THE SRI MARCH 2000 HUB-5 CONVERSATIONAL SPEECH TRANSCRIPTION SYSTEM

A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, J. Zheng

Speech Technology and Research Laboratory
SRI International, Menlo Park, California
http://www.speech.sri.com/

ABSTRACT

We describe SRI's large vocabulary conversational speech recognition system as used in the March 2000 NIST Hub-5E evaluation. The system performs four recognition passes: (1) bigram recognition with phone-loop-adapted, within-word triphone acoustic models, (2) lattice generation with transcription-mode-adapted models, (3) trigram lattice recognition with adapted cross-word triphone models, and (4) N-best rescoring and reranking with various additional knowledge sources. The system incorporates two new kinds of acoustic model: triphone models conditioned on speaking rate, and an explicit joint model of within-word phone durations. We also obtained an unusually large improvement from modeling cross-word pronunciation variants in "multiword" vocabulary items. The language model (LM) was enhanced with an "anti-LM" representing acoustically confusable word sequences. Finally, we applied a generalized ROVER algorithm to combine the N-best hypotheses from several systems based on different acoustic models.

1. Introduction

The goals in developing SRI's DECIPHER March 2000 Hub-5 evaluation system were twofold: first, we wanted to integrate several novel research efforts into an overall recognition system. We also wanted to significantly enhance the baseline performance of our system. The second goal was important since we felt that a competitive baseline was needed to demonstrate the benefits of new approaches, and because our previous, 1998 Hub-5 system's word error rate (WER) had lagged behind the best systems by about 10% absolute.

Hence, we decided to improve our system in as many aspects as possible, combining three strategies: (1) inclusion of known techniques not previously part of SRI's system, (2) improved implementation and tuning of previously used techniques, and (3) novel techniques not previously used in large vocabulary continuous speech recognition (LVCSR) systems. In this paper we summarize our efforts and results on all three fronts. We focus on novel methods that gave appreciable improvements in recognition accuracy, but also touch on some approaches that seemed promising but did not end up yielding improved results. We also hope that the results involving known techniques will be useful for other system developers.

Tables 1 and 2 outline the processing steps and error rates of the old and the new evaluation systems, respectively. As shown, the new system achieves a 12.5% absolute (24% relative) reduction in WER, and involves a larger number of processing steps, which will be detailed below. The system runtime on the Hub-5 2000 test set was 320 times real time on a 400 MHz Intel Pentium system. Unless stated otherwise, reported results pertain to a subset of the 1998 Hub-5 evaluation test set, consisting of 20 conversation sides (1143 utterances) that were balanced for difficulty and roughly for gender (11 females, 9 males).[1] Throughout the paper, WER changes are reported as absolute percentage point differences.

Table 1: 1998 Hub-5 system structure and performance

	Processing step	WER
1.	Gender detection	-
2.	Cepstral mean removal	-
3.	Bigram recognition with SI models,	-
	lattice generation	59.7
4.	Vocal-tract length normalization	-
5.	Phone-loop adaptation; lattice generation	55.9
6.	Transcription-mode adaptation	55.2
7.	N-best recognition, trigram rescoring	53.0
8.	Acoustic rescoring	52.6
9.	Confidence estimation	-

Table 2: 2000 Hub-5 system structure and performance

	Processing step	WER
1.	Gender detection	-
2.	Cepstral normalization; VTL normalization	-
3.	Phone-loop adaptation, bigram recognition	47.3
4.	N-best generation and trigram rescoring	45.6
5.	Transcription-mode adaptation (bigram rec)	45.5
6.	Bigram lattice generation; trigram expansion	-
7.	Adapt cross-word models; lattice recognition	43.2
8.	N-best generation	-
9.	Rescoring with class and anti-LM	41.5
10.	Rescoring with duration model	40.9
11.	N-best ROVER w/alternate acoustic models	40.1
12.	Confidence estimation	-

2. Acoustic Modeling

2.1. Front-end processing

As in the past, our system starts with gender detection using a two-state hidden Markov model (HMM) with 256 Gaussian mixtures, having one state each for male and female speech. The feature for this classification is an 8-dimensional cepstral vector. The gender

[1] For convenience, we also used a simplified scoring procedure based on the raw reference transcripts, which was less forgiving of spelling differences and optional nonlexical words than the standard NIST scoring protocol. We found that generally NIST scoring reduced the WER by about 2.5% on this development set. Only the results in Table 6 use the full NIST scoring procedure to enable comparison with ROVER.

with the higher likelihood over the entire conversation side is chosen.

After performing gender selection, we recompute the features by using a front end that was newly optimized for recognition. As did other researchers in the Hub-5 domain, we observed an improvement by widening the analysis bandwidth beyond that of the nominal telephone channel, to cover frequencies from 100 to 3760 Hz. We also increased the number of cepstral features from 9 to 13 (including C0, plus the corresponding first and second derivatives). We found that this reconfiguration of the front end alone reduced WER by about 4.4% absolute.

All features are then normalized to zero mean and unit variance for each conversation side. Variance normalization had not been part of previous systems, and was found to give a WER reduction of 0.6%. We also computed gender-dependent estimates of the speaker vocal-tract length (VTL), based on the algorithm reported in [24]. To compute the VTL, we use a 128-Gaussian mixture model trained on a subset of the training data using mean and variance normalized features. The VTL for each test conversation side is then estimated by maximizing the likelihood of the test data, searching over seven discrete VTL values in the interval [0.94, 1.06]. Once the VTL is estimated, we use it to recompute the features, which are now normalized for VTL, mean, and variance.

2.2. Cepstral modeling and adaptation

Our primary acoustic models consisted of genonic (bottom-up state-clustered), continuous density HMMs [5]. All models were gender dependent and trained from a combination of corpora: Switchboard (3094 conversation sides, 160 hours), English CallHome (100 conversations, 16 hours), and Macrophone (read telephone speech, 18 hours). About 10 hours of Switchboard material had been hand checked for transcription and segmentation errors at SRI; the remaining Switchboard transcripts were old segmentations prepared by BBN. After initial model training, all Switchboard and CallHome transcripts were subjected to a flexible realignment (similar to [7]) that allowed initial or final substrings to be skipped or replaced by a "reject" model, thus accommodating errors in segmentation. This procedure, plus an additional EM training iteration with the cleaned-up transcripts resulted in a WER improvement of 0.3%.

We observed no improvements from adding the 1996 and 1997 CallHome test sets to the training corpus. Also, we observed a small degradation in recognition accuracy when we replaced our traditional Switchboard training corpus with the retranscribed and resegmented transcripts from Mississippi State-ISIP [4], although this step also nearly doubled the amount of training material. This surprising result needs more investigation; one plausible reason is that the training set becomes excessively biased toward the characteristics of Switchboard-1 (as opposed to Switchboard-2 and CallHome). This hypothesis is consistent with the fact that others have observed improved results with an explicit stronger weighting of CallHome training data [11].

Initial N-best and lattice generation used within-word triphone models. Unlike in previous years, we also trained (and adapted) a set of cross-word triphone models, for the subsequent lattice decoding stage. The introduction of cross-word triphones reduced the WER by 1.3%.

Acoustic training resulted in 2,063 male genones and 2,348 female genones of 64 Gaussians each for the within-word triphone models. The rate-independent cross-words models used 3,064 male genones and 2,721 female genones. The rate-dependent cross-word models (see Section 2.3) comprised 3,323 male genones and 2,983 female genones. Adjusting the state clustering to produce larger models gave no improvements (although there is a possibility that this would change if we combined larger models with the added training data mentioned earlier).

Speaker-dependent acoustic models were created by a two-step adaptation process. First, we adapted the gender-dependent Gaussian means only, by maximizing the likelihood of a phone-loop model. This step does not require a prior recognition pass, yet it yields over 50% of the improvement of a transcription-mode adaptation. We combined the phone-loop adapted models with trigram N-best rescoring to obtain high-quality hypotheses for use in the subsequent transcription-mode adaptation. In this second step, we adapted the gender-dependent models again, this time using both a block-diagonal means transform [16] and variance scaling [19].

Relative to our previous system, the adaptation procedure was improved in several ways. The addition of variance scaling, which had previously been omitted, reduced WER by 0.2%. We then increased the number of phone classes (i.e., transforms) in the second adaptation pass from 3 to 7, yielding a 0.7% lower WER. Finally, we made the adaptation to transcriptions more robust to recognition errors by replacing low-confidence word hypotheses with a phone loop, similar to the one used in the first adaptation pass. For this purpose the word posterior estimates derived as a by-product of the trigram N-best rescoring were thresholded at 0.8. This combination of transcription and phone-loop adaptation reduced WER further by 0.2%.[2]

2.3. Duration and rate-of-speech modeling

Two new kinds of model were included in this year's system to specifically address duration-related aspects of conversational speech. The first of these models characterizes phone durations within a word, conditioned on both the word identity and the co-occurring phones within the word. This is achieved by modeling the joint phone duration distributions as word-dependent, multivariate Gaussians, backing off to triphone- and phone-conditioned distributions for cases of sparse training data. The phone duration model is applied as an additional knowledge source when rescoring the final N-best hypotheses, and achieved a 0.8% WER reduction at that stage. Details of the approach are described in a separate paper [17].

Duration, or local speaking rate variation, also affects the spectral properties of speech. This is accounted for in our system by having separate acoustic models for fast and slow realizations of each phone. The model is constrained to switch between fast and slow models only at word boundaries. This approach effectively combines speaking rate detection and rate-specific scoring as part of the decoding process. As described in [28], rate-dependent models lower the WER by 0.7% in our baseline system.

3. Pronunciation Modeling

3.1. Dictionary optimization

The dictionary in SRI's LVCSR system is based on version 0.4 of the CMU pronunciation dictionary. In previous systems we had simply stripped the lexical stress diacritics in the CMU phone set, based on experiments showing that stress-marked phones did not improve recognition accuracy. This year we systematically explored several

[2] Because of time constraints this last feature was not included in the evaluation system.

Table 3: Dictionary excerpt showing different kinds of multiword pronunciations: (1) reduced form, (2) concatenated canonical pronunciations, and (3) canonical pronunciations with pauses

(1)	a_lot_of	ax l aa dx ax
(2)	a_lot_of	ax l ao t ah v
(3)	a_lot_of	ax - l ao t - ah v

changes to the phone set, and settled on a variant in which unstressed [ah0] and [ih0] were coded as a separate schwa phone [ax]. We also replaced [t] and [d] in the appropriate contexts by a new flap phone [dx]. The dictionary thus modified yielded about 1% WER improvement.

3.2. Multiword modeling

Next, our goal was to model the substantial pronunciation changes, especially phone (or even syllable) reductions found in spontaneous speech [10]. Since these changes often involve phones at word boundaries and are predictable by word combinations, we decided to follow the "multiword" approach also used by other system developers (e.g., [7, 13]). Multiwords are straightforward to implement in a standard LVCSR system, since it only involves defining vocabulary items comprising multiple words (e.g., "going_to") and giving them idiosyncratic pronunciations where appropriate (e.g., "gonna").

We considered all bigrams and trigrams that occurred more than 200 times in the training data. A phonetician (Colleen Richey) examined the combined pronunciation entries and added possible idiosyncratic alternate forms.[3] Only multiwords that had pronunciations differing from the canonical forms were retained. This yielded 1,389 multiwords types with a total of 1,802 idiosyncratic pronunciations. To these, we added all canonical multiword pronunciations, taking care to include forms with pauses at word boundaries. This resulted in a total of 11,072 multiword pronunciations. Table 3 shows an example of the different kinds of dictionary entries created as a result. Overall, multiwords covered about 40% of all word tokens in the Switchboard and CallHome training transcripts. Finally, we retrained acoustic models with the new dictionary and estimated context-independent probabilities for all pronunciations, including those of multiwords.

We found that we could prune word pronunciations with probabilities smaller than 0.3 times those of the most probable variant without affecting recognition accuracy much. The pruned dictionary retained 3,652 multiword pronunciations and resulted in considerable speedup and memory savings during the initial bigram decoding phase, and actually yielded a small accuracy improvement in the lattice decoding runs, so we decided to use it in both recognition passes.[4]

To incorporate multiwords into the language model (LM), we simply replaced the appropriate bigrams and trigrams in the training transcripts with corresponding multiwords and otherwise used the standard LM training procedure. However, we obtained best results when these replacements excluded cases where noise markers or punctuation had occurred at word boundaries. We found that the multiword bigram LM performed better than the regular bigram, presumably because the multiwords effectively capture frequent higher-order N-grams. This is in agreement with [13], but runs counter to the results of [7], which could be due to the increased multiword coverage in our system, or the special transcript processing described above.

[3] This crucial step was informed by both linguistic knowledge and experience gained in the ICSI Switchboard Transcription Project [10].

[4] Post-evaluation we found a bug that had caused pronunciation probabilities to be ignored, although pruning had not been affected. While we did not rerun the entire recognition system we observed that lattice decoding results improved by 0.3% after we fixed the problem.

Table 4: Multiword experiments (male development test set)

Model	WER
No multiwords	49.0
Multiwords in LM only	48.3
Multiwords in dictionary (unpruned)	45.7
Multiwords in dictionary (probabilities)	44.5
Multiwords in dictionary (pruned)	44.8

Results Table 4 shows comparative results with various stages of multiword modeling on the male subset of the development set, using a bigram recognizer. We found that multiwords in the LM alone gave a 0.7% improvement, which increased to 3.3% when idiosyncratic multiword pronunciations were added without probability weighting. Adding pronunciation probabilities gave an additional 1.2% improvement, which was only slightly reduced by dictionary pruning. On the full development test set, the combined WER reduction with pruning was 4.4%. However, we found that later in the recognition system, the incremental win from decoding trigram lattices was reduced by 0.4%, consistent with the notion that the multiword-bigram LM already benefits from partial modeling of higher-order N-grams.

4. Language Modeling

4.1. Word- and class-N-grams

Initial decoding used a multiword bigram backoff LM containing about 1.3M bigrams. The LM was trained from all Switchboard-1 transcripts (3M words), 100 CallHome conversations (210K words), and the Broadcast News (Hub-4) LM training corpus (130M words). The recognition vocabulary contained 34,000 word types, including all those found in the spontaneous speech materials and the 10,000 most common words from the Broadcast News corpus. Considerable effort was spent in trying to make the Broadcast News transcripts conform to the Switchboard vocabulary (including the replacement of multiwords). Separate LMs were trained from each corpus and then statically interpolated into a single backoff model using the SRILM tools [21]. The interpolation weights had been optimized for perplexity on prior evaluation data. To save memory and time during initial decoding, we also pruned the LM of bigrams that caused less than 10^{-8} relative change in perplexity [20]. Lattice expansion used an unpruned, trigram backoff LM (4.8M bigrams, 11.5M trigrams) constructed in the same fashion. The compact trigram expansion technique described in [26] was employed for incorporating trigram LM scores into the lattices prior to the second decoding pass. Our 1998 evaluation system did not use trigrams in recognition from lattices, leaving them for rescoring of the final N-best lists. We estimate that the lattice-based trigram search reduced final WER by about 0.7%.

A further improvement was obtained by rescoring N-best lists with a class-based 4-gram, for which word classes had been automatically induced from the Switchboard and CallHome texts using a mutual-information criterion [3]. The class-LM probabilities were interpo-

Table 5: Anti-LM performance compared to a baseline LM

Model	Test set		Weight	
	Tuning	Held-out	Std. LM	Anti-LM
Baseline	41.9	43.5	9.1	n/a
Anti-LM	41.5	43.1	13.6	-2.4

lated with the standard trigram at the word level, for a win of 0.6%.

Two additional potential LM improvements we investigated were an explicit optimization of the vocabulary size, and tuning of the reject model probability (the reject model corresponds to unintelligible and out-of-vocabulary words and fragments in the training transcripts, but is otherwise treated as a regular word). However, neither of these experiments gave improved results.

4.2. Anti-language model

In previous years we had experimented with various discriminative LM training approaches, with the goal of making the LM sensitive to the acoustic model and to optimize for overall recognition error [23]. These experiments, based on maximum mutual information estimation and gradient descent in the LM parameter space, were largely unsuccessful because of data sparseness and overfitting problems. This year we pursued a similar goal with a more heuristic, but, as we hoped, more robust approach.

Our approach is to construct a separate "anti-LM" of those hypotheses that are acoustically confusable with correct transcriptions. The resulting N-gram LM gives a score that can be used to penalize likely misrecognitions. The idea is different from, yet similar in spirit to, other corrective modeling approaches that adjust model parameters away from recognition errors [12, 1] or that learn post-recognition error correction [18]. The anti-LM itself can be trained on the acoustic training corpus, and only a single penalty weight parameter needs to be estimated on held-out data, making the estimation very robust.

We implemented the anti-LM as follows: 500-best recognition hypotheses for a 1.6M word subset of the Switchboard and CallHome training corpora were generated. The hypothesized N-grams were weighted by the posterior probabilities (normalized N-best scores) of the hypotheses in which they occurred. N-grams with a total expected count of at least 1 were used in estimating a backoff trigram anti-LM. The Witten-Bell discounting scheme [27] was employed since it naturally generalizes to fractional counts. We then generated anti-LM scores for the 2000-best hypotheses from our development test set, and optimized the log-linear score combination weights relative to the standard acoustic and language models.

Results Table 5 shows the N-best rescoring performance on both the development set used for tuning and the held-out data. On both data sets the anti-LM reduces the WER by 0.4% compared to the baseline without anti-LM. Also shown are the optimized weights for the two LMs. The optimization of the anti-LM weight was allowed to use both positive and negative values, yet it settled on a negative value as intended.

We also experimented with variants of the training procedure. One variant was to remove the correct N-grams from the posterior N-best distribution, to emphasize incorrect outputs; another experiment used only acoustic scores to compute posterior expected counts, to let the anti-LM focus on acoustic confusability. However, both of these modifications degraded the results slightly.

Incidentally, N-best generation for training the anti-LM made use of a recognizer that was much poorer than our current system (it used the 1997 Hub-5 acoustic models, bigram LM, and no speaker-adaptive features). The effectiveness of the anti-LM under these conditions speaks for the robustness of the training approach; on the other hand, we can expect further improvements from a complete retraining with the current recognition system.

5. Model Combination

5.1. Progressive search organization

Our system follows the principle of progressive search [15], whereby successively more detailed (and computationally expensive) knowledge sources are brought to bear on the recognition search as the hypothesis space is narrowed down. Accordingly, we use within-word triphone acoustic models and a bigram LM for initial, unconstrained recognition; followed by trigram LMs and cross-word trigrams for decoding from lattices; followed by N-best rescoring with class-based 4-gram, anti-LM, and duration models. A revised lattice-generation and expansion algorithm [26] allowed us to apply cross-word models and the trigram LM earlier in the search than in previous systems.

Rate-dependent models were integrated into the evaluation system by generating a separate set of N-best lists and applying the system combination technique described in the next section. This approach proved superior to acoustic rescoring of N-best lists with the rate-dependent models.

5.2. N-best ROVER

The widely used ROVER approach to system combination [8] combines the 1-best output from several recognition systems by voting among the various hypotheses at the word level. In a related approach, it was been shown that WER can be reduced by letting the word hypotheses from a *single* recognizer vote with their posterior probabilities, since this reduces the expected word (rather than sentence-level) error [22, 14]. This leads to a natural generalization of both approaches, in which the N-best lists from *multiple* systems are combined. Word hypotheses can then compete on the basis of posterior probability estimates that are interpolated from multiple system outputs, and are therefore more accurate. This way, for example, two second-ranked hypotheses could override a 1-best hypothesis if the combined posterior is high enough.[5]

Algorithm The "N-best ROVER" algorithm starts by word-aligning N-best hypotheses h from multiple systems S_i. Each system computes its own word posterior estimates by log-linear score weighting, followed by normalization over all hypotheses:

$$P_i(w|X) = \frac{\sum_{h:w\in h} e^{\sum_j \lambda_{ij} s_{ij}(h|X)}}{\sum_{\text{all } h} e^{\sum_j \lambda_{ij} s_{ij}(h|X)}} \tag{1}$$

where w is a word hypothesis and $s_{ij}(h|X)$ is the jth log score for hypothesis h in system S_i. The combined posterior is computed as a linear combination

$$P(w|X) = \sum_i \mu_i P_i(w|X) \tag{2}$$

[5] The idea of combining multiple hypotheses spaces was independently developed by [6] for lattices, and again for N-best lists by [9]. Our implementation of N-best ROVER is available in [21].

Table 6: System combination results with N-best ROVER

System	WER
Rate-indep. cross-word + duration + anti-LM	37.6
Rate-dependent cross-word	39.2
Rate-independent non-cross-word	41.2
N-best ROVER	37.1
Standard ROVER	37.4

where μ_i are system weights, empirically chosen and summing to 1. As usual in N-best or lattice-based voting, the word hypotheses with the highest posterior at each position in the alignment are concatenated.

Results Table 6 shows comparative results with three individual systems, N-best ROVER combination, and the standard 1-best ROVER. The parameters of both ROVER methods had been optimized for the test set. The WER with N-best ROVER was about 0.3% below that of the standard ROVER, consistent with results by [6]. We also found that it was best to combine systems with minimal overlap in their knowledge sources. Accordingly, only the first system incorporated duration and anti-LM rescoring.

The N-best ROVER result (37.1%) was the final WER of our evaluation system on the development set, using the full NIST scoring protocol. The corresponding number on the March 2000 test set was 30.2%.

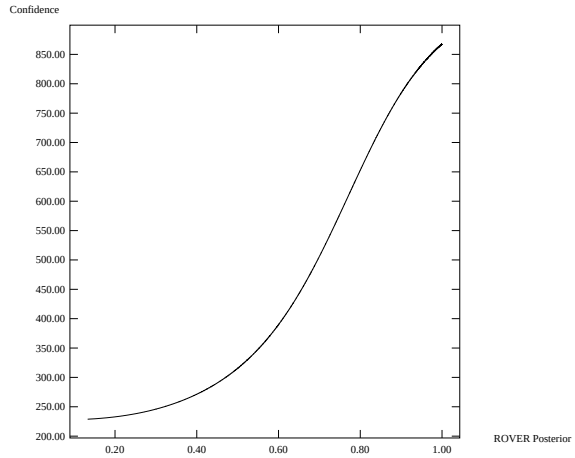
Weight optimization for word-level scoring The score combination weights λ_{ij} in Equation (1) need to be optimized discriminatively. In the past we achieved this approximately by carrying out the usual optimization for *sentence-level* hypothesis ranking, and then rescaling so that the LM receives weight 1. We recently developed a new approach that directly optimizes the weighting for *word-level* hypothesis selection, inspired by the discriminative model combination (DMC) algorithm [2]. However, because of the form of (1), the closed-form solution of DMC does not apply; instead, we optimize by gradient descent on a smoothed word-error function in the style of GPD [12].

6. Confidence Estimation

As in previous years, we used a neural network to estimate word correctness probabilities (confidences) from word-level features [25]. However, because of time constraints, we limited the number of input features severely. Only the combined word log posteriors from the N-best ROVER system were used, since this measure already constitutes a confidence measure that includes all knowledge sources used by the recognizer. The network simply adjusts the posterior estimates to compensate for the bias resulting from the limited hypothesis space represented in the N-best list, as shown in Figure 1.

We added two minor, readily available features that could help the network gauge the magnitude of the posterior overestimates. Using the overall number of words in the hypothesis and the relative position of the word within the utterance, we achieved a small (1% relative) reduction in cross-entropy. The normalized cross-entropy (NCE) achieved on the 1998 development subset was 0.207 (0.233 on the March 2000 test set). Since these values are considerably higher than in previous systems we conclude that the N-best ROVER approach significantly improves the preliminary posterior estimates compared to those of a simple N-best approach.

Figure 1: Word-posterior to confidence mapping by neural network



7. Summary

Overall, the combination of techniques described here reduced word error rate by about 12% absolute in our Hub-5 system. Table 7 gives an approximate breakdown of this improvement estimated from various contrastive experiments; the sum of individual WER reductions is smaller than the actual total reduction, partly because the baseline WERs for the various contrasts were higher than in the final system, and partly because some of the approaches overlap in what they model (e.g., cross-word acoustic modeling and multiword pronunciations).

We note that by far the largest improvements were achieved by a reconfigured, broader-band front end, and by extensive cross-word pronunciation modeling via multiwords. Three novel knowledge sources, a duration model, rate-dependent acoustic models, and the anti-LM each contributed small, but significant improvements. Finally, an N-best generalization of the ROVER technique gave an additional win when combining multiple system outputs, as well as yielding improved word posterior estimates for confidence estimation.

Table 7: Factors that improved recognition accuracy

What	Δ WER
Wider front end/more cepstral coeffs.	-4.4
Multiword dictionary	-4.0
Cross-word triphones	-1.3
Schwas and flaps in dictionary	-1.0
Duration model	-0.8
Rate-dependent model	-0.7
More detailed adaptation transform	-0.7
Trigram lattices	-0.7
Cepstral variance normalization	-0.6
Class LM	-0.6
N-best ROVER	-0.5
Anti-LM	-0.4
Training transcript cleanup	-0.3
Variance scaling transforms	-0.2
Total	-16.2
Actual	-12.5

Acknowledgments

We thank Erik McDermott, Mitch Weintraub, and Ananth Sankar for helpful discussions. The work reported here was supported by the Department of Defense and SRI Internal Research & Development funds, with additional funding from NSF grant IRI-9619921. The views herein are those of the authors and should not be interpreted as representing the policies of the funding agencies.

References

1. L. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer. Estimating hidden Markov model parameters so as to maximize speech recognition accuracy. *IEEE Trans. Speech Audio Process.*, 1(1):77–83, 1993.
2. P. Beyerlein. Discriminative model combination. In *Proc. ICASSP*, vol. I, pp. 481–484, Seattle, WA, 1998.
3. P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.
4. N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone. Resegmentation of Switchboard. In R. H. Mannell and J. Robert-Ribes, editors, *Proc. ICSLP*, pp. 1543–1546, Sydney, 1998. Australian Speech Science and Technology Association.
5. V. Digalakis, P. Monaco, and H. Murveit. Genones: Generalized mixture tying in continuous hidden Markov model-based speech recognition. *IEEE Trans. Speech Audio Process.*, 4(4):281–289, 1996.
6. G. Evermann and P. Woodland. Posterior probability decoding, confidence estimation, and system combination. In *Proceedings NIST Speech Transcription Workshop*, College Park, MD, 2000.
7. M. Finke and A. Waibel. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, *Proc. EUROSPEECH*, vol. 5, pp. 2379–2382, Rhodes, Greece, 1997.
8. J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 347–352, Santa Barbara, CA, 1997.
9. V. Goel and W. Byrne. Applications of minimum Bayes-risk decoding to LVCSR. In *NIST Speech Transcription Workshop*, College Park, MD, 2000.
10. S. Greenberg. Speaking in shorthand—A syllable-centric perspective for understanding pronunciation variation. In *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pp. 47–56, Kerkrade, The Netherlands, 1998.
11. T. Hain, P. C. Woodland, G. Evermann, and D. Povey. CU-HTK March 2000 Hub 5E transcription system. In *Proceedings NIST Speech Transcription Workshop*, College Park, MD, 2000.
12. S. Katagiri, C.-H. Lee, and B.-H. Juang. New discriminative training algorithms based on the generalized probabilistic descent method. In B. H. Juang, S. Y. Kung, and C. A. Kamm, editors, *Proceedings IEEE Workshop on Neural Networks for Signal Processing*, pp. 299–308, 1991.
13. K. Ma, G. Zavaliagkos, and R. Iyer. BBN pronunciation modeling. In *9th Hub-5 Conversational Speech Recognition Workshop*, Linthicum Heights, MD, 1998.
14. L. Mangu, E. Brill, and A. Stolcke. Searching for consensus to improve recognition output. In *9th Hub-5 Conversational Speech Recognition Workshop*, Linthicum Heights, MD, 1998.
15. H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub. Large-vocabulary dictation using SRI's DECIPHER speech recognition system: Progressive search techniques. In *Proc. ICASSP*, vol. II, pp. 319–322, Minneapolis, 1993.
16. L. Neumeyer, A. Sankar, and V. Digalakis. A comparative study of speaker adaptation techniques. In J. M. Pardo, E. Enríquez, J. Ortega, J. Ferreiros, J. Macías, and F. J. Valverde, editors, *Proc. EUROSPEECH*, Madrid, 1995.
17. V. R. Rao Gadde. Modeling word duration for better speech recognition. In *Proceedings NIST Speech Transcription Workshop*, College Park, MD, 2000.
18. E. K. Ringger and J. F. Allen. Error correction via a post-processor for continuous speech recognition. In *Proc. ICASSP*, vol. 1, pp. 427–410, Atlanta, 1996.
19. A. Sankar and C.-H. Lee. A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Trans. Speech Audio Process.*, 4(3):190–202, 1996.
20. A. Stolcke. Entropy-based pruning of backoff language models. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pp. 270–274, Lansdowne, VA, 1998. Morgan Kaufmann.
21. A. Stolcke. SRILM—the SRI Language Modeling Toolkit version 1.0.1. http://www.speech.sri.com/projects/srilm/, 2000.
22. A. Stolcke, Y. Konig, and M. Weintraub. Explicit word error minimization in N-best list rescoring. In G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, *Proc. EUROSPEECH*, vol. 1, pp. 163–166, Rhodes, Greece, 1997.
23. A. Stolcke and M. Weintraub. Discriminative language modeling. In *9th Hub-5 Conversational Speech Recognition Workshop*, Linthicum Heights, MD, 1998.
24. S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin. Speaker normalization on conversational telephone speech. In *Proc. ICASSP*, vol. 1, pp. 339–341, Atlanta, 1996.
25. M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke. Neural-network based measures of confidence for word recognition. In *Proc. ICASSP*, vol. 2, pp. 887–890, Munich, 1997.
26. F. Weng, A. Stolcke, and A. Sankar. New developments in lattice-based search strategies in SRI's Hub4 system. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pp. 138–143, Lansdowne, VA, 1998. Morgan Kaufmann.
27. I. H. Witten and T. C. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. Inform. Th.*, 37(4):1085–1094, 1991.
28. J. Zheng, H. Franco, and A. Stolcke. Rate-dependent acoustic modeling for large vocabulary conversational speech recognition. In *Proceedings NIST Speech Transcription Workshop*, College Park, MD, 2000.