Experiments on context awareness and phone error propagation in human and machine speech recognition

Amit Juneja[1]*, Mark Hasegawa-Johnson*[2]

[1]Think A Move Ltd., Beachwood OH, USA
[2]Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, USA
amjuneja@gmail.com, jhasegaw@gmail.com

Abstract

A comparison of human speech recognition (HSR) and automatic speech recognition (ASR) is presented using a noisy continuous-speech corpus of null grammar or uniformly distributed unigram sentences, focusing on the differential tendency of machines vs. humans to propagate errors from an unclear phone to its neighbors. It is shown using controlled experiments that when given the same context for recognition - in this case a vocabulary of a limited number of known words - ASR makes as much as an order of magnitude more errors than HSR. The study provides evidence to contradict the claim made in recent literature that narrowing down the context of conversation and modeling of exceptional ordering of words is vital in achieving human-like accuracy by ASR. Using Chebyshev confidence intervals it is shown that ASR, but not HSR, propagates a phone recognition error from the phone to its neighbors at a rate significantly higher than chance.

Index Terms: automatic speech recognition, human speech recognition, null grammar, phone errors

1. Introduction

In this paper experiments in human speech recognition (HSR) and automatic speech recognition (ASR) are presented using a medium vocabulary continuous speech database containing recordings of null grammar sentences. A null grammar is defined as a uniformly distributed unigram grammar where any word can follow any other word with the same probability. A null grammar corpus provides the advantage that the bottom-up word recognition accuracy can be measured for both HSR and ASR with no contribution from a real-life language model. Word error rates and phone error rates of ASR can thus be compared HSR to assess the accuracy of just the acoustic modeling and the pronunciation modeling modules of ASR. It should be noted that ASR systems can be assessed with null grammar on any speech corpus by switching the language model. But the accuracy on grammatically sound sentences cannot be compared to HSR since humans cannot be asked to ignore grammar while transcribing such a corpus. Limited null grammar comparisons of ASR and HSR using different data for the two conditions were done by Lippmann [1] where HSR was only tested with monosyllabic words and there were no comparisons in noisy null grammar speech.

A more important aspect of ASR and HSR comparison in this work is that HSR was constrained to recognize from a limited vocabulary of words just like ASR as described in Section 2.1. In effect, HSR and ASR were provided exactly the same language model: uniform distribution over a finite vocabulary. Not only was the recognition context of HSR and ASR the same, the identically constrained recognition also meant that there were no exceptions in the occurrence of words in the speech sentences. It has been recently suggested [2] that two of the reasons for the superiority of HSR over ASR are that humans can narrow down the context on the current conversation or scenario and that humans are particularly good at catching exceptions so that if some unexpected words are spoken humans can still recognize them with good accuracy. The current experiments thus provide a control experiment for this hypothesis by comparing accuracy in identical contexts and with no exceptions. A wide gap in performance in ASR and HSR in this setup would mean that a major deficiency in ASR is in acoustic and/or pronunciation modeling.

The null grammar corpus also allows us to explore the relative contributions of the acoustic model and pronunciation model, as follows. ASR tends to propagate errors: given true word sequence $[w_1, w_2, \ldots]$ and recognized word sequence $[\hat{w}_1, \hat{w}_2, \ldots]$, the bigram error probability is higher than the unigram error probability: $p(\hat{w}_n \neq w_n | \hat{w}_{n-1} \neq w_{n-1}) > p(\hat{w}_n \neq w_n)$ (e.g., [3]). We do not know of any similar published results concerning phone sequences, and we do not know of any similar results for HSR. There is some reason to believe that ASR and HSR may have different patterns of error propagation. Only about half of all word tokens in spontaneous speech are produced with the canonical dictionary pronunciation [4], but in most cases, if one tries to represent the full range of pronunciation variability in an ASR dictionary, the increased entropy of the pronunciation model

causes increased word error rate [5]. We propose the following hypotheses. (H_1): ASR acoustic models are less accurate than HSR acoustic phone detection and recognition processes. (H_2): ASR compensates, in part, by using a pronunciation model with lower entropy, i.e., a dictionary that fails to represent typical human knowledge about pronunciation variability and/or about the apparent phone changes caused by noise masking. (H_3): Perhaps because H_1 and H_2 are true, or perhaps for other reasons, ASR propagates errors at a higher rate than HSR. Hypothesis H_2 is true by design: this paper uses the CMU speech recognition dictionary, which contains, on average, fewer than two pronunciation variants per word.. Hypothesis H_1 has been tested previously [6, 7, 8], and can be confirmed by measuring phone error rates on the null-grammar corpus. Hypothesis H_3 has not been previously tested (to our knowledge), and is considered one of the main contributions of this paper.

2. Method

2.1. Database Design

The NullG-Eval0 Speech Corpus [11] was used for evaluations and the TIMIT [12] and the WSJ corpora [13] were used for training the ASR system. The NullG-Eval0 corpus for human and machine speech recognition consists of (1) text prompts containing nonsense sentences, (2) continuous speech recordings from the text prompts, and (3) human transcriptions of those nonsense speech recordings. The corpus contains 1440 recorded utterances from 9 speakers (6 males and 3 females) such that there are 160 recordings from each speaker. Of those 160 sentences there are forty sentences each with words picked randomly from the following vocabulary sizes - 1000, 2000, 4000 and 8000. An example of a nonsense sentence from the speech corpus is "finally especially dad learn I hand", that is, the randomly picked words are spoken in a continuous manner similar to a proper sentence in English. Since there were only 40 sentences per vocabulary size not all the words in each vocabulary were spoken by each speaker but the words were randomly picked from those vocabularies. The vocabulary size is the same as the perplexity for this data because of the uniform distribution of the unigrams. For each of the nine speakers and each of the four vocabularies the 40 sentences are split such that there are 10 sentences each with SNRs = Quiet, 20dB, 10dB and 0dB (added white Gaussian noise).

2.1.1. Human transcription setup

To obtain human transcriptions, recordings from each speaker were transcribed independently by two normal-hearing human listeners [9]. Unlike machines, humans cannot remember an exact vocabulary of thousands of words. To work around this shortcoming the human transcribers had access to the vocabulary in a graphical user interface while transcribing each audio recording. If they entered an out-of-vocabulary (OOV) word the user interface presented them with suggestions of words in the vocabulary that are close to the OOV word in either pronunciation or spelling. The transcribers first transcribed the 1000-word vocabulary sentences which in turn were arranged in sets of ten sentences at each SNR - Quiet, 20dB, 10dB and 0dB - in that order. The process was then repeated for 2000, 4000 and 8000 word vocabulary sentences.

2.2. ASR Architecture

A speech recognition system based on cross-word context dependent triphones and mel-frequency cepstral coefficients (MFCCs) was built using Hidden Markov Model Toolkit (HTK) [10]. Cepstral mean subtraction and unsupervised adaptation were applied for noise robustness. The triphone models were trained using three read-speech corpora - TIMIT, WSJ0 and WSJ1 [14].

Humans adapt well to new speakers and environments. Therefore the ASR system in this work was also adapted to the NullG-Eval0 corpus using maximum likelihood linear regression (MLLR)[18] with one regression class in an unsupervised manner. To minimize the chances of adaptation depending heavily on a few sentences, block adaptation was applied in the following manner. For the 1000-word vocabulary, unsupervised MLLR was run on the quiet sentences to get the adapted models. The adapted models were then applied again to the quiet 1000-word vocabulary sentences to get the accuracy for that condition. The adapted models were then adapted to the 20 dB 1000-word vocabulary sentences using MLLR, and the accuracies were obtained using the adapted models for that condition. This process was repeated for the 10 dB and the 0 dB conditions at the 1000-word vocabulary. The process for 1000-word vocabulary was then repeated for 2000, 4000, and 8000 word vocabularies such that the initial models for each vocabulary were the baseline models trained on TIMIT and WSJ corpora without adaptation. The ordering of utterances was thus the same for both ASR and HSR.

2.3. Phone Error Propagation

Errors were analyzed at the phone level to compare how errors propagate from a phone to adjacent phones. For human transcriptions, ground truth sentences and ASR transcriptions, the sentences were mapped to phones using the CMU pronunciation dictionary [16] and the HTK tool HLEd. The following error rates and associated statistics were then computed using the output of the NIST scoring toolkit [15]:

1. $P(e)$ or probability of phone error was estimated as PER/100 where PER is the percent phone error rate in-

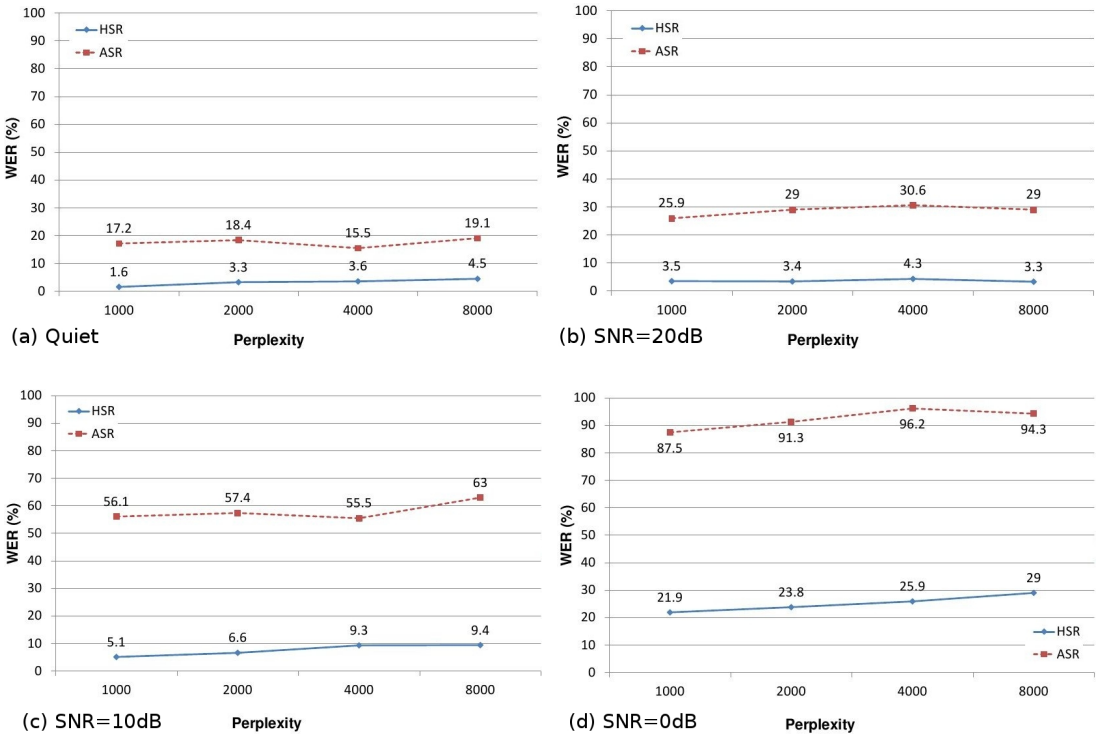

Figure 1: Variation of WER for ASR and HSR with perplexity for different SNRs. For null grammar, perplexity = vocabulary size

cluding phone deletions, substitutions and insertions. The phone errors were then modeled as a random variable $E = E'/N$ where E' is a binomial random variable with N trials such that N is the total number of reference phones. The mean of E is then $\mu = P(e)$ and the variance is $\sigma^2 = P(e)(1 - P(e))/N$.

2. $P(e|e_p)$ or probability of phone error given an error on the previous phone. $P(e|e_p)$ was estimated as a ratio of the number of consecutive phone error pairs to the number of single phone errors. The phone error given previous phone error was modeled as a random variable $E_p = E'_p/N_e$ where E'_p is a binomial random variable with N_e trials such that N_e is the total number of single phone errors. The mean of E_p is then $\mu_p = P(e|e_p)$ and the variance is $\sigma_p^2 = P(e|e_p)(1 - P(e|e_p))/N_e$.

3. $P(e|e_{2p})$ or probability of phone error given errors on both of the previous two phones. $P(e|e_{2p})$ was estimated as a ratio of the number of three consecutive phone errors to the number of two consecutive phone errors. The phone error given previous two phone errors was modeled as a random variable $E_{2p} = E'_{2p}/N_p$ where E'_{2p} is a binomial random variable with N_p trials such that N_p is the total number of phone error pairs. The mean of E_{2p} is then $\mu_{2p} = P(e|e_{2p})$ and the variance is $\sigma_{2p}^2 = P(e|e_{2p})(1 - P(e|e_{2p}))/N_p$.

If phone errors did not spread from one phone to adjacent phones $P(e)$ and $P(e|e_p)$ would be identical. To identify significant differences between $P(e)$ and $P(e|e_p)$, we used confidence intervals based on the Chebyshev bound. The Chebyshev bound states that $P\{|X - \mu| > d\} < \sigma^2/d^2$ for a random variable X with mean μ and variance σ [19]. The null hypothesis $H_0 : P(e) = P(e|e_p)$ can be rejected with at least 95% confidence (at most 5% false rejection rate) if the corresponding confidence intervals $P(e) \pm d$ and $P(e|e_p) \pm d_p$ do not overlap, where $d = \sigma/\sqrt{0.05} = 4.47\sigma$. Thus it is possible to conclude that errors propagate at a rate $P(e|e_p)$ significantly different than the chance rate of $P(e)$ if $|P(e) - P(e|e_p)| > 4.47(\sigma + \sigma_p)$. Similarly, we can conclude with at least 95% confidence that $P(e) \neq P(e|e_{2p})$ if $|P(e) - P(e|e_{2p})| > 4.47(\sigma + \sigma_{2p})$.

3. Results

3.1. ASR Control Experiments

To present a meaningful comparison between ASR and HSR, the validity of the ASR system needs to be established. For this purpose ASR was tested on the 5K word WSJ Nov 92 task using the bigram language model provided in the WSJ corpus. WERs of 5.5% and 4.7% were obtained without and with unsupervised MLLR, respectively. The WERs are in the ballpark of the best reported WERs for this task that are between 3% and 4% [17]. MLLR was run separately for each speaker using all of the sentences from that speaker. In another control experiment the bigram language model was replaced by a null grammar with the same vocabulary. WERs of 23.5% and 20.0% were obtained without and with adaptation on the Nov 92 task showing a significant absolute drop in accuracy from the bigram model. Homophone substitution

Table 1: *Percent WER for ASR is shown at different noise levels with and without MLLR on NullG-Eval0 corpus on the 1000-word vocabulary sentences*

	Quiet	20dB	10dB	0dB
Before MLLR	23.9	38.1	76.8	98.3
After MLLR	17.2	26.1	56.1	87.5

errors were not counted while scoring the transcriptions with null grammar.

3.2. HSR vs ASR Results

All of the ASR results shown in this section are from the adapted models. Using adaptation the percent WER on 1000-word vocabulary sentences dropped as shown in Table 1. Similar improvements were obtained for all vocabulary sizes.

Figure 1 shows WER comparisons for ASR and HSR as the vocabulary size is varied. Four different plots are shown - one for each SNR. ASR makes as much as an order of magnitude more errors than HSR in the quiet environment. These results show that even when given the same context - in this case a list of words - ASR performance is significantly worse than HSR for read speech at any SNR. ASR error rates double between 20dB and 10dB SNR, and seem to saturate near 100% at 0dB SNR. HSR error rates also double between 20dB and 10dB SNR, and double again between 10dB and 0dB; HSR error rates at 0dB SNR are better than ASR error rates at 20dB. For both ASR and HSR the WER increase with increasing perplexity is rather slow. Taken as an average over all SNR levels the difference in WER from the perplexity of 8000 to a perplexity of 1000 is about 4% for ASR and about 3.5% for HSR. The relative insensitivity of both ASR and HSR to an eight-fold increase in perplexity suggests that HSR and ASR differ more in their modeling of acoustic variability than in their ability to effectively apply contextual constraints. It is also interesting to note that HSR has only 3.6% and 4.5% WER in the null grammar perplexity of 4000 and 8000, respectively, in quiet environment. Those errors are close to the lowest known WER of 3.4% obtained by an ASR system using a trigram model with a similar sized 5,000 word vocabulary.

3.3. Phone Error Propagation Results

Analysis of propagation of phone errors to neighboring phones was done as described in section 2.3. Table 2 shows the values of $\mu_p - \mu$ and $4.47(\sigma + \sigma_p)$ for the 8000 word vocabulary and different SNRs; Table 3 shows a similar comparison of $\mu_{2p} - \mu$. In all cases, there is a trend toward error propagation rates greater than chance ($\mu_p > \mu$ and $\mu_{2p} > \mu$). The trend reaches significance for ASR, but not for HSR (except at 0dB SNR). It is therefore possible to conclude that ASR propagates phone errors at a rate significantly higher than chance. HSR has a tendency to propagate errors, but the tendency is not significant with a Chebyshev test at 95% confidence. At 0dB SNR the exception may have occurred because (1) it is possible that for humans two phone errors may trigger an error in the third phone with much higher probability than at lower noise levels, and (2) at 0DB SNR, ASR exhibits 87.5% WER which means both $P(e)$ and $P(e|e_{2p})$ are close to 1.0 and hence similar to each other.

Table 2: *Error propagation to one adjacent phone is significantly above chance whenever $\mu_p - \mu > 4.47(\sigma + \sigma_p)$; these columns are highlighted with the symbol ">."*

	ASR		HSR	
SNR	$\mu_p - \mu$	$4.47(\sigma + \sigma_p)$	$\mu_p - \mu$	$4.47(\sigma + \sigma_p)$
Quiet	0.29	> 0.18	0.15	0.22
20dB	0.29	> 0.14	0.19	0.25
10dB	0.16	> 0.11	0.12	0.14
0dB	0.11	> 0.09	0.24	> 0.10

Table 3: *Error propagation to 2 adjacent phones is significantly above chance whenever $\mu_{2p} - \mu > 4.47(\sigma + \sigma_{2p})$; these columns are highlighted with the symbol ">."*

	ASR		HSR	
Noise	$\mu_{2p} - \mu$	$4.47(\sigma + \sigma_{2p})$	$\mu_{2p} - \mu$	$4.47(\sigma + \sigma_{2p})$
Quiet	0.32	> 0.29	0.29	0.65
20dB	0.24	> 0.21	0.39	0.70
10dB	0.26	> 0.14	0.14	0.40
0dB	0.04	0.09	0.30	> 0.17

4. Discussion and Conclusions

ASR exhibits phone error rates as much as an order of magnitude greater than HSR even if supplied with the same context (null grammar, known vocabulary, no words are exceptional or out of vocabulary). ASR and HSR show a similar near-insensitivity to an eight-fold change in the perplexity of the task. Both ASR and HSR show a tendency to propagate phone recognition errors at a rate higher than chance, i.e., $p(e|e_p) > p(e)$. This tendency reaches significance for ASR at all noise levels, but reaches significance for HSR only at 0dB SNR.

There are at least two aspects of ASR that tend to encourage error propagation. First, ASR models acoustic context using triphone models, therefore distortion of one phone may cause errors in the recognition of neighboring phones. Second, each word in the dictionary corresponds to exactly one phone sequence. Consider two words, w_1 and w_2, with a Levenshtein distance of $D(w_1, w_2) = 2$ phones. If the ASR has poor models of the distortions caused by noise or pronunciation variability, then a production of w_1 with one distorted phone might be misrecognized as w_2, generating two phone errors. Results in this paper are therefore compatible with the claim that

ASR's acoustic models are worse than HSR's acoustic models, and that ASR overcompensates by dependence on overly constrained triphone and/or pronunciation context models.

5. References

[1] R. P. Lippmann, "Speech Recognition by Machines and Humans", Speech Communication, 22(1), 1-15, 1997.

[2] J. Baker, L. Deng, J. Glass, S. Khudanpur, C. Lee, N. Morgan, and D. O'Shaughnessy, "Research Developments and Directions in Speech Recognition and Understanding, Part 1," IEEE Signal Processing Magazine, 75-80, May 2009.

[3] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," Comp., Speech and Lang. 14(4):373-400, 2000.

[4] K. Livescu, "Feature-Based Pronunciation Modeling for Automatic Speech Recognition," Ph.D. Thesis, MIT, 2005

[5] A. Kantor, "ASR using segmental models and context dependent pronunciation models," Ph.D. Thesis, University of Illinois, 2009

[6] S. Greenberg and S. Chang, "Linguistic dissection of switchboard-corpus automatic speech recognition systems", ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium, pp. 195-202, 2000

[7] B. T. Meyer, T. Brand, and B. Kollmeier, "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes", J. Acoust. Soc. of Am. 129, pp. 388-403, 2011

[8] S. A. Phatak and J. B. Allen, "Consonant and vowel confusions in speech-weighted noise", J. Acoust. Soc. of Am., 121, pp. 2312-2326, 2007

[9] A. Juneja, "A comparison of automatic and human speech recognition in null grammar", J. Acoust. Soc. Am. Volume 131, Issue 3, pp. EL256-EL261, 2012

[10] S. J. Young et. al., "The HTK Book", Cambridge University Engineering Department, 2006

[11] Signomics LLC, "Null grammar speech recognition evaluation corpus", http://www.signomics.net/corpora

[12] J. S. Garofolo, et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus", Linguistic Data Consortium, Philadelphia 1993

[13] J. Garofalo, et al., "CSR-I (WSJ0) Sennheiser" and "CSR-II (WSJ1) Complete", Linguistic Data Consortium, Philadelphia, 1994, 2007

[14] K. Vertanen, "Baseline WSJ Acoustic Models for HTK and Sphinx: Training Recipes and Recognition Experiments", Technical Report, Cavendish Laboratory, 2006.

[15] "NIST Scoring Toolkit SCTK", http://www.itl.nist.gov/iad/mig/tests/rt/2002/software.htm

[16] "The CMU Pronouncing Dictionary", http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[17] W. Macherey, L. Haferkamp, R. Schlter, and H. Ney, "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition,, Interspeech, 2005, pp. 2133-2136.

[18] M.J.F. Gales , P.C. Woodland, "Mean and Variance Adaptation within the MLLR Framework", Computer Speech and Language, Volume 10, Issue 4, October 1996, Pages 249-264,1996

[19] A. Papoulis, "Probability, random variables, and stochastic processes," McGraw-Hill, New York, 1965.