SPEECH RECOGNITION IN SRI'S RESOURCE MANAGEMENT AND ATIS SYSTEMS

Hy Murveit, John Butzberger, Mitch Weintraub

SRI International, Menlo Park, CA 94025

OVERVIEW

This paper describes improvements to DECIPHER, the speech recognition component in SRI's Air Travel Information Systems (ATIS) and Resource Management systems. DECIPHER is a speaker-independent continuous speech recognition system based on hidden Markov model (HMM) technology. We show significant performance improvements in DECIPHER due to (1) the addition of tied-mixture HMM modeling (2) rejection of out-of-vocabulary speech and background noise while continuing to recognize speech (3) adapting to the current speaker (4) the implementation of N-gram statistical grammars with DECIPHER. Finally we describe our performance in the February 1991 DARPA Resource Management evaluation (4.8 percent word error) and in the February 1991 DARPA-ATIS speech and SLS evaluations (95 sentences correct, 15 wrong of 140). We show that, for the ATIS evaluation, a well-conceived system integration can be relatively robust to speech recognition errors and to linguistic variability and errors.

Introduction

The DARPA ATIS Spoken Language System (SLS) task represents significant new challenges for speech and natural language technologies. For speech recognition, the SLS task is more difficult than our previous task, DARPA Resource Management, along several dimensions: it is recorded in a noisier environment, the vocabulary is not fixed, and, most important, it is spontaneous speech, which differs significantly from read speech. Spontaneous speech is a significant challenge to speech recognition, since it contains false starts, and non-words, and because it tends to be more casual than read speech. It is also a major challenge to natural language technologies because the structure of spontaneous language differs dramatically from the structure of written language, and almost all natural language research has been focused on written language.

SLS Architecture

SRI has developed a spoken language system (SLS) for DARPA's ATIS benchmark task [1]. This system can be broken up into two distinct components, the speech recognition and natural language components. DECIPHER, the speech recognition component, accepts the speech waveform as input and produces a word list. The word list is processed by the natural language (NL) component, which generates a data base query (or no response). This simple serial integration of speech and natural language processing works well because the speech recognition system uses a statistical language model to improve recognition performance, and because the natural language processing uses a template matching approach that makes it somewhat insensitive to recognition errors. SRI's SLS achieves relatively high performance because the SLS-level system integration acknowledges the imperfect performance of the speech and natural language technologies. Our natural language component is described in another paper in this volume [2]. This paper focuses on the speech recognition system and the evaluation of the speech recognition and overall ATIS SLS systems.

Resource Management Architecture

SRI has also evaluated DECIPHER using DARPA's Resource Management task [3,4]. The system architecture for this task is simply the speech recognition system with no NL postprocessing. There are two language models used in the evaluation: a perplexity 60 word-pair grammar, and a perplexity 1000 all-word grammar. The output is simply an attempted transcription of the input speech.

DECIPHER

This section reviews the structure of the DECIPHER system [5]. The following sections describe changes to DECIPHER.

Front End Analysis

DECIPHER uses an FFT-based Mel-cepstra front end. Twenty-five FFT-Mel filters spanning 100 to 6400 Hz are used to derive 12 Mel-cepstra coefficients every 10-ms frame. Four features are derived every frame from this cepstra sequence. They are

- Energy-normalized Mel-cepstra
- Smoothed 40-ms time derivatives of the Mel-cepstra
- Energy
- Smoothed 40-ms energy differences.

We use 256-word speaker-independent codebooks to vector-quantize the Mel-cepstra and the Mel-cepstral differences. The resulting four-feature-per-frame vector is used as input to the DECIPHER HMM-based speech recognition system.

Pronunciation Models

DECIPHER uses pronunciation models generated by applying a phonological rule set to word baseforms. The techniques used to generate the rules are described in [6] and [5]. These generate approximately 40 pronunciations per word as measured on the DARPA Resource Management vocabulary and 75 per word on the ATIS vocabulary. Speaker-independent pronunciation probabilities are then estimated using these bushy word networks and the forward-backward algorithm in DECIPHER. The networks are then pruned so that only the likely pronunciations remain—typically about 4 per word for the resource management task and 2.6 per word on the ATIS task. This modeling of pronunciation is one of the ways that DECIPHER is distinguished from other HMM-based systems. We have shown in [6] that this modeling reduces error rate.

Acoustic Modeling

DECIPHER builds and trains word models by using context-dependent phone models arranged according to the pronunciation networks for the word being modeled. Models used include unique-phone-in-word, phone-in-word, triphone, biphone, and generalized biphones and triphones, as well as context-independent models. Similar contexts are automatically smoothed together, if they do not adequately model the training data, according to a deleted-estimation interpolation algorithm similar to [7]. The acoustic models reflect both inter-word and across-word coarticulatory effects. Training proceeds as follows:

- Initially, context-independent boot models are estimated from hand-labels in the TIMIT training database.
- The boot models are used as input for a two-iteration context-independent model training run, where context-independent models are refined and pronunciation probabilities are calculated using the full word networks. These large networks are then pruned by eliminating low probability pronunciations.
- Context-dependent models are then estimated from a second two-iteration forward-backward run, which uses the context-independent models and the pruned networks from the previous iterations as input.

ACOUSTIC MODELING IMPROVEMENTS

Tied Mixtures

We have implemented tied-mixture HMMs (TM-HMMs) in the DECIPHER system. Tied mixtures were first described by Huang[9] and more recently in by Bellegarda and Nahamoo[8]. TM-HMMs use Gaussian mixtures as HMM output probabilities. The mixture weights are unique to each phonetic model used, but the set of Gaussians is shared among the states. The tied Gaussians could be viewed as forming a Gaussian-based VQ codebook that is reestimated by the HMM forward -backward algorithm.

Our implementation of TM-HMMs has the following characteristics:

- We used 12-dimensional diagonal-covariance Gaussians. The variances were estimated and then smoothed with grand variances.
- Computation can be significantly reduced in TM-HMMs by pruning either the mixture weights or the Gaussians themselves. We found that shortfall threshold Gaussian pruning--discarding all Gaussians whose probability density of input at a frame is less than a constant times the best probability density for that frame—works as well for us as standard top-N pruning (keeping the N best Gaussians) and requires less computation.
- We use two separate sets of Gaussian mixtures for our TM-HMMs; one for Mel cepstra and one for Mel-cepstral derivatives. We retained our discrete distribution models for our energy features.
- Corrective training [5,10,11] was used to update the mixture weights for the TM-HMMs. The algorithm is identical to that used for discrete HMMs. That is, the mixture weights are updated as if they were discrete output probabilities. No mixture means or variances were corrected.

We evaluated TM-HMMs on the RM task using the perplexity 60 word-pair grammar. Our training corpus was the standard 3990 sentence training set. We used the combined DARPA 1988, February 1989, and October 1989 test sets for our development set. This contains 900 sentences from 32 speakers. We achieved a 6.8 percent word error rate using our discrete HMM system on this test set. The TM-HMM approach achieved an error rate of 5.5 percent. Thus, the TM-HMMs improved word recognition error rate by 20 percent compared to discrete HMMs.

System Type	Word Error (percent)
Discrete DECIPHER	6.8
Discrete+sex separation	6.3
TM-HMM for recognition only	6.4
TM-HMM	5.5
TM-HMM + sex separation	4.9
TM-HMM + corrective training	4.7
TM-HMM +sex +corrective	4.5

TABLE 1. Error rate improvements with TM-HMMs with our 900-sentence RM development set

Male-Female Separation

In the June 1990 DARPA Speech and Natural Language meeting [5], we reported a 20 percent reduction in RM word-error rate by training separate male and female recognizers, decoding using recognizers from both sexes, and then choosing the sex according to the recognizer with the highest probability hypothesis. This improvement was achieved using a recognizer trained on 11,190 sentences. We did not achieve a significant improvement using male-female separation on the smaller 3990 sentence training set. We set out to see, as has been claimed in [8], whether TM-HMMs can take advantage of male-female separation with smaller (3990 sentence) training sets. Our results were mixed. Although performance did improve from 5.5 percent word error with combined models, to 4.9 percent word error with separate male-female models (a 10 percent improvement) we note that 2/3 of the overall improvement was due to the dramatic improvement for speaker HXS. Aside from this one speaker, the performance gain was not significant. Based on our last study, however, we are confident that male-female separation does improve performance with sufficient training data. The table below shows performance for tied-mixture HMMs using combined and sex-separated models.

Name	Standard Models Errs	Wds	%Err	Male-Female Models Errs	Wds	%Err
ESG	2	241	0.83	4	241	1.66
TAB	4	178	2.25	3	178	1.69
CEW	11	241	4.56	5	241	2.07
AJC	10	253	3.95	6	253	2.37
HXS	36	222	16.22	6	222	2.70
DMS	6	179	3.35	5	179	2.79
GMB	3	246	1.22	7	246	2.85
HLM	11	296	3.72	9	296	3.04
BEF	5	226	2.21	7	226	3.10
TJS	9	265	3.40	9	265	3.40
DAS	14	203	6.90	7	203	3.45
JDH	12	246	4.88	9	246	3.66
EWM	12	272	4.41	10	272	3.68
KLS	8	244	3.28	9	244	3.69
DTD	10	233	4.29	10	233	4.29
AEO	9	229	3.93	10	229	4.37
DML	18	272	6.62	12	272	4.41
PGH	13	204	6.37	9	204	4.41
ERS	11	212	5.19	10	212	4.72
GAW	15	244	6.15	12	244	4.92
AEM	8	302	2.65	17	302	5.63
DTB	7	227	3.08	13	227	5.73
CTW	17	253	6.72	15	253	5.93
CMH	18	230	7.83	15	230	6.52
CRZ	23	302	7.62	20	302	6.62
DWA	19	270	7.04	19	270	7.04
CMR	19	231	8.23	17	231	7.36
JDM	16	271	5.90	21	271	7.75
LNS	21	272	7.72	22	272	8.09
GAG	22	296	7.43	24	296	8.11
JWS	16	222	7.21	21	222	9.46
RKM	22	209	10.53	21	209	10.05
AVG	427	7791	5.48	384	7791	4.93

TABLE 2. Performance with and without sex-separation

There was no significant additional gain from using corrective training in addition to male-female separation. Performance improved from 4.9 percent error (male-female only) or 4.7 percent error (corrective training only) to 4.5 percent error (both methods). This lack of further improvement is due to the reduction in training data.

Speaker Adaptation

We have begun experiments into speaker-adaptation, converting speaker-independent models into speaker-dependent ones. Our experiment involved using VQ codebook adaptation via tied-mixture HMMs as proposed by Rtischev [13]. That is, we adjusted VQ codeword locations based on forward-backward alignments of adaptation sentences. However, since we are using a tied-mixture recognition system, we adapted the Gaussian means instead of the codebook.

We selected 21 of the speakers in our development test set for use in an adaptation experiment. We had either 25 or 30 Resource Management sentences recorded for each of these speakers. We chose to use their first 20 sentences for adaptation, and the other 5 or 10 sentences for adaptation testing.

Using our original TM-HMM models, we achieved an error rate of 7.4 percent (114 errors in 1541 reference words) on this adaptation test set. After adjusting means for each speaker using the 20 adaptation sentences, we achieved an error rate of 6.1 percent (94 errors in 1541 reference words) on the adaptation test sentences.

This improvement with adaptation leads to performance that is still quite short of speaker-dependent accuracy (the ultimate goal of adaptation). Thus, it does not seem worth the added inconvenience of obtaining 20 known sentences from a potential system user, though it is promising for on-line adaptation. We plan to look into several areas for further improvement. For example:

1. Rtischev et al. [14] have shown that adapting mixture weights is at least as important as adapting means.
2. Kubala [15] et al. have shown that adapting speaker-dependent models can be superior to adapting from speaker-independent models.
3. It is possible that the adaptation sentences need not be supervised given the relatively good (7.4 percent error) initial performance.

Rejection of Out-of-Vocabulary Input

We implemented a version of DECIPHER that rejects false input as well as recognizing legal input (our standard recognizer attempts to classify all the input). In addition to standard word models, it uses an out-of-vocabulary word model to recognize the extraneous input. The word model has the following pronunciation network similar to [17].

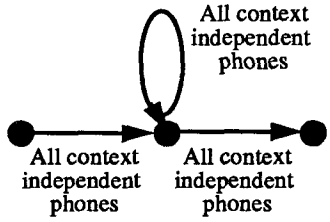

FIGURE 1. Out-of-vocabulary word model

There are 67 phonetic models on each of the arcs in the above word network. All phonetic transition probabilities in this word network are equal, and are scaled by a parameter that adjusts the amount of false rejection vs. false acceptance.

Thus far, we have performed a pilot study that shows this method to be promising. We gathered a database of 58 sentences total from six people. About half of the sentences are digit strings and the other half are digits mixed with other things. There are a total of 426 digits in the database, and 176 additional non-digit words. Example sentences are outlined in Table 3.

We considered correct recognition for these sentences to be the digits in the string without the rest of the words (i.e. 2138767287, 3876541104, 33589170429 are the correct answers for the top three sentences in Table 3).

We trained a digit recognizer with rejection from the Resource Management training set and achieved a word error rate of 5.3 percent for the 27 sentences that contained only digits (13 errors = 1 insert 3 delete 9 subs in 243 reference words), which is within one error of the system without rejection. Thus, in this pilot study, using rejection didn't hurt performance for "clean" input. The overall error rate was 11.7 percent (26 inserts 15 deletes 9 subs in 426 reference words). That is, 402 of 426 digits were detected, and at least 141 of the 176 extraneous words were rejected.

my parents number is 2 1 3 um 8 7 6 ok 7 2 8 7
if you have questions please dial extension 3 8 7 6 at 5 4 1 1 oh 4
please call 3 3 5 8 9 1 um 7 oh 4 2 9
hmm let's see what's this 1 2 3 4 5 uh that's not right 2 3 4 5
1 2 3 oh no that's wrong 2 4 5 8 9 yeah i think that's it
this is a test 1 2 3 4 5 8 7 this was only a test
<grunt> 1 2 <cough> 3 4 5 <sneeze> 8 7 <mic-noise>
4 1 dollars and 3 1 8 cents
what's this oh 4 1 0 8
well let's see 3 1 4 7 8 ok

TABLE 3. Sample sentences for the rejection study

LANGUAGE MODELING

Bigram Language Modeling

We used a bigram language model to constrain the speech recognition system for the ATIS evaluation. A back-off estimation algorithm [16] was used for estimation of the bigram parameters. The training data for the grammar consisted of 5,050 sentences of spontaneous speech from various sites—1,606 from MIT's ATIS data collection project, 774 from NIST CD-ROM releases, 538 from SRI's ATIS data collection project, and 2,132 from various other sites.

Robust estimates for many of the bigram probabilities cannot be achieved since the vast majority of them are seen very infrequently (because of the lack of sufficient training data). Furthermore, frequencies of words such as months and cities were biased by the data collection scenarios and the time of year the data was collected. To reduce these effects, words with effectively similar usage were assigned to groups, and instead of collecting counts for the individual words, counts were collected for the groups. After estimation of the bigram probabilities, the probabilities of transitioning to individual words were assigned the group probability divided by the number of words in the group. This scheme not only reduced some of the problems due to the sparse training data, but also allowed some unseen words (other city names, restriction codes, etc.) to be easily added to the grammar. The table below contains the groups of words tied together.

months, days, digits, teens, decades, date-ordinals, cities, airports, states, airlines, class-codes, restriction-codes, fare-codes, airline-codes, aircraft-codes, airport-codes, other-codes

TABLE 4. Tied Groups

Using our back-off bigram on our ATIS development set (most of the June 1990 DARPA-ATIS test set), we achieved a 14.1 percent word error rate with a test-set perplexity of 19 (not counting 6 words not covered by the grammar). When we applied this grammar to the February 1991 ATIS evaluation test set (200 sentences) the perplexity was 43, excluding 26 instances of words not covered in our vocabulary. For the 148 Class A sentences, the recognition word error rate was 17.8 percent.

We also explored various class-grammar implementations. These grammars were generated by interpolating word-based bigrams with class-based bigrams. We were able to vary the grammars and their perplexities by varying the interpolation coefficients. However, recognition performance never improved over that for the back-off bigram. In fact, accuracy remained relatively constant throughout a large range of perplexities.

Table 5 illustrates recognition accuracy using bigrams with different perplexities on our ATIS development test set. A preliminary set of models was used for recognition (with 442 words in the vocabulary) and the grammars were estimated using 2,909 sentences.

	Perplexity	Word Error (percent)
Backed-off Bigram	19	14.1
Interpolated Bigrams	20	14.5
	24	15.3
	71	14.9
	89	14.7
	91	14.5
	113	14.9
	442	29.2

TABLE 5. Perplexity vs. word error on the ATIS development set

These tables also illustrate that recognition performance did not depend strongly on the test-set perplexity. Clearly, other factors are dominating performance. We believe that one of our most pressing needs in this research is to understand what this bottleneck is, and to develop ways that express it better than perplexity.

Multi-Word Lexical Units

Many words occur with sufficient frequency and with significant cross-word coarticulation that a better acoustic model might be made by training these word combinations as a single word model. These words include "what-are-the," "give-me," etc., which can have a variety of pronunciations best modeled with a network of phones representing the phonetic and phonological variation of the whole sequence ("what're-the," "gimme," etc.) instead of each word separately.

Also, when considering class grammars, multiple word sequences allow classes which could not be constructed by considering every word separately. For instance, having distinct models of all the restriction codes (e.g. "v-u-slash-one") might be more appropriate than modeling *alpha->alpha->slash->number* in the bigram. The latter form would allow all the alphabet letters to transition to all the alphabet letters, with probabilities as prescribed by the bigram, and would incorrectly increase the probability for invalid restriction codes.

This multi-word technique allows all the probabilities of all the restriction codes to be tied together, so that all are equally covered at the appropriate place in the grammar, instead of depending completely on the individual words' statistics estimated from sparse training data. The multi-word approach resulted in only a slight performance improvement compared to a system where non-coarticulatory multi-words were left separated. That is, for the "separate words" system, words like "a p slash eighty" were separate words, but coarticulatory word models like "what-are-the" and "list-the" were retained. On a

119-sentence subset of the June 90 evaluation set, the results were as shown in Table 6.

Development Set Performance

	Perplexity	Word Error (percent)
Multi-Word	26	9.6
Separate Words	20	10.7

February 1991 Class-A Evaluation Performance

	Perplexity	Word Error (percent)
Multi-Word	43	17.8
Separate Words	34	18.3

TABLE 6. Effectiveness of multi-word modeling

Note that the higher perplexity of the multi-word system is deceiving since high probability grammar transitions are now hidden within the multi-word models, and are not seen by the grammar. Tables 7 and 8 list the various multi-word units.

flights-from, what-is-the, show-me-the, show-me-all, show-me, how-many, one-way, what-are-the, give-me, what-is, i-would-like, i'd-like-to, what-does

TABLE 7. Coarticulatory Multi-Words

CITIES:	san-francisco, washington-d-c, ...
AIRLINES:	a-l, c-o, t-w-a, u-s-air, ...
AIRCRAFT:	d-c-ten, seven-forty-seven, ...
AIRPORTS:	a-t-l, b-o-s, s-f-o, d-f-w, ...
CLASS CODES:	q-x, f-y-b-m-q, k-y, y-n, ...
RESTRICT CODES:	a-p-eighty, a-p-slash-eighty,...
COLUMN HEADS:	d-u-r-a, e-q-p, r-t-n-max, ...

TABLE 8. Semantic Multi-Words

EVALUATION

RM Evaluation

SRI evaluated the DECIPHER system on DARPA's February 1991 speaker-independent test set. The characteristics of the evaluated system were:

- Speaker-independent recognition
- 3990 sentence DARPA-RM training
- 3 state, left-to-right, context-dependent hidden Markov model using deleted-interpolation estimation of parameters
- Input features were 12 Mel-cepstra and delta-Mel-cepstra and scalar quantized energy and delta-energy
- Tied-mixture modeling for Mel cepstra and delta-Mel-cepstra
- 256 diagonal covariance Gaussians for each
- Independent discrete density HMM models for energy and delta energy
- Multiple pronunciation trained phonological modeling, about 4 pronunciations per word on average
- Cross-word acoustic and phonological modeling
- Sex-consistent modeling
- Corrective training on mixture weights
- Resource Management all-word and word-pair grammars used with 992-word Resource Management vocabulary.

We achieved the performance shown in Table 9.

Speaker	P=60	P=1000
ALK03	9.7	20.8
CAL15	2.5	11.9
CAU07	2.6	14.7
EAC02	10.2	22.0
JLS04	1.6	11.1
JWG05	7.5	19.5
MEB03	2.9	17.6
SAS05	2.2	10.4
STK01	4.1	21.2
TBR01	5.2	27.8
Average	**4.8**	**17.6**

TABLE 9. DARPA-RM February 1991 speaker-independent evaluation

Our performance is severely limited by training data[5], and many further improvements for the RM task may only be ways to work around RM's artificial limit on training data. Thus, we expect to develop and evaluate our system in the future with the ATIS task which both has more training data available and uses more realistic (spontaneous) speech.

SLS Evaluation

We evaluated on DARPA's February 1991 ATIS test set using a system similar to the one described above except:

- The system was trained on 17,042 sentences (3990 RM-SI, 4200 TIMIT, 7932 read ATIS, 920 spontaneous ATIS).
- 1,139 word vocabulary (the test set vocabulary was not revealed in advance) using multi-word units.
- Discrete distribution HMM modeling was used for all features.
- A back-off bigram language model [16] with tied word-groups was used, with a test set perplexity of 43 (not counting 26 words out of vocabulary).
- A template-matcher natural language component [2] was used to generate ATIS database queries based on the speech recognition output.

We achieved the performance shown in Table 10.

SPKR	Corr	Sub	Del	Ins	Err	Sent Err
CL	93.6	5.1	1.3	1.7	8.1	42.3
CJ	92.0	6.9	1.0	0.7	8.7	46.2
CO	92.0	3.7	4.3	1.2	9.3	56.2
CP	90.7	7.5	1.8	2.5	11.8	59.3
CK	83.3	8.8	7.8	1.0	17.6	58.3
CH	84.2	5.3	10.5	5.3	21.1	100.0
CE	81.5	12.0	6.5	3.2	21.8	70.0
CI	73.1	24.0	2.9	5.8	32.7	90.0
CM	75.0	23.5	1.5	26.5	51.5	100.0
Average	**86.5**	**10.3**	**3.1**	**4.3**	**17.8**	**60.1**
All-word (Perplexity 1139)						
Average	**86.5**	**23.9**	**3.7**	**8.0**	**35.5**	**91.2**

TABLE 10. DARPA-ATIS February 1991 speech evaluation 148 Class A Sentences

SPKR	Corr	Sub	Del	Ins	Err	Sent Err
CJ	91.9	6.5	1.6	0.8	8.9	54.5
CP	91.7	6.6	1.7	1.7	10.0	55.2
CL	91.4	6.7	1.9	1.9	10.4	44.8
CK	85.0	8.7	6.3	0.5	15.5	64.0
CE	83.0	11.8	5.2	2.6	19.6	73.9
CO	79.4	13.7	6.9	1.4	22.0	75.9
CH	78.6	13.1	8.3	3.6	25.0	100.0
CI	67.1	27.3	5.6	5.6	38.6	92.9
CM	72.5	25.2	2.3	23.9	51.4	100.0
Average	**83.5**	**12.6**	**3.9**	**4.2**	**20.7**	**66.5**

TABLE 11. DARPA-ATIS February 1991 speech evaluation All sentences

As can be seen, speakers CI and CM contributed significantly to the overall error rate. Furthermore, many of the errors occurred despite their relatively small bigram probabilities, indicating that the grammar is still not completely effective in overriding poor acoustic matches.

Table 12 describes overall spoken language system performance.

System	Right	Wrong	NA[1]	WErr[2]	Score[3]
NL Only	109	9	27	31.0	69.0
SLS	96	11	38	41.4	58.6

TABLE 12. DARPA-ATIS February 1991 SLS evaluation 148 Class A sentences

Discussion

The most interesting result of this evaluation (see the paper by Pallett in this proceedings) was that, though SRI along with BBN achieved the best speech recognition accuracy, and SRI along with CMU had the best natural-language-only performance, the accuracy of SRI's combined speech and natural language systems was far better than that for the other sites. We attribute this to the error tolerant nature of our speech/natural-language interface. For instance, note that performance using spoken language is not much worse than the performance of the NL component given transcribed input (i.e. given a perfect speech recognition component) even though the SLS speech recognition component had a 60 percent sentence error rate (at least one word was wrong in 60 percent of the sentences).

The above results indicate to us that steady progress in the speech recognition and natural language technologies, together with error-tolerant speech/natural-language interfaces can lead to practical spoken language systems in the near future.

1. NA is no answer
2. WErr or weighted error is percent no answer plus two times the percent wrong.
3. Score = 100 - Werr

REFERENCES

1 Price, P., "The ATIS Common Task: Selection and Overview," *Proceedings DARPA Speech and Natural Language Workshop*, June 1990.

2 Jackson, E., D. Appelt, J. Bear, R. Moore, and A. Podlozny, "A Template Matcher for Robust NL Interpretation," *Proceedings DARPA Speech and Natural Language Workshop*, June 1991.

3 Pallet, D., "Benchmark Tests for DARPA Resource Management Database Performance Evaluations," *Proceedings ICASSP-89.*

4 Price, P., W.M. Fisher, J. Bernstein, and D.S. Pallet, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," *Proceedings ICASSP-88.*

5 Murveit, H., M. Weintraub, M. Cohen, "Training Set Issues in SRI's DECIPHER Speech Recognition System," *Proceedings DARPA Speech and Natural Language Workshop*, June 1990.

6 Cohen, M., H. Murveit, J. Bernstein, P. Price, M. Weintraub, "The DECIPHER Speech Recognition System," *Proceedings ICASSP*, April 1990.

7 Jelinek, F. and R. Mercer, "Interpolated Estimation of Markov Source Parameters from Sparse Data," pp. 381-397 in E.S. Gelsima and L.N. Kanal (editors), *Pattern Recognition in Practice*, North Holland Publishing Company, Amsterdam, The Netherlands.

8 Bellegarda, J., D. Nahamoo, "Tied Mixture Continuous Parameter Modeling for Speech Recognition," *IEEE Trans. ASSP*, December 1990.

9 Huang, X.D.,"Semi-continuous hidden Markov models for speech recognition," *Computer Speech and Language*, 3 pp. 239-251 (1989)

10 Bahl, L., P. Brown, P. De Souza, and R. Mercer, "A New Algorithm for the Estimation of Hidden Markov Model Parameters," *Proceedings ICASSP-88.*

11 Lee, K-F., and S. Mahajan, *Corrective and Reinforcement Learning for Speaker-Independent Continuous Speech Recognition*, Technical Report CMU-CS-89-100, Carnegie Mellon University, January 1989.

12 Huang, X., F. Alleva, S. Hayamizu, H.-W. Hon, M.-Y. Hwang, and K.-F. Lee, "Improved Hidden Markov Modeling for Speaker-Independent Continuous Speech Recognition," *Proceedings DARPA Speech and Natural Language Workshop*, June 1990.

13 Rtischev, Dimitry, *Speaker Adaptation in a Large-Vocabulary Speech Recognition System*, Master's Thesis, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

14 Rtischev, D., D.. Nahamoo, and M. Picheny, "Speaker Adaptation via VQ Prototype Modification," submitted to *IEEE Trans. Signal Processing*.

15 Kubala, Francis, Richard Schwartz, and Chris Barry, "Speaker Adaptation from a Speaker Independent Training Corpus," *Proceedings ICASSP-90*.

16 Katz, S., "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. ASSP*, March 1987.

17 Asadi, A., R. Schwartz, and J. Makhoul, "Automatic Detection of New Words in a Large Vocabulary Continuous Speech Recognition System," *Proceedings DARPA Speech and Natural Language Workshop*, Oct. 1989.