorical and pronunciation features combined.

In mixed-effects models, it is possible to extract estimates of the by-word adjustments to the model predictions, that is, the amount by which each lexical item's odds of an error deviates from the mean. Figure 7 lists, for each system, the 30 words with the greatest negative deviation from the mean. As we might expect given the similarities between the two systems in our other results, the two lists

ACCEPTED MANUSCRIPT

(a) SRI: *yup*, *yep*, *a.*, *halloween*, *phones*, *into*, *half*, *though*, *then*, *after*, *wanted*, *watched*, *whether*, *happened*, *them*, *says*, *than*, *worked*, *started*, *something*, *foreign*, *theater*, *island*, *r.*, *room*, *tastes*, *space*, *salad*, *called*, *goes*.

(b) Cambridge: *yup*, *yep*, *something*, *phones*, *him*, *after*, *then*, *though*, *ask*, *couple*, *wanted*, *half*, *into*, *tried*, *faith*, *than*, *whether*, *them*, *space*, *happened*, *watched*, *already*, *worked*, *four*, *thinking*, *stay*, *god*, *thanks*, *yes*, *probably*.

Fig. 7. The 30 lexical items for each system with the greatest estimated negative effect on the probability of correct recognition.

contain many words in common. In fact, the correlation between the estimated effect of each lexical item in the two different systems (over all 3867 lexical items) is fairly high: r=0.69 linear correlation.

Some of these errors are clearly due to inconsistencies in the reference transcriptions that are not covered by the normalization rules used in the NIST evaluation. The two words with the highest estimated error in both systems, *yup* and *yep*, are orthographic variants of the same word. Similarly, the most frequent misrecognition of *theater* is a substitution by *theatre* (for the (American) SRI system as well as the (British) CU-HTK system). Both systems frequently substituted hypothesis *yeah* for reference *yes*; this is likely another problem with inconsistent transcription in the reference.

Many of the other high-error words involve morphological substitutions, particularly between the bare stem and the past tense forms. The language model is often insufficient to distinguish these two forms, since they can occur with similar neighboring words (e.g., *they watch them* and *they watched them* are both grammatical and sensible), and they are also similar acoustically. Examples of this kind of error, with their most frequent substitution in parentheses, include the following reference words: *called* (*call*), *asked* (*ask*), *asks* (*asked*), *happened* (*happen*), *says* (*said*), *started* (*start*), *thinking* (*think*), *tried* (*try*), *wanted* (*want*), *watched* (*watch*), *tastes* (*taste*), *phones* (*phone*), and *goes* (*go*).

In addition to these morphological substitutions, several other high-error words are also frequently substituted with homophones or near-homophones that can occur in similar contexts, in particular *than* (*and*), *then* (*and*), *him* (*them*), and *them* (*him*). The high error rates found for these words may explain why we did not find strong effects for neighborhood density overall. In most situations, the context in which a word is spoken is sufficient to disambiguate between acoustically similar candidates, so competition from phonetically neighboring words is not usually a problem. Errors in ASR are caused not by words with large numbers of similar neighbors, but by words with one or two strong competitors that can occur in similar contexts. Put another way, acoustically confusable words are not typically a problem, but *doubly confusable pairs* — word pairs with similar language model scores in addition to similar acoustic scores — can be a source of errors. This finding also

ACCEPTED MANUSCRIPT

suggests that the effects of neighborhood density in human word recognition might also be significantly reduced when words are recognized in context rather than in isolation, as is typical in experimental settings.

Finally, we note that the words in the previous paragraph (*than*, *then*, *him*, and *them*) are very frequently deleted as well as being substituted. This is probably due to a combination of lack of disambiguating context (e.g., *would want him to be* and *would want to be* are both acceptable, *and then* and *and* mean essentially the same thing) and the fact that these words are subject to massively reduced pronunciations, often due to cliticization (*him* and *them* are generally cliticized to the preceding verb; *then* is often cliticized to the following word). Other words with high error that are known to be subject to massive reduction include *something*, *already*, and *probably* suggesting that all these examples may be due to pronunciation factors beyond those captured by simple duration.

4.2.7 Differences between speakers

As we have already mentioned, ASR error rates are known to differ greatly between speakers. Using the mixed-effects logistic regression methodology presented here, it is possible to examine the extent to which these differences can be explained by variation in speakers' patterns of lexical choice, prosody, or disfluency. We first used the same method described above to analyze the overall importance of the random effect for speaker identity in our fitted models. As shown in Table 7, removing the random effect for speaker identity from the full models results in a much worse fit to the data. That is, the lexical, prosodic, and disfluency 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 models for both the SRI and Cambridge data sets, and is more important than any other feature group aside from the prosodic features. Note that, as with the other features we analyzed, the error rates of different speakers are similar in the two data sets, with a linear correlation of .92. Figure 8 illustrates. Thus, whichever factors are responsible for the speaker-specific differences in recognition, they seem to behave similarly with respect to both systems, unlike some of the speaker differences found in the work of Nusbaum and Pisoni (1987) with much earlier systems.

In addition to analyzing the overall importance of speaker identity in our models, we can also examine the estimated effects of individual speakers, just as we did for individual words in the previous section. Figure 9 shows the by-speaker adjustments to the model predictions, with the actual probability of error for each speaker given for reference. Notice that the estimated adjustments do not completely track speaker error rates, as they would if speaker identity were the only factor in the model. This indicates that the other factors in our model do explain some of the variation in error rates between speakers, just not all.

ACCEPTED MANUSCRIPT

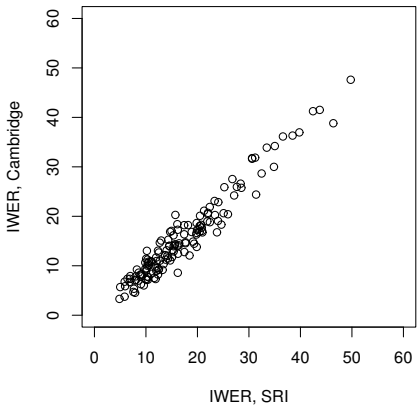


Fig. 8. A comparison of speaker error rates in the two systems. Each point represents a single speaker.

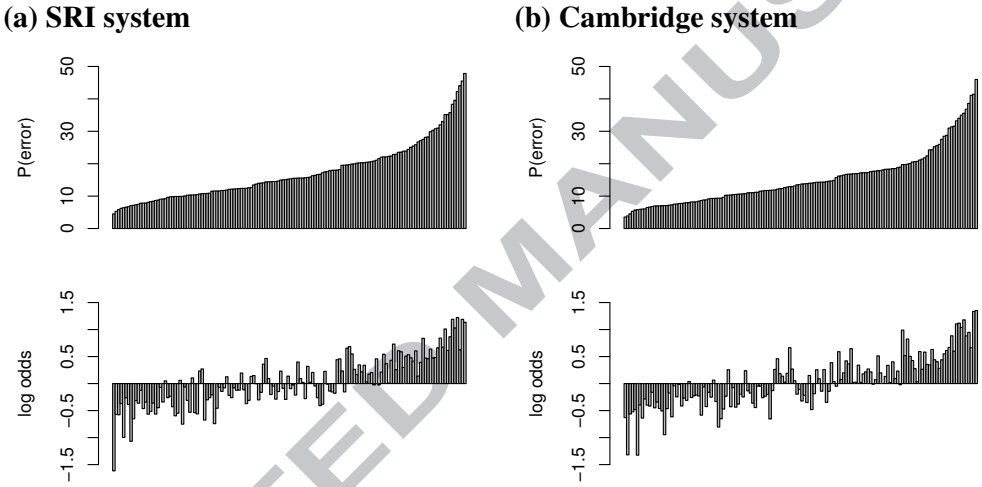


Fig. 9. Empirical probability of a recognition error for each speaker (top) and estimated change in the log odds of an error for each speaker (bottom). Each bar corresponds to a single speaker, with both graphs for a single system sorted according to the speakers' error probability under that system.

At this point, it seems natural to ask whether different speakers might not only have different overall error rates, but different patterns of errors – that is, does changing the values of certain features affect error rates differently for different speakers? The models presented here assume that each speaker has a different baseline error rate, but that the effects of each feature are the same for each speaker. Using techniques similar to those used here, it would theoretically be possible to introduce additional random effects for the intercepts (or even slopes) of each feature on a speaker-by-speaker basis, and to test for the significance of these additional parameters. However, the number of possible models for comparison would be enormous, so a purely exploratory analysis (similar to our own) is infeasible at present. To our knowledge, there are currently no automated model selection tools for mixed-

ACCEPTED MANUSCRIPT

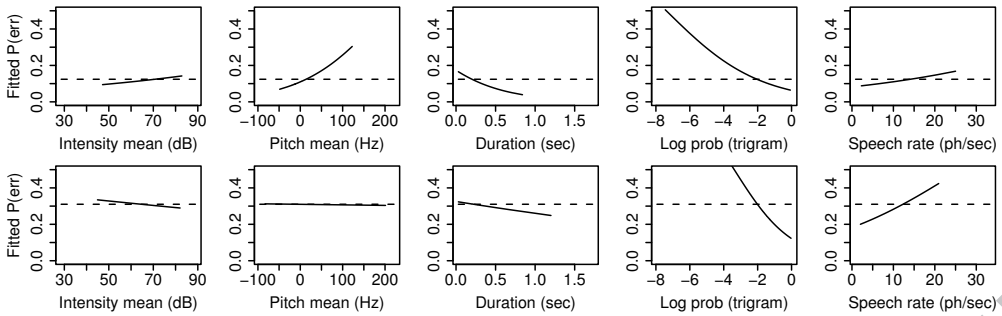


Fig. 10. Estimated effects of various features on the error rates of two different speakers (top: speaker fsh_60682_b, bottom: speaker sw_47282_b) using the SRI system. Dashed lines illustrate the baseline probability of error for each speaker. Solid lines were obtained by fitting a logistic regression model to each speaker's data, with the variable labeled on the x-axis as the only predictor.

effects models with multiple random effects, so analysis involves a human in the loop. Moreover, the complexity of our current models already pushes the boundary of what can be done with reasonable time and accuracy using the numerical optimization procedures that are available to fit the models.[14] Nevertheless, it is possible to get some sense of the variability between speakers by fitting separate logistic regression models to each speaker's data and plotting the results. Figure 10 illustrates some of the differences between two speakers chosen fairly arbitrarily from our data set, showing that the estimated effects of various features are quite different for the two speakers. For example, the estimated error rate increases dramatically for speaker fsh_60682_b as mean pitch increases, while speaker sw_47282_b shows almost no effect of pitch. Similar kinds of variability appear in the rest of the data set in both systems and for many of the features we examined. Although we do not know whether these differences are statistically significant, they are certainly suggestive that the effects of many features may vary considerably between speakers.

Since our models assume that features behave similarly across speakers, this suggestion might cause some readers to question the validity of our analysis and conclusions. However, we emphasize that the trends we have found are still an accurate reflection of the average behavior of the systems across a number of speakers. Statistical analyses of complex data sets have always been limited by the available technology, and we can only hope to incrementally improve our understanding of the phenomena in question by making use of the best tools available at the time. The mixed-effects models used here are a step up from previous work in which speaker differences were not modeled at all, and even correlations between features were often ignored. As new statistical modeling tools become available, we may be able to further refine our understanding of speaker differences. However, it is already

[14] Our largest model takes about an hour to fit on a 2.66 GHz workstation. Introducing more random effects would increase this time significantly, and could create problems with the accuracy of the final model fit as well.

ACCEPTED MANUSCRIPT

clear that, despite the speaker adaptation models used in the systems we analyzed, speaker differences remain an important source of error in ASR and an important challenge for future research.

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 a large variety of lexical, disfluency, contextual, and prosodic features in two different ASR systems, both individually and in a joint model. We found that despite differences in the overall performance of the two systems, the effects of the factors we examined were extremely similar. In particular, 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. However, only the former effect persists in both systems after accounting for the effects of other factors. (2) Disfluencies heavily impact error rates: IWER for non-final repetitions and words preceding fragments rises by up to 14% absolute, while IWER for final repetitions and words following repetitions decreases by up to 5.4% absolute. After accounting for the effect of prosodic features, the latter benefit is nearly eliminated, and a negative effect for words before filled pauses is revealed, suggesting that the effects of these disfluencies are normally obscured by the greater duration of nearby words. (3) For most acoustic-prosodic features (including pitch mean and range, intensity mean, jitter, and speech rate) there is little effect on recognition performance over a range of typical values, but errors increase for words with more extreme values in one or both directions (other factors being equal). The exception is duration, for which higher-than-average values yield the best performance. (4) After accounting for the effects of other factors, both unigram and trigram probability have strong independent effects on error rates, with the odds of an error increasing nearly linearly as probability decreases. (5) The probability of misrecognizing a word is only very weakly correlated with the number of neighbors of that word (similar-sounding words), and is uncorrelated with the number of homophones (identical-sounding words). However, these factors seem to be more important when contextual cues (language model probabilities) are insufficient to disambiguate similar-sounding words. (6) Although the factors we examined can account for some of the variance in error rates between speakers, unexplained differences between speakers are still a major factor in determining word error rates.

Our results have a number of implications for automatic speech recognition. The first concerns the role of disfluencies. About 15% of the words in our conversational telephone speech corpora are either disfluencies or adjacent to disfluencies, underscoring the importance of understanding how disfluencies contribute to error rates. We find that in fact, only some types of disfluencies are problematic — specifi-

ACCEPTED MANUSCRIPT

cally, fragments, non-final repetitions, and words preceding fragments. These kinds of words constitute about 4% of our corpora, but nevertheless cause a significant number of errors due to their hugely inflated error rates. Taken together, these results highlight the importance of continued research on disfluencies for decreasing recognition error rates in spontaneous speech, and also provide a guide as to which types of disfluencies might be more profitable to study.

Similarly, the fact that extreme prosodic values led to more errors, as well as the large individual differences we found, suggests that our current systems are not doing a good job of adapting to prosodic variation within and among speakers. Current algorithms for speaker-adaptive training such as MLLR and VTLN, focused as they are on cepstral values, are capable only of adjusting for speaker differences in segmental (phone) realization. While prosodic factors in speech processing have traditionally been studied in the context of speech synthesis rather than speech recognition, augmenting speaker-adaptive training to deal with prosodic variation may require explicit representation of prosodic aspects of the speech signal in recognition.

Our results suggest not only that speaker variation is an important remaining source of errors, but also provide at least a first step toward refining the search for the possible locus of this variation. Even after accounting for prosodic factors like pitch, intensity, and rate of speech, as well as language model, pronunciation, and disfluency factors, we found speaker differences to have a large role in determining error rates. This shows that none of the other basic factors we examined is the crucial source of speaker differences affecting recognition errors. Better understanding of speaker differences must remain a major direction for future research.

Finally, our analysis of the random effect for word identity suggests a new important factor that increases error when a word is a member of a *doubly confusable pair*: a pair of similar-sounding words that can also occur in very similar contexts. Such pairs include morphological substitutions between bare stem and past tense (preterite) forms like *ask/asked, says/said*, *watch/watched* and *want/wanted*, or pairs that are homophones when reduced, like *than/and* and *him/them* . Because examples like *they ask him* and *they asked him* are acoustically similar and similarly grammatical, neither the acoustic model nor the language model has sufficient evidence to distinguish them.

One way to improve discrimination of these pairs in recognition might be to build sets of binary classifiers that are designed to disambiguate exactly these situations when run as a rescoring procedure on lattice or confusion network output. For example, a *him/them* classifier might be able to make use of sophisticated natural language features, such as coreferring singular or plural nouns that occur elsewhere in the discourse. A stem/preterite classifier could make use of adverbial or other hints about tense or aspect in the clause.

ACCEPTED MANUSCRIPT

Our results on these lexical effects may also have implications for the study of human speech recognition. As we mentioned above, one of the most replicated studies in human speech recognition shows that humans subjects have difficulty recognizing similar-sounding words (Luce and Pisoni, 1998; Vitevitch and Luce, 1999). This result has been modeled by proposing that recognition of a target word is affected by the number and frequency of similar-sounding words, the frequency-weighted neighborhood density. But previous studies of human word recognition generally consist of isolated words. Our findings suggest the possibility that such difficulties may disappear when words are presented in context. In most situations, the context in which a word is spoken is sufficient to disambiguate between acoustically similar candidates, and indeed we saw that for ASR, competition from phonetically neighboring words is not usually a problem. Instead, we suggest that difficulties in human word recognition in context are caused not by words with large numbers of similar neighbors, but by doubly confusable pairs, i.e., homophones or neighbors with similar contextual predictability. The fact that our error analysis of automatic speech recognition helped us to develop a hypotheses about human speech recognition highlights the importance of the joint study of human and machine language processing.

Acknowledgments

This work was supported by the Edinburgh-Stanford LINK and ONR MURI award N000140510388. We thank Andreas Stolcke for providing the SRI recognizer output, language model, and forced alignments; Phil Woodland for providing the Cambridge recognizer output and other evaluation data; and Katrin Kirchhoff and Raghunandan Kumaran for datasets used in preliminary work, useful scripts, and additional help.

References

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

Baayen, R. H., 2008. Analyzing Linguistic Data. A Practical Introduction to Statistics. Cambridge University Press, Cambridge, UK.

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

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., Gildea, D., 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.

ACCEPTED MANUSCRIPT

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

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

Bulyko, I., Ostendorf, M., Stolcke, A., 2003. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In: Proceedings of the conference on Human Language Technologies.

Dahan, D., Magnuson, J., Tanenhaus, M., 2001. Time course of frequency effects in spoken-word recognition: Evidence from eye movements. Cognitive Psychology 42, 317–367.

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

Doddington, G., Schalk, T., 1981. Speech recognition: Turning theory to practice. IEEE Spectrum 18, 26–32.

Evermann, G., Chan, H. Y., Gales, M. J. F., Hain, T., Liu, X., Wang, L., Mrva, D., Woodland, P. C., 2004a. Development of the 2003 CU-HTK conversational telephone speech transcription system. In: Proceedings of ICASSP.

Evermann, G., Chan, H. Y., Gales, M. J. F., Jia, B., Liu, X., Mrva, D., Sim, K. C., Wang, L., Woodland, P. C., Yu, K., 2004b. Development of the 2004 CU-HTK English CTS systems using more than two thousand hours of data. In: Proceedings of the Fall 2004 Rich Transcription Workshop (RT-04f).

Evermann, G., Chan, H. Y., Gales, M. J. F., Jia, B., Mrva, D., Woodland, P. C., Yu, K., 2005. Training LVCSR systems on thousands of hours of data. In: Proceedings of ICASSP. Philadelphia, PA, pp. 209–212.

Fiscus, J., Garofolo, J., Le, A., Martin, A., Pallett, D., Przybocki, M., Sanders, G., 2004. Results of the fall 2004 stt and mde evaluation. In: RT-04F Workshop.

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

Fougeron, C., Keating, P., 1997. Articulatory strengthening at edges of prosodic domains. Journal of the Acoustical Society of America 101 (6), 3728–3740.

Goldwater, S., Jurafsky, D., Manning, C., 2008. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase ASR error rates. In: Proceedings of the Association for Computational Linguistics.

Good, P. I., 2004. Permutation, Parametric, and Bootstrap Tests of Hypotheses, 3rd Edition. Springer.

Hain, T., Woodland, P. C., Evermann, G., Gales, M. J. F., Liu, X., Moore, G. L., Povey, D., Wang, L., 2005. Automatic transcription of conversational telephone speech. IEEE Transactions on Speech and Audio Processing 13 (6), 1173–1185.

Harrell Jr., F., 2007. Design Package. R package version 2.1-1.

Heike, A., 1981. A content-processing view of hesitation phenomena. Language and Speech 24 (2), 147–160.

Hirschberg, J., Litman, D., Swerts, M., 2004. Prosodic and other cues to speech

ACCEPTED MANUSCRIPT

recognition failures. Speech Communication 43, 155– 175.

Howes, D., 1954. On the interpretation of word frequency as a variable affecting speech recognition. Journal of Experimental Psychology 48, 106–112.

Ingle, J., Wright, R., Wassink, A., 2005. Pacific northwest vowels: A Seattle neighborhood dialect study. The Journal of the Acoustical Society of America 117 (4), 2459–2459.

Keating, P., Cho, T., Fougeron, C., Hsu, C., 2003. Domain-initial articulatory strengthening in four languages. In: Local, J., Ogden, R., Temple, R. (Eds.), Phonetic Interpretation (Papers in Laboratory Phonology 6). Cambridge University Press, pp. 143–161.

Luce, P., Goldinger, S., Auer, E., Vitevitch, M., 2000. Phonetic priming, neighborhood activation, and PARSYN. Perception and Psychophysics 62 (3), 615–625.

Luce, P., Pisoni, D., 1998. Recognizing spoken words: the Neighborhood Activation Model. Ear and Hearing 19, 1–36.

Marcus, M., Santorini, B., Marcinkiewicz, M. A., Taylor, A., 1999. Treebank-3. Linguistic Data Consortium (LDC), catalog #LDC99T42.

Marslen-Wilson, W., 1987. Functional parallelism in spoken word-recognition. Cognition 25, 71–102.

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

Nusbaum, H., DeGroot, J., Lee, L., 1995. Using speech recognition systems: Issues in cognitive engineering. In: Syrdal, A., Bennett, R., Greenspan, S. (Eds.), Applied Speech Technology. CRC Press, Ch. 4.

Nusbaum, H., Pisoni, D., 1987. Automatic measurement of speech recognition performance: A comparison of six speaker-dependent recognition devices. Computer Speech and Language 2, 87–108.

Pennock-Speck, B., 2005. The changing voice of women. In: de Leonardo et al., J. C. G. (Ed.), Actas del XXVIII Congreso Internacional de AEDEAN. Servei de Publicacions de la Universitat de Valencia, pp. 407–415.

Povey, D., Woodland, P., 2002. Minimum phone error and I-smoothing for improved discriminative training. In: Proceedings of the IEEE ICASSP.

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. Available from http://www.R-project.org.

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

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

Shriberg, E., 1995. Acoustic properties of disfluent repetitions. In: Proceedings of the International Congress of Phonetic Sciences. Vol. 4. pp. 384–387.

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

Stolcke, A., Chen, B., Franco, H., Gadde, V. R. R., Graciarena, M., Hwang, M.-Y.,

ACCEPTED MANUSCRIPT

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

Vergyri, D., Stolcke, A., Gadde, V. R. R., Ferrer, L., , Shriberg, E., 2002. Prosodic knowledge sources for automatic speech recognition. In: Proceedings of the IEEE ICASSP. Vol. 1. pp. 208–211.

Vitevitch, M., Luce, P., 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. Journal of Memory and Language 40, 374–408.

Wang, W., Harper, M., 2002. The SuperARV language model: Investigating the effectiveness of tightly integrating multiple knowledge sources. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP. pp. 238–247.

ACCEPTED MANUSCRIPT