IEEE Instrumentation and Measurement
Technology Conference
Budapest, Hungary, May 21–23, 2001

Evolution of the performance of automatic speech recognition algorithms in transcribing conversational telephone speech

M. Padmanabhan G. Saon, G. Zweig, J. Huang, B. Kingsbury and L. Mangu
IBM T. J. Watson Research Center
Yorktown Heights, NY, 10598, USA
Phone: (914) 945-2929, email: (mukund,gsaon,zweig,yellow,bedk,mangu)@us.ibm.com

Abstract – [1] ***Research in the speech recognition speech-to-text conversion) area has been underway for a couple of decades, and a great deal of progress has been made in reducing the word error rate (WER). In this paper, we attempt to summarize the state of the art in speech recognition algorithms. The algorithms we describe span the areas of lexicon design, feature extraction, classifier design, combination of hypotheses, and speaker adaptation of acoustic models. We will benchmark the algorithms on two main sources of speech, the first being Voicemail (conversational telephone speech from a single speaker) and the second being Switchboard (conversational telephone speech between two speakers). We also present the results of some cross-domain experiments which highlight the "brittleness" of speech recognition systems today and illustrates the need to focus research effort on improving cross-domain performance.***

Keywords – ***Speech Recognition, Spontaneous speech, Telephone speech, Discriminant Transforms, Boosting, Consensus, Formant frequencies, Spectral peaks***

I. INTRODUCTION

Research in the speech recognition area has been underway for a couple of decades, and a great deal of progress has been made in reducing the word error rate (WER) on some specific categories of speech such as Broadcast News and Switchboard transcription and Voicemail transcription. For instance, the WER on the DARPA sponsored Voicemail transcription task has dropped by more than a factor of two over the last three years. Further, the basic techniques that are used in most of these tasks are similar. In this paper, we attempt to summarize the state of the art in speech recognition algorithms. We will benchmark the algorithms on two main sources of speech, the first being Voicemail (conversational telephone speech from a single speaker) and the second being Switchboard (conversational telephone speech between two speakers).

A number of algorithms were developed in the context of these tasks that contributed significantly to reducing the WER. In the following sections, we describe some of these algorithms, spanning the areas of lexicon design, feature extraction, classifier design, combination of hypotheses, and speaker adaptation of acoustic models. These algorithms were instrumental in reducing the word error rate on Voicemail data to around 28%. Further, though Voicemail and Switchboard both represent spontaneous conversational telephone speech, we will show that there are significant differences between the two. For instance, systems trained on one of these databases do not provide good performance on the other database. The "brittleness" of speech recognition systems today is apparent in the results of some of the cross-domain experiments we describe, and it illustrates the need to focus research effort on improving cross-domain performance.

II. BACKGROUND

A. Training/Test data

Voicemail

The Voicemail training database comprises 70 hours of speech, which corresponds to approximately 700k words of text. We will refer to this training database as T-VM1. The size of the testing vocabulary is 11k words. The development test set for this database comprises 43 messages (D-VM) and the evaluation test set (E-VM) comprises 86 messages. The language model is a trigram built from the 700k words of text.

Switchboard

We used 2378 of the 2438 Switchboard I conversations [1] as our training set, and the 19 conversations used in the 1997 Johns Hopkins Workshop as the test set. This training set represents around 250 hours of speech and 2 million words of text. We will refer to this training database as T-SWB1 and to the test database as E-SWB. The size of the vocabulary used for testing was 18k words. The language model is a trigram built from the 1.2 million words of text.

B. System Description

The speech recognition system uses a phonetic representation of the words in the vocabulary. Each phone is modelled with a 3-state left-to-right HMM. Further, we identify the variants of each state that are acoustically dissimilar by asking questions about the phonetic context in which the state occurs. The questions are arranged hierarchically in the form of a decision tree, and its leaves correspond to the basic acoustic units that we model. For futher details, see [2]. A feature vector is extracted

[1] ***We would like to acknowledge the support of DARPA under Grant MDA972-97-C-0012 for funding part of this work.***

0-7803-6646-8/01/$10.00 ©2001 IEEE

System	FSP	D	#L	#G	Trg
S-VM1 f433	Ceps	39	2313	134k	T-VM1
S-VM2 f708	Proj (1)	39	2313	134k	T-VM1
S-VM3 f708	Proj (2)	39	2313	134k	T-VM1
S-VM4 f608	Ceps	39	2313	36k	T-VM1
S-VM5 f526	MSG	26	3527	154k	T-VM1
S-VM6 f8101	Ceps	39	2307	130k	T-VM1
S-VM7 f844.v60	Ceps	39	2778	279k	T-VM1
S-VM8 f844.v70	Proj (1)	39	2778	279k	T-VM1
S-SWB1 f844.v28	Ceps	39	3140	277k	T-SWB1
S-SWB2 f901	Proj (3)	60	3140	277k	T-SWB1
S-VM9 f844.v50	Ceps	39	2778	260k	T-VM1 + T-SWB1
S-VM10 f844.v40	Proj (4)	39	2778	260k	T-VM1 + T-SWB1

TABLE I

SYSTEMS DESCRIPTION: FSP INDICATES THE TYPE OF FEATURE SPACE, D INDICATES THE DIMENSIONALITY OF THE SPACE, $\#L$ INDICATES THE NUMBER OF LEAVES IN THE DECISION TREE, $\#G$ INDICATES THE NUMBER OF GAUSSIANS, AND Trg INDICATES THE TRAINING DATA THAT WAS USED TO BUILD THE SYSTEM

every 10 ms, and we model the pdf of the feature vector for each leaf of the decision tree with a mixture of gaussians. The baseline feature vector is the Mel cepstrum augmented with its 1st and 2nd temporal derivatives (which we refer to as deltas). We will refer to this as the cepstral feature space. Some of the systems that we experimented with spliced together 9 frames of cepstra (the cepstra at the current frame and 4 frames before and after the current frame) and projected the spliced feature vector down to a lower dimension by means of a linear transform. We will refer to this feature space as the projected feature space.

During the course of running these experiments, we built a number of "baseline" acoustic models using both Voicemail and Switchboard training data. The improvements accruing from specific algorithms are benchmarked on these baseline systems. We summarize the models that we worked with in Table I.

III. LEXICON DESIGN

One observation in connection with voicemail data is that crossword co-articulation is very common in this data because of the casual nature of the speech and the fast speaking rate. For instance, the phrase 'going to take' would often be pronounced as 'gontake = G AO N T AE KD'. We chose to model such effects by constructing compound words. A similar technique was used in [4] to obtain performance improvements on the Switchboard task - our work differs from [4] in the measure used to select the compound words.

System	It.	Nb.	Examples	D-VM
S-VM1	0	0		34.7%
S-VM1	3	70	AREA-CODE, GIVE-ME, A-CALL, E-MAIL, TAKE-CARE GIVE-ME-A-CALL, LET-ME-KNOW, AS-SOON-AS, THANK-YOU-VERY-MUCH TALK-TO-YOU-LATER-BYE, THANKS-A-LOT, PLEASE-GIVE-ME-A-CALL	32.3%

TABLE II

WORD ERROR RATES FOR LM MEASURE

The use of these compound words serves multiple purposes. First, it is generally the case that decoding errors are more common in shorter words, hence, as the compound words have relatively long baseforms, there are fewer errors in the compound words. Second, as mentioned earlier, they enable the modelling of cross-word co-articulation effects. Third, stereotypical phrases such as "hi-this-is" are very common in Voicemail, and can serve as a trigger for detecting quantities of interest such as names.

We present a novel algorithm for automatically selecting compound words from a training corpus. The algorithm ranks all pairs of words in the corpus on the basis of a linguistic measure, and makes the highest ranking pairs compound words. The measure is defined in terms of the *direct bigram* probability between the words w_i and w_j, $P_f(W_{t+1} = w_j|W_t = w_i)$, and a *reverse bigram* probability between the words as $P_r(W_t = w_i|W_{t+1} = w_j)$.

The measure that we introduced was the geometrical average of the direct and the reverse bigram:

$$LM_2(w_i, w_j) = \sqrt{\hat{P}_f(w_j|w_i)\hat{P}_r(w_i|w_j)} = \frac{P(w_i, w_j)}{\sqrt{P(w_i)P(w_j)}} \quad (1)$$

A. Results

This measure was applied iteratively to the corpus resulting in an increasing number of compound words per iteration. Table II summarizes the total number of new compound words, examples of such words, the word error rate.

In summary, it may be seen that adding compound words based on the LM measure results in a 7 % relative improvement in the word error rate. This vocabulary (with compound words) and the associated trigram LM will be used in all Voicemail related experiments in subsequent sections.

IV. FEATURE EXTRACTION

In this section, we report on the results of experiments in feature extraction. As mentioned earlier, most systems extract Mel cepstra every 10 ms from the sampled speech. Though the Mel cepstra are perceptually motivated, they do not explicitly attempt to discriminate between different phonetic classes. Further, it is possible to augment the Mel cepstra by using additional knowledge related to the speech production process in the hope that this will better help discriminate between phonetic classes. In this section, we describe a process of computing a linear transformation on the Mel cepstra that separates the phonetic classes out. We also describe the utility of adding spectral peak related information to the Mel cepstra.

A. Linear transformations of the feature space

In this subsection, we report on experiments related to designing a linear transformation that can be applied on the Mel cepstra to better discriminate between phonetic classes. Linear discriminant analysis is a standard technique for dimensionality reduction with minimal loss of discrimination information. However, the LDA formulation makes certain assumptions that are not true. Chief among these is the assumption that all the classes have the same covariance matrix. A second assumption is that the classes are modelled with full covariance gaussians (an assumption that is not true in most speech recognition systems). We experimented with two variants of LDA as described below.

A.1 Maximum likelihood discriminant (MLD) transformation

Let $\{x_i\}_{1 \leq i \leq N}$ denote a sequence of D dimensional feature vectors, where each of the vectors belongs to a single class $j \in \{1, \cdots, J\}$. Let N_j, μ_j, Σ_j denote the sample count, mean and covariance of the j^{th} class. The class information may be condensed into two matrices called

$$\text{within-class scatter:} \quad W = \frac{1}{N}\sum_{j=1}^{J} N_j \Sigma_j$$

$$\text{between-class scatter:} \quad B = \frac{1}{N}\sum_{j=1}^{J} N_j \mu_j \mu_j^T - \bar{\mu}\bar{\mu}^T$$

The LDA objective function tries to find a PxD projection, θ, such that the ratio of the following determinants is maximized

$$J(\theta) = \frac{|\theta B \theta^T|}{|\theta W \theta^T|} \qquad (2)$$

However, the assumption of equal class covariances in LDA can lead to a serious degradation in classification performance In [5] we presented a HDA formulation that modified the LDA objective function (2) to take into account the different covariance matrices of the different classes. The modified objective function is given by

$$\frac{|\theta B \theta^T|^N}{\prod_{j=1}^{J} |\theta \Sigma_j \theta^T|^{N_j}} \qquad (3)$$

System	Feature space	D-VM	E-VM
S-VM1	Cepstra + deltas	32.3 %	39.6 %
S-VM2	HDA+MLLT	30.2 %	35.3 %

TABLE III
WORD ERROR RATE FOR CEPSTRA, HDA+MLLT, AND DHDA FEATURES

and taking the log of the above objective yielded the HDA objective function

$$H(\theta) = \sum_{j=1}^{J} -N_j \log |\theta \Sigma_j \theta^T| + N \log |\theta B \theta^T| \qquad (4)$$

This objective function does not yield a closed form solution as for the case of LDA, however, it may be optimized using non-linear optimization techniques.

The discrimination between classes provided in the HDA feature space requires the use of full covariance gaussian models for the classes. This is generally too computationally expensive to be practical in most speech recognition systems, consequently, the models are replaced with gaussians that have diagonal covariances. If the HDA feature space is characterized by dimensions that are highly correlated, the modeling approximation inherent in the diagonal covariance assumption negates any beneficial effect that the HDA may have. Consequently, we applied a further transformation (MLLT) that tries to diagonalize the HDA feature space [7]. The application of this transform does not change the HDA objective function value. We refer to this final feature space as the HDA+MLLT space.

A.2 Results

The word error rates obtained on the development and evaluation voicemail test sets for the cepstral and projected feature spaces are shown in Table III. The acoustic models were described in Section II-B, and the language model and vocabulary were described in Section III-A. In summary, the HDA+MLLT space is seen to provide a relative improvement of 10-15% over the baseline cepstral space.

B. Augmenting cepstra with spectral peak information

One of the most commonly used acoustic observations are the Mel cepstra, which are extracted from the speech signal every 10 ms. The Mel cepstra are based on perceptual studies and attempt to emulate the way in which the human auditory system works. It is possible to augment this information further by incorporating additional knowledge about the speech production process into the process of feature extraction. One source of information is represented in the spectral peak trajectories of speech. In this section, we attempt to add information related to spectral peak trajectories and energies to the baseline Mel cepstral observations. A similar idea was proposed in [8] with the objective of providing robustness to noise. However, no attempt was made in [8] to quantify the amount of information provided by the new features. This analysis [6] (and the subsequent speech recognition experiments) indicate that the useful information is not in the locations of the spectral peaks, but rather in the energy at those spectral peaks.

B.1 Tracking spectral peaks

We experimented with extracting features that track the spectral peak locations in some predefined bands, and quantified the amount of information contained in these new features over and above the cepstra. Subsequently, we incorporated these new features into the speech recognition system using feature fusion, and obtained experimental results that indicate an improvement in the word error rate due to the addition of these features. The spectral peaks were obtained by first bandpass filtering the speech signal using two bandpass filters with passbands of 250-750 Hz and 850-2300 Hz (we assumed that there was only one dominant spectral peak in each bandpass signal) [2]. Subsequently we used an adaptive filter [10] to isolate the spectral energy peaks in each of the filtered signals.

B.2 Results

We incorporated the new features into the speech recognition system using feature fusion i.e., the cepstral features in the speech recognition system were augmented with the new features. The new features were specifically either the frequency estimate of the peaks, or the energy at these peaks. We further augmented the "fused" feature with its first and second temporal derivatives. The word error rate results computed on the D-VM test set are summarized in Fig 1. The x-axis indicates the dimensionality of the extracted feature (either number of cepstra, or number of cepstra + e_1, e_2, or number of cepstra + s_1, s_2).

The acoustic models were described in Section II-B, and the language model and vocabulary were described in Section III-A. For reasons of quick turnaround time, the S-VM4 system, which represents a smaller version of the S-VM1 system, was used as the baseline for these experiments. The figure shows that the (e_1, e_2) estimates do contain more information than the higher order (13th) cepstra and can be used to improve the performance of the system (by 5.7% on the dev test and 5.2% on the eval test).

V. CLASSIFIER DESIGN

The basic speech recogntion problem could be interpreted as a classification problem, where the goal is to predict the class corresponding to an acoustic observation. In most instances, this problem is converted to a likelihood problem by the application of Bayes' rule, which requires the evaluation of the probability of an acoustic observation belonging to a class. The classes generally correspond to context dependent phonetic states. The probability density of the observations for each of these classes is very often modeled using mixtures of multi-dimensional gaussians. In this section, we look at how to improve the performance of a classifier based on mixtures of gaussians by applying an iterative scheme that successively focuses on the regions of the acoustic space that are difficult to classify.

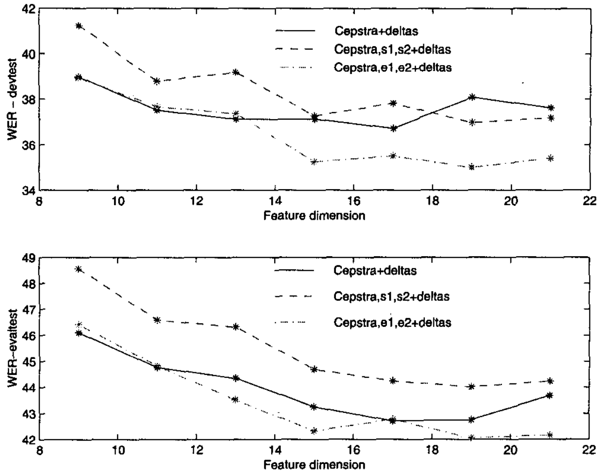

Fig. 1. Word error rate vs feature dimension

Boosting is a technique for sequentially training and combining a collection of classifiers in such a way that the later classifiers make up for the deficiencies of the earlier ones. Many variants exist [12] but all follow the same basic strategy. There is a sequence of iterations, and at each iteration a new classifier is trained on a weighted set of the training examples. Initially, every example gets the same weight, but in subsequent iterations, the weights of hard-to-classify examples are increased relative to the easy ones. The outputs of the classifiers are then combined in such a way as to guarantee certain bounds on both training and testing error [12]. We report results here using an extension to Adaboost that was presented in [11] and that allows for large speedups in training time. The extension was motivated by the scale of the problem, where we have tens of millions of labeled training pairs, thousands of classes, and hundreds of thousands of gaussians that model the probability density of the classes.

A. Results

The experimental results obtained by boosting the system are summarized in Table IV for the E-VM test set. The starting point was the S-VM1 system described in Section II-B, and the

[2] These passband ranges are motivated by physio linguistic observations [9] that state that the spectral peaks in the speech signal correspond to formants, and the range of movement of the first two formants are respectively 250-750 Hz and 850-2300 Hz for the average American speaker.

E-VM				
1st It.	2nd	3rd	4th	5th
39.6	39.5	39.2	39.1	38.9

TABLE IV

WORD ERROR RATE FOR DIFFERENT ITERATIONS OF BOOSTING

	D-VM		E-VM	
System	Baseline	Consensus	Baseline	Consensus
S-VM2	30.2 %	28.9 %	35.2%	33.8%
S-VM6	33.7 %	31.2 %	39.4%	37.7%
S-VM5	42.4 %	41.6 %	47.7%	46.9%
Rover	29.2 %	28.5 %	34.2%	33.3%

TABLE V

WORD ERROR RATES FOR VARIOUS SYSTEMS USING 1-BEST AND CONSENSUS HYPOTHESIS

System	Training	Test	
Cross domain-Cepstral feature space			
		E-VM	E-SWB
S-VM7	T-VM1	39.5 %	62.2 %
S-SWB1	T-SWB1	53.5 %	45.8 %
Cross domain - Projected feature space			
S-VM8	T-VM1	36.3 %	57.3 %
S-SWB2	T-SWB1	46.8 %	38.5 %
Joint Training - Cepstral feature space			
S-VM9	T-VM1 + T-SWB1	41.7 %	48.7 %
Joint Training - Projected feature space			
S-VM10	T-VM1 T-SWB1	36.9 %	44.7 %

TABLE VI

WER PERFORMANCE FOR CROSS-DOMAIN CONDITION

language model and vocabulary were described in Section III-A. The word error rate numbers indicate a small but consistent improvement with an increasing number of iterations.

VI. COMBINING MULTIPLE HYPOTHESES

The most commonly used decoding paradigm for speech recognition is the maximum a-posteriori (MAP) rule which is used to guide the hypothesis search.

$$\underline{w}^* = argmax_{\underline{w}} p(\underline{w}/\underline{y}) = argmax_{\underline{w}} p(\underline{y}/\underline{w}) p(\underline{w})/p(\underline{y}) \quad (5)$$

where $\underline{w}$ represents the sequence of decoded words and $\underline{y}$ denotes the acoustic observations corresponding to the sentence. In [13] a novel decoding rule was applied to a word lattice (that was produced by a MAP decoder) to obtain a "consensus hypothesis" as follows: the word lattice (graph) produced by a MAP decoder is first converted into a chain-like structure by merging different paths in the graph. The components of the chain represent parallel sequences of words. The criterion for merging two paths in the graph is related to the time overlap between the paths and the phonetic similarity between the word sequences in the two paths. The decoding rule was equivalent to picking the most probable word in each component. The concatenation of these words represents the consensus hypothesis. Further details are given in [13].

A. Results

We evaluated the performance of this technique on the E-VM test set with a number of systems (denoted S-VM2, S-VM5, S-VM6). The acoustic models were described in Section II-B, and the language model and vocabulary were described in Section III-A. Subsequently, we combined the consensus hypotheses of these three systems using ROVER [14]. The results are presented in Table V (baseline results refers to the 1-best hypothesis of the corresponding system) and show a consistent improvement (of approximately 3% relative) by using the consensus hypothesis rather than the 1-best hypothesis.

VII. CROSS-DOMAIN EXPERIMENTS

Finally, we examine the difference between two different sources of telephone speech, as typified in Voicemail and Switchboard conversations. Specifically we examined the performance on the Switchboard test set using acoustic models trained on Voicemail and vice-versa. Superficially, as Voicemail and Switchboard both represent telephone bandlimited conversational speech, one would expect the performance on either test set to be independent of what database it is trained on, but the results show that this is not the case. The language model and vocabulary were NOT mismatched in these experiments. The difference in performance also appears to depend on the feature space that is used. We present results here for several systems.

From Table VI, the performance degradation from the matched condition (shown underlined) due to a mismatch in the acoustic models ranges from 35-36% for the cepstral feature space to 29-49% for the projected feature space. The degradation appears to be worse for the Switchboard test set. Training the acoustic models on data from both domains does reduce the degradation to a large extent (6% for the cepstral feature space, 1% for the projected feature space). The results show that the individual systems built on either training database are relatively domain-dependent, and that our current modeling techniques are not as robust as one might desire and should be the focus of future algorithm development. Further details of these experiments are given in [15].

VIII. CONCLUSION

In this paper we report on the evolution of the word error rate (WER) on a large vocabulary telephone speech recognition task, as typified in voicemail. We report results on a nu,mber of algorithms spanning the areas of lexicon design, feature extraction, classifier design, and combination of hypotheses, which acoustic models and were instrumental in reducing the word error rate on Voicemail data to around 28%. More specifically, the algorithms and their relative contributions were :

- a lexicon design technique that yields a 7% relative improvemnt in performance
- a novel linear projection (HDA+MLLT) that improves performance on the baseline cepstral feature space by approximtely 10% relative
- a novel feature fusion technique that augments the cepstra with spectral peak energy information and yields a relative improvement of 2.5%
- use of boosting techniques for gaussian mixtures that yields 3% relative improvement
- use of a consensus hypothesis algorithm that provides a 3% relative improvement

Finally, we also reported on the results of some cross-domain experiments that underline the "brittleness" of the speech recognition systems we use today and highlight the need to focus research attention on improving cross-domain performance. In particular

- the cross-domain experiments that show the sensitivity of system performance to training data
- the crude approach of making the system more robust by training on the union of all data sets does seem to work

References

[1] Proceedings of the LVCSR Workshop, Oct 1996.

[2] L. R. Bahl et al., "Performance of the IBM Large Vocabulary Speech Recognition System on the ARPA Wall Street Journal task", Proceedings of ICASSP, 1995.

[3] G. Saon and M. Padmanabhan, "Data-driven Approach to Designing Compound Words for Continuous Speech Recognition", Proceedings of IEEE ASRU Workshop, 1999.

[4] M. Finke and A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition", Proceedings of Eurospeech 1997, vol. 5, pp 2379-2382.

[5] G. Saon, M. Padmanabhan, R. Gopinath, S. Chen, "Maximum likelihood discriminant feature spaces", Proceedings of ICASSP, 2000.

[6] M. Padmanabhan "Use of spectral peak infrmation use in speech recognition", Prceedings of DARPA Speech Transcription Workshop, May 2000.

[7] R. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification", Proceedings of ICASSP, 1998.

[8] K. Paliwal, "Spectral subband centroid features for speech recognition", Proceedings of ICASSP, 1998.

[9] P. Ladfoged, "A course in Phonetics", Harcourt Brace College Publishers, 1993.

[10] M. Padmanabhan, K. Martin, "Resonator-based filter-banks for frequency domain applications", IEEE Trans. Circuits and Systems, Oct 1991.

[11] G. Zweig and M. Padmanabhan, "Boosting gaussian mixtures in an LVCSR system", Proceedings of ICASSP, 2000.

[12] R. Schapire, Y. Freund, P. Bartlett, W. S. Lee, "Boosting the margin: A new explanation for rge effectiveness of voting methods", Annals of Statistics, 26(5): 1651-1686, 1998.

[13] L. Mangu, E. Brill and A. Stolcke, "Finding consensus among words: lattice-based word error minimization", Proceedings of Eurospeech, 1999.

[14] J. G. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER)", Proceedings of IEEE ASRU Workshop, pp. 347-352, Santa Barbara, 1997.

[15] J. Huang et al., "Performance improvement in Voicemail Transcription", Proceedings of DARPA Speech Transcription Workshop, May 2000.