CAN CONTINUOUS SPEECH RECOGNIZERS HANDLE ISOLATED SPEECH?

Fil Alleva, Xuedong Huang, Mei-Yuh Hwang and Li Jiang

Microsoft Research
One Microsoft Way
Redmond Washington 98052, USA

ABSTRACT

Continuous speech is far more natural and efficient than isolated speech for communication. However, for current state-of-the-art of automatic speech recognition systems, isolated speech recognition (ISR) is far more accurate than continuous speech recognition (CSR). It is a common practice in the speech research community to build CSR systems using only CSR data. In doing this we ignore the fact that isolated (a.k.a. discrete) speech is a special case of continuous speech. A slowing of the speaking rate is a natural reaction for a user faced with the high error rates of current CSR systems. Ironically, CSR systems typically have a much higher word error rate when speakers slow down since the acoustic models are usually derived exclusively from continuous speech corpora. In this paper, we summarize our efforts to improve the robustness of our speaker-independent CSR system without suffering a recognition accuracy penalty. In particular the multi-style trained system described in this paper attains a 7.0% word error rate for a test set consisting of both isolated and continuous speech, in contrast to the 10.9% word error rate achieved by the same system trained only on continuous speech.

1. INTRODUCTION

Most commercially available large vocabulary speech recognition systems require users to speak with pauses between words. As acoustic modeling improves and as computational resources grow the constraint that requires speakers to pause between words will be relaxed. While the introduction of continuous speech recognition will greatly enhance the acceptance of automatic speech recognition we cannot expect that users will always speak fluently, yet this is the expectation of current laboratory CSR systems. Typically, the user of a CSR system tends to speak fluently until the system makes an error or until the user begins pondering the composition of the document. At this point the speaker may slow down, often to the point of pausing between words. Other users familiar with discrete word recognition systems may naturally (or is it unnaturally?) begin to speak to the system in the halting manner required by ISR systems. In each case the user believes he is helping the system by speaking slowly and distinctly, with pauses between words, when in fact they are stressing the system beyond its capabilities.

In this paper, we summarize our efforts to improve the robustness of our speaker-independent CSR system without suffering a recognition accuracy penalty when users switch to isolated speech. To that end, we have collected an isolated speech corpus that we have used to address the problem of poorly trained silence context models in our CSR acoustic models. We have reduced the error rate on a mixed style test corpus by 36% over our baseline CSR system. By carefully modeling silence contexts, we can now claim that isolated speech is indeed a special case of continuous speech. This result means that system performance is now inline with user expectations; that is the accuracy of the CSR system actually improves, rather than deteriorates, when the user slows down.

2. SYSTEM DESCRIPTION

The system being discussed processes 16kHz PCM data using a MEL-scale cepstrum along with its dynamics into a multi-dimensional feature vector. The acoustic model is a gender dependent HMM model with continuous-density output probabilities, consisting of 6000 senones [5]. A mixture of 20 Gaussian densities with diagonal covariances is used for each senone. The phonetic modeling in the system consists of position and context dependent within-word and crossword triphones. Viterbi maximum likelihood training is performed, followed by maximum a posteriori smoothing [7] at each iteration. The North American Business News corpus[1] was used to derive a 60,000-word trigram language model. The recognition system is based on the pronunciation prefix-tree decoder [2]. A more complete description of the Whisper speech recognition system can be found in [3].

Two test and two training acoustic corpora were used in this study. The CSR training and test corpora correspond to the widely used SI-284 [4] training set and the 1994 H1 development test set [4] (denoted as H1dev94). The ISR training and test corpus (denoted as ISR) was collected locally and consists of 133 training speakers and 19 test speakers. The content of the ISR training corpus was partially taken from the Wall Street Journal, and partially from some simple stories designed to obtain broad phonetic coverage. The ISR test corpus content came exclusively from Wall Street Journal text. The number of phonetic units (except for silence) contained in the ISR training corpus is about three-quarters of that in the CSR training corpus. However, on the basis of duration the amount of training

Experiment	Test Set		
	ISR (Local)	CSR (H1dev94)	Average
(1) CSR Model, CSR CART	13.3	8.5	10.9
(2) ISR Model, ISR CART	4.9	42.5	23.7
(3) Parallel Model (CSR ‖ ISR)	4.9	8.6	6.8
(4) Robust CSR Model	5.1	9.2	7.2
(4a) Robust CSR Model + ISR Decoder	3.8		
(5) Robust CSR Duration Model	4.6	9.3	7.0

Table 1: Word recognition error rate for experiments run against the ISR and CSR test sets. The robust CSR model was derived from the combined CSR and ISR data.

data in the ISR training corpus exceeds that of the CSR corpus by a factor of 1.2 to 2.1 on a phone by phone basis since the word duration of isolated speech is typically longer than that of continuous speech. The sizes of the testing sets, in terms of word counts, are similar. H1dev94 consists of ~7400 words; ISR has ~7300 words.

Unless otherwise indicated all experiments were performed using our CSR search architecture. The parameters of the system are tuned based on continuous speech, in the hope of achieving the best performance for the most common usage of the SR system --- that is when users speak fluently.

3. EXPERIMENTAL RESULTS

3.1 Baseline Experiments

We built two models for our baseline experiments, the CSR model trained exclusively with continuous speech data and the ISR model trained exclusively with isolated speech data. These models are gender dependent and each model was built using its own set of gender-independent classification and regression trees (CART) for senone sharing [5].

As the first row of Table 1 shows, the word error rate on H1dev94 was 8.5%. However, its performance degraded significantly to 13.3% on ISR test data. Using the same search engine, the ISR model achieved an error rate of 4.9%. This demonstrates the irony that a typical CSR system violates the user's expectation of improved performance when their speaking rate slows. The second row of Table 1 also shows how badly an ISR system performs on the cross condition --- the error rate goes up to more than 40% on continuous speech. This is because there is no crossword co-articulation in the ISR training data. Overall, the CSR model has much better performance on the cross condition than does the ISR model. We therefore combined the results obtained with the CSR model for the test sets H1dev94 and ISR to establish the baseline error rate of 10.9% for this study.

3.2 Continuous vs. Isolated Speech Decoding

It is well know that, isolated speech requires far fewer computational resources than continuous speech. The primary reason for this is that word boundaries are easily identified allowing the decoding process to tightly constrain its search space. For this study, side information indicating the style of speech (isolated or continuous) is not made available to the decoder. Row (2) of Table 1 gives the results on the ISR and CSR test sets with the ISR Model. For these comparable test sets, it can be seen that isolated speech has a much lower error rate when contrasted with continuous speech. The error rate for isolated speech is about half that of continuous speech.

3.3 Parallel Model

To obtain the best performance on both continuous and isolated speech, we used both the CSR and the ISR models in parallel. To decode the input speech, we chose the recognized string with the highest likelihood. Running both models in parallel is certainly not resource efficient but it does provide a lower bound on the error rate we might expect to achieve with a simpler and more efficient model. For the purposes of this study, we assume that the speaking style is unknown and therefore the CSR decoder is used with both models. This yields an error rate of 6.8% on the combined test data set, Table 1, row (3). Note that using the parallel model had a small negative impact on the H1dev94 test set. This was because two utterances in the H1dev94 test set matched the ISR model better than the CSR model. These two sentences had a large number of pauses that caused the acoustic likelihood to favor the ISR model. Unfortunately, additional errors were introduced in the fluent portions of the utterances leading to the slightly increased error rate.

3.4 Robust CSR Model

In order to build a robust CSR model that gets good performance on both styles of speech we used the continuous and isolated training data together to derive both the senone mapping table and the HMM Gaussian parameters. We decided not to increase the total number of senones for the combined isolated and continuous speech model in order to see if our probabilistic models can absorb the variations of both isolated and continuous speech, without an increase in the number of model parameters. Row (4) in Table 1 shows that the robust CSR model suffers minimal degradation compared to the recognition performance of the parallel model that has twice the number of parameters. Comparing the baseline CSR model (row 1), the robust CSR model reduced the

error rate from 10.9% to 7.2%. Although the Robust model does not quite obtain the performance of the Parallel model (7.2% vs. 6.8%) it has a significant advantage over the Parallel model in that it can accept mixed mode input, i.e. speech that is a mixture of fluent and discrete words.

3.5 Likelihood Analysis of Test Data

To better understand how well each model predicts crossword triphones, we compute the average likelihood of generating the test data given a specific model. In particular, we choose the CSR model, ISR model, and the robust CSR model, *all using the robust CSR CART tree*. For simplicity, we report only the male test set and do not include the HMM transition probabilities here. Moreover, those utterances that contain out-of-vocabulary (OOV) words are removed. While computing the likelihood, the HMM phonetic state label for each data frame is determined by the Viterbi alignment algorithm, using each set of models against the correct transcription.

Triphone vs. Model	Silence-context Cross-word	Non-silence-context Cross-word	Within-Word	Context-indep. Silence
CSR model	5.51e+0	1.64e+0	5.77e+0	1.10e-8
ISR model	2.37e-1	5.32e-2	2.31e+0	1.44e-9
ISR model*	1.19e+0	6.80e-4	2.05e+0	3.82e-8
Robust CSR model	4.08e+0	1.63e+0	6.11e+0	8.21e-7

Table 2 : Average likelihood on the CSR male test set (3144 words). The third row (*) shows the likelihood assuming the HMM state segmentation is given by the robust CSR model.

For each model, Table 2 shows the average likelihoods of the CSR male test set without OOV words for silence-context crossword triphones, non-silence-context crossword triphones, within-word triphones, and context-independent silence.

As we can see from Table 2, the robust CSR model is about as good as the CSR model in all contexts, while in silence itself, the robust CSR model outperforms the baseline CSR model significantly since the ISR training corpus provides additional data to train silence. We believe the degradation in the silence-context crossword triphones is because some silences in fluent speech are very short and thus could have different effects on the neighboring phonemes from a real silence pause.

Also notice that the likelihood of row 2 is very low since the model is significantly different from the test data. When the HMM state segmentation is given by the robust CSR model, the likelihood of generating the test data with the ISR model is as shown in row 3. Here we see two points of interest: One, the silence likelihood is higher than that generated by the CSR model and thus the silence speech in the ISR training corpus will provide further positive data for training the silence model of the robust CSR model. This is shown by the silence likelihood of row 4. Two, the dramatically lower likelihood for non-silence context crossword triphones of the ISR model is as predicted and explains the ~40% error rate in Table 1.

Triphone vs. Model	Silence-context Crossword	Within-word	Context-indep. Silence
CSR model	2.29e+2	1.40e+2	8.44e-8
ISR model	1.77e+3	4.91e+2	1.46e-5
Robust CSR model	1.07e+3	3.06e+2	1.18e-6

Table 3: Average likelihood on the ISR male test set (3333 words).

Similarly, Table 3 shows the average likelihoods of the ISR male test set. Again, the likelihood table here supports the results shown in Table 1 for the ISR test set. The robust CSR model is better than the baseline CSR model on ISR test data because of the better silence model and the better silence-context crossword triphone models.

4. IMPROVING THE ROBUST CSR MODEL

One striking aspect in our experiments is the difference in error rate on the ISR test set if we impose silence context constraints in the decoding process. Using the ISR only version of our decoder the error rate was reduced from 5.1% to 3.8% (Table 1, line 4a), a 25% error reduction for the isolated test set using the robust CSR model.[1]

Clearly, if an oracle could provide knowledge regarding the style of a spoken word then the gap between these two experiments could be closed since the isolated word constraint could then be enforced. While the existence of such an oracle may seem unlikely, it turns out that controlling applications may in fact can act as the oracle. For instance, applications that have voice editing and correction modes would be in a position to inform the decoding process that exactly one word is expected.

Lacking a similar oracle for the *dictation mode* we have worked to close this gap by modeling the average duration of words in each discrete phrase of word count n, in order to discriminate multi-word phrases from single-word phrases. A discrete phrase of word count n is a sequence of n fluently spoken words beginning and ending with silence. The segmentation of the training set into discrete phrases is obtained using the word segmentation created by the training procedure. In Figure 1, we see the distribution of words with respect to discrete phrase lengths for the SI-284 data set. As can be seen the distribution has a mode of 6 and approximates a gamma distribution.

[1] A simplified version of the ISR dictation demonstration system can be downloaded from the Microsoft Research web [6].

On the same graph, we also plot the average word duration with respect to discrete phrase length. From this graph, we see the difference in average duration is significant as the discrete phrase word count grows from phrases of length one up to phrases of length six or seven. Based on this information we created a model of word duration using the gamma distribution parameterized on the word count of the discrete phrase. During decoding the output likelihood of the duration model was applied at the boundaries of discrete phrases. When this model was applied to the ISR test set the error rate was reduced by about 10% (See the last line of Table 1). Coincidentally when the same model was used on the CSR test set the error rate rose only by 1%. Several obvious improvements could be made to this model, including the use of word and phonetic identities to parameterize the model. Nevertheless, the basic observation is that the number of words in a discrete phrase is an important parameter for word duration models.

5. CONCLUSION

We have described a system capable of recognizing both continuous and isolated speech, at their respective expected performances. The error rate of the ISR data is about half of the corresponding CSR data. We have improved a typical CSR system by 36% on a combined test set consisting of equal amount of continuous and isolated words, by pooling continuous and isolated training corpora together. To further improve the performance of the system on the mixed training set we have introduced a model of word duration based on the number of words in a discrete phrase.

Another observation from our study is that the new CSR model has a slight degradation on continuous speech, compared with the baseline CSR system. This may be because the HMM parameters are overweighted by the silence-context triphone as the total number of senones remains identical in three systems compared. It is very likely that we don't have enough parameters to model the style change from the isolated speech to continuous speech.

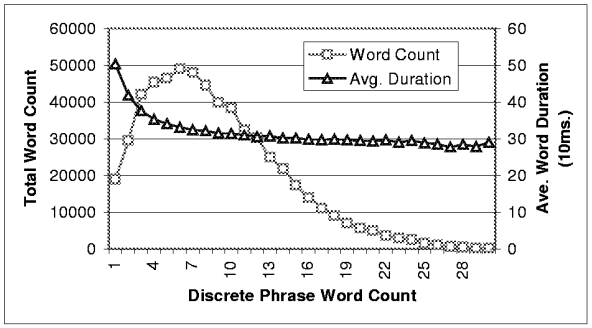

Figure 1: The word count series shows the distribution of words in discrete phrases of word count n in the SI-284 training set. The avg. duration series shows the average duration of words in discrete phrases of word count n.

REFERENCES

[1] Linguistic Data Consortium. *CSR-III Text Language Model,* University of Pennsylvania, 1994.

[2] F. Alleva, X. Huang, and M. Hwang. "Improvements on the Pronunciation Prefix Tree Search Organization". *IEEE ICASSP,* Atlanta, GA, 1995

[3] X. Huang, A. Acero, F. Alleva, M. Hwang, L. Jiang, and M. Mahajan. "From Sphinx-II to Whisper - Making Speech Recognition Usable". *Speech and Speaker Recognition-Advanced Topics*, Kulwer Publisher, 1994

[4] Linguistic Data Consortium. *CSR-II ARPA Continuos Speech Recognition Corpus,* University of Pennsylvania, 1994.

[5] M. –Y. Hwang, X. D. Huang and F. A. Alleva "Predicting Unseen Triphones with Senones". *IEEE Trans. On Speech and Audio Processing.* Vol. 4; No. 6, p412. Nov, 1996

[6] Microsoft Research Speech Technology Group : http://research.microsoft.com/stg/install.htm

[7] J.-L. Gauvain and C.-H. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Trans. On Speech and Audio Processing*, Vol.2, No. 2, pp. 291-298, April 1994.