New Results with the Lincoln Tied-Mixture HMM CSR System[1]

Douglas B. Paul

Lincoln Laboratory, MIT
Lexington, Ma. 02173

ABSTRACT

The following describes recent work on the Lincoln CSR system. Some new variations in semiphone modeling have been tested. A very simple improved duration model has reduced the error rate by about 10% in both triphone and semiphone systems. A new training strategy has been tested which, by itself, did not provide useful improvements but suggests that improvements can be obtained by a related rapid adaptation technique. Finally, the recognizer has been modified to use bigram back-off language models. The system was then transferred from the RM task to the ATIS CSR task and a limited number of development tests performed. Evaluation test results are presented for both the RM and ATIS CSR tasks.

INTRODUCTION

The following experiments are all carried out in the context of the Lincoln tied-mixture (TM) hidden Markov model (HMM) continuous speech recognition (CSR) system. This system uses two observation streams (TM-2) for speaker-dependent (SD) recognition: mel-cepstra and time differential mel-cepstra. For speaker-independent (SI) recognition, a second differential mel-cepstral observation stream is added (TM-3). The system uses Gaussian tied mixture [1, 2] observation pdfs and treats each observation stream as if it is statistically independent of all others. Triphone models [14], including cross-word triphone models [10, 7, 16], are used to model phonetic coarticulation. These models are smoothed with reduced context phone models [14]. Each phone model is a three state "linear" (no skip transitions) HMM. The phone models are trained by the forward-backward algorithm using an unsupervised monophone (context independent phone) bootstrapping procedure. The recognizer extrapolates (estimates) untrained phone models and recognizes using a Viterbi beam search. The initial implementation uses finite-state grammars, contains an adaptive background model, and allows optional inter-word silences. All RM1 development tests use the designated SD development test set (100 sentences x 12 speakers) and all RM2 tests use the designated development test set (120 sentences x 4 speakers).

[1]This work was sponsored by the Defense Advanced Research Projects Agency.

SEMIPHONES

One difficulty with the current triphone-based HMM systems with cross-word triphone models is that the number of triphones becomes very large (~60K triphones) when used in a large (20K word) vocabulary task[11]. This requires estimation of very large numbers of parameters and makes execution of the trainer and recognizer inefficient on practical hardware. We have previously proposed semiphones as a modeling unit because they significantly reduce the number of elemental phonetic models by as much as an order of magnitude. (Semiphone models split each phone into a triplet of left and right context dependent models[11]. Semiphones include triphones and "classic" diphones—which extend from the center of one phone to the center of the next—as special cases.) On the Resource Management (RM) task, they reduced the number of unique states by about a factor of 5 at the cost of a performance penalty of about 20% for the speaker-independent (SI) task and 30% for the speaker-dependent (SD) task.

The initial semiphone system used 1 left state, 1 center state, and 1 right state (notation: 1-1-1) system [11]. (In this notation, a triphone system is designated by 0-x-0 and a classic diphone system is designated by x-0-y.) We have recently explored a number of other variations on the semiphone scheme subject to the constraint of three states per phone. The performance of 2-0-1 and 1-0-2 systems is shown in table 1. The lower error rate of the 1-0-2 system suggests that, on the average, the anticipatory coarticulation is stronger than the backward coarticulation. This agrees with an assertion by Ladefoged that English is dominantly an anticipatory coarticulation language [6].

We have also tested a hybrid triphone-semiphone system. This hybrid used 1-0-2 semiphones for the cross-word models and triphones for the word-internal models. (50K of the above mentioned 60K triphones were cross-word-context phones.) Its performance was the same as the 1-0-2 system. This suggests that the less detailed modeling of the word boundary phones is the primary site where information is lost in the semiphone systems compared to the triphone

systems.

These results may be affected by the lack of richness in the RM database—there were 1752 word-internal (WI) semiphones and 2413 WI triphones and therefore only 27% of the WI triphones were merged in transitioning to the semiphone models. Similarly there were 1891 cross-word (XW) semiphones and 3580 XW triphones and therefore 47% of the cross-word (XW) triphones were merged in the transition. Thus the transition to semiphones would be expected to affect the XW modeling more than the WI modeling. All of the XW semiphone systems, however, outperform the corresponding non-XW triphone systems.

Attempts to improve semiphone results by smoothing the mixture weights with occurrence based smoothing weights[14] proved unsuccessful. (This form of smoothing significantly improved the triphone system results [11].) This correlates with the reduced number of single occurrence models in the semiphone system (1340=37% of the semiphones) compared to the triphone system (3094=52% of the triphones).

IMPROVED DURATION MODELING

The standard HMM system suffers from the difficulty that an incorrect phone can minimize its scoring penalty by minimizing the dwell time of the path through its model. The current CSR uses three states per phone and can suffer from this problem for long duration phones. Since there are no skip arcs within the phone model, a path can traverse a phone in 30 msec (3 time steps). Some phones are essentially never produced with this short a duration and therefore an incorrect short segment matched to this phone can have too high a score.

One way to minimize this problem is alter the phone model to increase the minimum path dwell time to a time commensurate with the minimum duration of the phone. Since this system does not adapt in any way to the speaking rate, the desired minimum would be the minimum duration at the fastest speaking speed. Since the available training data is not fast speech, a pragmatic estimate of the minimum might be the shortest observed duration times a safety factor. An additional difficulty in estimating the minimum duration is that some phones are observed only a very few times in the training data thereby making such an estimate less reliable.

For this experiment, a much simpler estimate of the minimum duration was chosen. The system was trained normally with three states per phone, which has the dual advantages of maintaining a uniform phone topology to allow smoothing between different phone models and of not increasing the number of parameters to be estimated. Finally, states whose average duration (as computed from the stay transition probability) was above a constant were split into a linear sequence of states until each final state had an average duration below the constant. Each of the split states shared the same observation pdf—only the stay and move transition probabilities were altered on the split states. Since no skip transitions were allowed in the phone models, the minimum duration was proportional to the final number of states in the phone.

This simple strengthening of the duration model improved the triphone system results by about 10% for both SI and SD systems (Table 2). This result is in agreement with a similar improvement obtained adding minimum phone duration constraints to a large vocabulary IWR[8]. The overall amount of computation was not significantly changed. Essentially all of the word error rate reduction was a result of reduced word insertion and deletion error rates.

NEW TRAINING STRATEGY WITH IMPLICATIONS FOR ADAPTATION

A modified multi-speaker/speaker-independent training strategy was tested. The standard strategy used to date has been:

1. Monophone bootstrap
2. Train triphones (all parameters trained on all speakers)

The new strategy is:

1. Monophone bootstrap (single set of Gaussians)
2. Train triphones (transition probabilities and mixture weights trained on all speakers, speaker-specific Gaussians)
3. (Optional) Fix transition probabilities and mixture weights and train a single set of Gaussians on all speakers

This new multi-speaker (MS)/SI strategy (without the option), in effect, implements a theory to the effect that all persons speak alike except that each uses a different section of the acoustic space, perhaps due to differently sized and shaped vocal tracts.

The new strategy without the option uses more data to train the mixture weights and might therefore, with the speaker-specific Gaussians, provide better SD recognition than the old method. It was significantly worse than the standard SD training for the RM1 database (12 speakers, Table 3), but slightly better for the RM2 database (4 speakers, Table 4). In both cases the new procedure was better than the SI-109 system.

The new strategy with the option is a new method for training a MS or SI system. The mixture weights are again trained in the context of speaker-specific Gaussians, but then the weights are fixed and a single set of MS or SI Gaussians trained. In all cases, the systems using SD Gaussians outperformed the MS/SI Gaussians. On the RM1 database, the old training method outperformed the new with the option respectively for both the MS-12 and the SI-109 training condition. Similarly, when training on the RM1 database and testing on the RM2 database, the old training method outperformed the new with the option respectively for the SI-12 and SI-109 training conditions. (The MS-12 models from RM1 become SI-12 when tested on the RM2 database because the RM2 database uses speakers which are not included in RM1.)

The controls for this experiment (SI-109 and SI-12), when tested on the RM2 database, confirm BBN's result [4] that similar SI performance can be obtained by training on large amounts of data from a small number of speakers as

with the traditional method of training on small amounts of data from a large number of speakers. However, unlike the BBN results, the SI-12 training condition yielded better results than did the SI-109 training condition. However, since the SI-12 condition used almost twice as much training data as did the SI-109 condition (7200 vs. 3990 sentences), the test is biased toward the SI-12 condition. Thus the better result for SI-12 may reflect the bias rather than any inherent advantage.

The tests using the SD Gaussians bear a similarity to a speaker adaptation method developed by Rtischev[13]. Rtischev transferred the codebook (Gaussians) and observation probabilities (mixture weights) from a SD discrete observation system to a TM system, adapted the Gaussians to a new speaker keeping the mixture weights fixed, and transferred the parameters back to the discrete observation system. Using only very limited amounts of training data, he was able to more than halve the error rate of the new speaker relative to the error rate obtained by using the parameters from the original speaker. While the experiments performed here do not actually test this adaptation procedure, the improved results for SD Gaussians over MS Gaussians for otherwise identical systems are consistent with Rtischev's results and suggest that this adaptation method should be tested in a TM system.

BIGRAM LANGUAGE MODELS

The recognizer has been modified to use bigram back-off language models [3] as well as finite state language models. The bigram language models, unlike the word-pair grammar (WPG) used with the RM database, have non-zero probabilities for all possible word transitions. In some preliminary testing, the minimum word error rate was found to occur at a lower language model weight than for minimum sentence error rate. (The language model weight is a scaling factor applied to the log word probabilities to balance the contribution of the acoustic and language model information sources.) This appears to be due to the higher weight reducing the sentence error rate by increasing the word constraints, but once a word error occurs, the increased constraint increases the word error propagation. The minimum word error rate occurred with a weight of about five, but comparison with another site [15] suggests that the optimal value is dependent upon other aspects of the recognizer. Recognition performance comparisons between the RM word-pair grammar and bigram back-off language models are presented in [12].

INFORMAL ATIS CSR BASELINE EVALUATION TEST DEFINITION

The value of well defined common evaluation testing has been amply demonstrated on the RM task over the past few years. We felt that the ATIS CSR evaluation tests would be more meaningful to system developers if a common baseline test were performed in addition to any tests demonstrating site specific variations. (The official evaluation test specified only the test data—there was no specification of an official training data set, vocabulary, or testing language model.) We also discovered that the SNOR (Standard Normalized Orthographic Representation) transcriptions of many of the read training files were formatted inconsistently.

In collaboration with BBN and NIST, we cleaned up the SNOR transcriptions of the spontaneous and read data and made them consistent with the transcription conventions used in the June 1990 tests. For example, numbers were variously specified by digits, words, and hyphenated words. All representations were changed to words separated by white space. This cleanup was performed on the data on CD-ROMs ATIS0_5-1.1 (June 90 spontaneous speech, 36 speakers), ATIS0_5-2.1 (read versions of sentences from 20 of the June 90 speakers), ATIS0_5-3.1 and ATIS0_5-4.1 (read sentences from a list of 2900 sentences and 40 adaptation sentences, 10 speakers), and 8mm. tape ATIS1 recorded at SRI (spontaneous speech).

Next, we designated the June 90 spontaneous and read test speakers (bd, bf, bm, bp and, bw) as development test speakers and all other ATIS data from the above sources to be language model training data. This training data contains an (observed) vocabulary of 921 words. BBN then extended the vocabulary by closing obvious classes (e.g. days and months) [5] and we verified the additions. The final lexicon contained 1065 words. This lexicon and the training sentences were used to generated a bigram back-off language model [3]. A bigram model, rather than a stronger trigram model was chosen because it could be easily integrated into the current CSR implementations at most sites. This language model included estimates of the probabilities of the unobserved words in the lexicon and an estimate of the probability of unknown (out of lexicon) words[12]. The June 90 test set perplexity of this language model was 17.8. Note that since this language model was trained solely on SNOR format transcriptions, it knows nothing about disfluencies and non-speech phenomena. The language model was specified in a simple machine-independent text format.

Finally, we designated acoustic training and development test sets. The June 90 spontaneous test set mentioned above was also designated as the acoustic development test set. (It was suggested, but not required, that the read versions of the test sentences also be reserved for testing.) All of the non-testing data distributed on the CD-ROMs listed above was designated as acoustic model training data. (Therefore, this also included the adaptation sentences.) This data was chosen because it was available to all sites and was distributed to all interested sites simultaneously. The ATIS1 acoustic data was omitted from the acoustic training set because there was a microphone problem with some of the acoustic data. This yielded a total of 5020 training sentences. Three of the five test speakers are included in the training set.

The corrected SNOR sentence lists, lexicon, language model, and data set designations were made available to all sites via anonymous FTP on January 8 and a minor correction made on January 10.

ATIS PILOT DEVELOPMENT TESTS

The initial ATIS development tests were performed before the all of the above data was available. Therefore only

the June 90 spontaneous training (774 sentences) and test data was used. Due to the limited amount of time available before the evaluation tests, no attempt was made to model the open vocabulary, disfluencies partial words, thinking noises and extraneous noises. Thus the SNOR transcriptions of the acoustic data were used for both training and testing. The lexicon (548 words) and a bigram back-off language model were generated from the training data which produced a test set perplexity of 23.8 with 1.3% out-of-vocabulary words.

The first system was as described in the introduction except that the system used SI TM-2 non-cross word triphone models and the improved duration modeling described above. Recognition was performed using the perplexity 23.8 bigram language model. The pilot tests were all SI trained with two observation streams. The closest RM system showed an SI-109 WPG word error rate of 10.4% [11]. After fixing some pruning difficulties in training due to the large silences in the training data, the system produced a word error rate of 37.5% (Table 5). Enabling optional inter-word silences in training reduced the pruning difficulties and improved the recognition performance to 33.3% (Table 5). (Optional inter-word silences during training had been tested on the RM task and found not to help the performance.) Finally, this system was tested using the perplexity 17.8 baseline language model and the error rate was reduced to 30.9% (Table 5).

ATIS BASELINE DEVELOPMENT TESTS

When the baseline test definition became available, the best pilot system was trained on the baseline training data. The error rate improved to 26.4% (Table 6). The additional data, which consisted of read in-task sentences and read adaptation sentences, increased the number of training sentences by a factor of 6.5, but produced a surprisingly small performance improvement. Cross-word triphone modeling was added which reduced the word error rate to 23.0%. (The closest corresponding system RM SI-109 WPG error rate is 8.5% [11].) Next, the third observation stream (second differential mel-ceptsra) was added (TM-3) which increased the error rate to 25.3%. In contrast, a 30% error rate reduction on the SI RM task occurred when the third observation stream was added[11]. Finally, a TM-3 1-0-2 semiphone system yielded 24.0% word error rate, which is between the results obtained with the TM-2 and TM-3 triphone systems.

EVALUATION TESTS

The SD and SI-109 RM evaluation tests were run with WPG and no grammar (NG). The systems are identical to the systems tested in the last set of evaluation tests[11] except the enhanced duration models were used. The SD system used two observation streams and the SI-109 system used three observation streams. The average word error rates with the WPG are 1.77% and 4.39% respectively (Table 7).

Due to the limited time between the distribution of the ATIS development data and the deadline for the evaluation tests, it was not possible to test all desired systems nor was it possible to adequately set the recognition parameters such as the grammar weight and word insertion penalty. As noted earlier, the open vocabulary, disfluencies, partial words, thinking noises, and extraneous noises were not modeled. The tested system is an SI TM-2 XW triphone system with the improved duration model. The test set perplexity of the class A test data was 24 with .8% out-of-vocabulary words using the informal baseline language model and the recognition word error rate was 26.5% (Table 8). The non-Class A test sets were also tested. Their results and perplexities are shown in Table 8. The recognition output sentences (top-1) were sent to Unisys to be input to their natural language system[9].

DISCUSSION AND CONCLUSIONS

While the additional work on semiphone models has not yielded any improvements over the original semiphone systems, they still represent a potentially useful tradeoff. They still yield a 20-30% higher error rate than do triphone models, but provide more than an order of magnitude reduction in the number of states required in a large vocabulary recognition system.

The improved duration model, as tested here, is extremely simple way to reduce the error rate by about 10%. A better method for determining the minimum state durations might be to perform a Viterbi alignment of the training data and determine the desired splitting factor from the observed minima.

The new training strategy, while it did not improve performance as tested, did yield results consistent with a method of rapid speaker adaptation. This method of speaker adaptation, which is performed by a modified TM trainer, is well suited to the current DARPA applications.

The bigram back-off language model was added to the Lincoln CSR. This made the system operational with a more practical class of language models than the previously implemented finite state grammars. In particular, it made testing on the ATIS CSR task feasible.

The tripling of error rates obtained on the ATIS task compared to the RM task is quite reasonable. A perplexity 25.7 bigram back-off language model trained on 8K RM sentences resulted in an approximate doubling of the error rate compared to the WPG[12] and the perplexity 17.8 ATIS bigram language model was trained on only 4K sentences. Thus, only a factor of about 1.5 increase occurred due to the extemporaneous speech and the less controlled environment.

Given the limited time between distribution of the data and the evaluation tests, it has not been possible to adequately study the difficulties unique to the ATIS database nor has it been possible to adequately test our systems. There are some known difficulties with the systems reported here (a bug in the recognition network generation has been found) and some known phenomena have not been modeled. We tested our best system-to-date and hope to be able to improve the modeling and cure the system difficulties in the near future.

REFERENCES

1. J.R. Bellegarda and D.H. Nahamoo, "Tied Mixture Continuous Parameter Models for Large Vocabulary Isolated Speech Recognition," Proc. ICASSP 89, Glasgow, May 1989.
2. X. D. Huang and M.A. Jack, "Semi-continuous Hidden Markov Models for Speech Recognition," Computer Speech and Language, Vol. 3, 1989.
3. S. M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," ASSP-35, pp 400-401, March 1987.
4. F. Kubala and R. Schwartz, "A New Paradigm for Speaker-Independent Training and Speaker Adaptation," Proc. DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, June 1990.
5. F. Kubala, S. Austin, C. Barry, J. Makhoul, P. Placeway, R. Schwartz, "BYBLOS Speech Recognition Benchmark Results," Proc. DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, Feb. 1991.
6. P. Ladefoged, *A Course in Phonetics*, Harcourt Brace Javanovich, New York, 1982.
7. K. F. Lee, H. W. Hon, M. Y. Hwang, S. Mahajan, and R. Reddy, "The SPHINX Speech Recognition System," Proc. ICASSP 89, May 1989.
8. M. Lennig, V. Gupta, P. Kenny, P. Mermelstein, D. O'Shaughnessy, "An 86,000—Word Recognizer Based on Phonemic Models," Proc. DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, June 1990.
9. L. M. Norton, M. C. Linebarger, D. A. Dahl and N. Nguyen, "Augmented Role Filling Capabilities for Semantic Interpretation of Spoken Language," Proc. DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, Feb. 1991.
10. D. B. Paul, "The Lincoln Robust Continuous Speech Recognizer," Proc. ICASSP 89., Glasgow, Scotland, May 1989.
11. D. B. Paul, "The Lincoln Tied-Mixture HMM Continuous Speech Recognizer," Proc. DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, June 1990.
12. D. B. Paul, "Experience with a Stack Decoder-Based HMM CSR and Back-Off N-Gram Language Models," Proc. DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, Feb. 1991.
13. D. Rtischev, "Speaker Adaptation in a Large-Vocabulary Speech Recognition System," Masters Thesis, MIT, 1989.
14. R Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul, "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech," Proc. ICASSP 85, Tampa, FL, April 1985.
15. R. Schwartz, personal communication.
16. M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin, and D. Bell, "Linguistic Constraints in Hidden Markov Model Based Speech Recognition," Proc. ICASSP 89, May 1989.

Table 1: SD RM TM-2 XW Semiphone Results

System		States per Phone	Total States	Wd Err
Triphone		0-3-0	24000	1.7% (.13%)
Semiphone		1-1-1	3800	2.2% (.14%)
Semiphone		1-0-2	5500	2.2% (.14%)
Semiphone		2-0-1	5300	2.5% (.15%)
Mixed	wd bdry	1-0-2	9300	2.2% (.14%)
	wd int	0-3-0		

Table 2: Improved Duration Model
RM1 % Word Error Rates (s-d) with WPG

System	Models	Improved Dur Model	
		without	with
TM-2 SD*	XW triphone	1.74% (.13%)	1.55% (.12%)
TM-3 SI-109*	XW triphone	5.64% (.23%)	5.20% (.22%)

* Evaluation test systems

Table 3: New Training Strategy: RM1 Tests Using a TM-2 XW Triphone Systems

System	Training Procedure	Mixture Weights	Gauss	Training	Wd Err (s-d)
SD	old	SD	SD	SD	1.7% (.13%)
MS-12 (SDG)	new	MS	SD	SD-12	2.6% (.16%)
MS-12	old	MS	MS	SD-12	3.4% (.18%)
MS-12 (MSG)	new, opt	MS	MS	SD-12	5.2% (.22%)
SI-109	old	SI	SI	SI-109	7.8% (.27%)
SI-109 (MSG)	new, opt	SI	SI	SI-109	8.6% (.28%)

(Codes: SD=speaker dependent, MS=multi-speaker, SI=speaker independent, -12=all 12 RM1 SD speakers combined, -109=109 RM1 SI training speakers, SDG=SD Gaussians, MSG=MS Gaussians)

Table 4: New Training Strategy: RM2 Tests Using a TM-2 XW Triphone Systems

System	Training Procedure	Mixture Weights	Gauss	Training Set	Wd Err (s-d)
MS-4 (SDG)	new	MS	SD	SD-4	.8% (.14%)
SD	old	SD	SD	SD (RM2)	1.0% (.16%)
MS-4 (MSG)	new,opt	MS	MS	SD-4	1.8% (.21%)
SI-12*	old	SI	SI	SD-12	6.4% (.39%)
SI-12 (SIG)*	new,opt	SI	SI	SD-12	7.0% (.40%)
SI-109	old	SI	SI	SI-109	7.6% (.42%)
SI-109 (SIG)	new,opt	SI	SI	SI-109	8.3% (.44%)

* These systems are the same as the corresponding MS systems in Table 3 but are actually SI in these tests because the test speakers are not in the training set. (-4, -12, and -109 are all disjoint speaker sets.)

(Codes: SD=speaker dependent (2400 training sentences for RM2), MS=multi-speaker, SI=speaker independent, -4=all 4 RM2 speakers combined, -12=all 12 RM1 SD speakers combined, -109=109 RM1 SI training speakers, SDG=SD Gaussians, MSG=MS Gaussians)

Table 5: ATIS Pilot Development Tests: SI, non-cross word triphones, 774 June 90 training sentences

system	opt silences	bigram perplexity	wd err (s-d)
TM-2 triphone	no	23.8	37.5% (1.2%)
TM-2 triphone	yes	23.8	33.3% (1.2%)
TM-2 triphone	yes	17.8	30.9% (1.2%)

Table 6: ATIS Baseline Development Tests: SI, 5020 training sentences, opt silences, perplexity 17.8

system	cross-word models	observation streams	wd err (s-d)
TM-2 triphone	no	2	26.4% (1.1%)
TM-2 triphone*	yes	2	23.0% (1.1%)
TM-3 triphone	yes	3	25.0% (1.1%)
TM-3 semiphone	yes	3	24.0% (1.1%)

* Evaluation test system

Table 7: RM Evaluation Test Results: XW triphones, improved duration model
% Word Error Rates (std dev)

		Word-pair Grammar (p=60)					No Grammar (p=991)*				
System	Training	sub	ins	del	word (s-d)	sent	sub	ins	del	word (s-d)	sent
TM-2	SD	1.0	.1	.7	1.77 (.26)	12.0	5.8	1.3	1.7	8.73 (.55)	44.0
TM-3	SI-109	2.8	.6	1.0	4.39 (.41)	23.3	14.2	2.9	2.7	19.73 (.80)	71.7

* Homonyms equivalent

Table 8: ATIS Baseline Evaluation Test Results: SI, 5120 training sentences
% Word Error Rates with Bigram Back-off Language Model

System	Models	Test Class	Nr Sent	Test Set perplexity	out of vocab wds	sub	ins	del	word (s-d)	sent
TM-2	XW triphone	A	148	22.6	.8%	16.2	5.9	4.0	26.1 (1.1)	88.5
TM-2	XW triphone	D1	58	27.2	1.4%	22.2	3.9	7.1	33.2 (1.9)	88.5
TM-2	XW triphone	A opt	11	73.7	1.4%	22.8	13.1	2.9	38.8 (3.4)	100.0
TM-2	XW triphone	D1 opt	4	23.8	.0%	15.8	21.1	3.5	40.4 (6.5)	100.0
TM-2	XW triphone	all	200	27.5	1.1%	19.1	6.5	4.8	30.4 (1.0)	90.5