# TRANSCRIPTION OF NEW SPEAKING STYLES - VOICEMAIL

M. Padmanabhan, B. Ramabhadran, E. Eide, G. Ramaswamy,
L. R. Bahl, P. S. Gopalakrishnan, S. Roukos
IBM T. J. Watson Research Center
P. O. Box 218, Yorktown Heights, NY 10598

## 1 INTRODUCTION

In this paper we describe a new testbed for developing speech recognition algorithms - a VoiceMail transcription task, analogous to other tasks such as the Switchboard, CallHome [1] and the Hub 4 tasks [2] which are currently used by speech recognition researchers. Spontaneous speech occurring in day-to-day life can broadly be classified into two categories (i) where the speaker does not receive any external feedback to direct his/her speech, and (ii) where the speaker receives external feedback from another person/machine/audience. Examples of the former category are radio broadcast news, voicemail etc., and examples of the latter category are telephone conversations, natural language transaction systems (eg. ATIS), seminars, etc. In general to obtain the best performance in transcribing a certain style of speech, it is necessary to train the speech recognition system on similar style of training data. Some of the speech categories mentioned above are quite well represented in currently existing databases. However, voicemail data is not well represented in any database, even though it represents a very large volume of real-world speech data. Consequently there is a need for a Voicemail database in order to improve transcription performance on a voicemail transcription task, and also to establish a new test bed for speech recognition algorithms.

Similar to the Switchboard/CallHome databases, the Voicemail database comprises telephone bandwidth spontaneous speech. However the difference with respect to the Switchboard and CallHome tasks is that the interaction is not between two humans, but rather between a human and a machine. Consequently, the speech is expected to be a little more formal in its nature, without the problems of cross-talk, barge-in etc. This eliminates some of the variables and provides more controlled conditions enabling one to concentrate on the aspects of spontaneous speech and effects of the telephone channel. In this paper, we will describe the modality of collection of the speech data, and some algorithmic techniques that were devised based on this data. We will also describe the initial results of transcription performance on this task.

## 2 DATA COLLECTION

For details of the data collection scheme see [3]. Briefly, some of the characteristics of the voicemail data are as follows:

- The data represents extremely spontaneous speech.
- The data contains both long-distance and local calls.
- Each voicemail message typically has a click at the beginning and/or end of the message arising from the caller hanging up.
- The data is subject to the compression of the phonemail system, which leads to a small degradation in accuracy.
- The average length of a voicemail message is 31 seconds, however, the peak of the histogram of voicemail durations occurs at 18 seconds.
- The average rate of the speech is approximately 190 words per minute.
- The topics covered in the collected data ranged from personal messages to messages with technical or business-related content.
- The database was not quite gender balanced, with the percentage of male speakers being 38 %.

## 3 SYSTEM OVERVIEW

We will first briefly describe the IBM large-vocabulary speech recognition system. Essential aspects of the system used in the experiments here have been described earlier [4]; however, we will summarize the main features here :

The acoustic features used are 13-dimensional cepstra and their first and second differences, and a feature vector is extracted every 10 msec from the 8KHz sampled voicemail data. Words are represented as sequences of phones. Each phone is further divided into 3 sub-phonetic units which correspond roughly to the beginning, middle, and end of each phone. The system uses context-dependent HMM acoustic models for these sub-phonetic units. For each sub-phonetic unit a decision tree is constructed from the training data [4]. Each leaf of the tree corresponds to a different set of contexts. The acoustic observations that characterize the training data at each leaf are modeled as a mixture of gaussian pdf's, with diagonal covariance matrices. The systems used in this paper had approximately 2700 leaves, and anywhere from 17000 to 170000 gaussians. The system also uses an envelope-search algorithm [4] to hypothesize a sequence of words corresponding to the utterance. A simple word N-gram (bigram or trigram) model is used to compute the language model probabilities.

## 4 ACOUSTIC MODELS

In this section, we will describe the construction of the acoustic models for this task. The first step is the construction of the decision trees to model context-dependent variations of the sub-phonetic units. The goal here is to model variations in pronunciation arising from context. However, as the voicemail data contains data from different environments, use of this data during the tree growing process may result in trees that try to model the environment variations rather than pronunciation variations. Further, the amount of voicemail data currently available is only around 20 hours. Consequently, we decided to bandlimit the Wall Street Journal SI-284 primary microphone data (WSJ-P) to 200-3400 Hz using a linear-phase 200 tap Lerner filter [5], and used this data to construct the decision trees and the gaussians modelling the leaves of the tree. The parameters of the acoustic model were then re-estimated via the E-M algorithm using the voicemail data.

In order to model the clicks in the voicemail messages, we decided to augment the phone alphabet by adding a 'click' phone. We also added a 'mumble' phone to model inarticulate segments of the messages. Both the 'click' and 'mumble' phones were modelled with 3-state HMM's just as for the other phones.

### 4.1 Clean-up of transcriptions

The initial transcriptions that we started off with for the 20 hours of voicemail data were not very clean, and had a fair number of transcription errors. As it would have been impractical to verify all these transcriptions manually, we devised an automatic scheme to identify possible transcription errors. This tagged around 1 % of the data, and we then corrected these transcriptions manually. Very briefly, the main idea used in the tagging scheme was to viterbi align the speech data against the (possibly incorrect) transcription, and then identify regions where the log-likelihood assigned to a phone by the alignment process was particularly low. For more details see [3]. This process identified script errors as well as baseform errors - for example

(i) we originally only had one baseform for IRA, AY AA R EY (the acronym baseform). In the recorded data IRA occurred as a name with pronunciation AY R AA, and was flagged as an error

(ii) there were several instances where disfluencies such as 'UH' and 'UM' had not been transcribed, and the technique flagged a number of these errors

### 4.2 Compound words

An additional observation arising from the tagged segments of the acoustic data was that crossword co-articulation was 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 OW N T EY KD', in which case at least one of the phones in the phonetic representation for 'going to take' would be flagged. This was clearly not a transcription error, but we needed some mechanism to model such cross-word co-articulation effects (degemination, palatization etc.).

For our initial experiments, we chose to model such effects by constructing compound words [9, 10]. For instance going-to-take would be a compound word, with several possible baseform representations, one of which would be 'G OW N T EY KD'. We selected these compound words based on the tagged segments of the acoustic training data. Some examples of the compound words and their pronunciations is given in Table I

Table I

| | |
|---|---|
| $CAN-WE$ | $K\ AX\ W\ IY$ |
| $FOR-YOU$ | $F\ AX\ Y\ UW$ |
| $GIVE-ME$ | $G\ IH\ M\ IY$ |
| $GOOD-MORNING$ | $G\ UH\ M\ AA\ N\ IX\ N$ |
| $IT-WAS$ | $IX\ W\ AX\ Z$ |
| $SO-IF$ | $S\ OW\ F$ |
| $TO-YOU$ | $CH\ Y\ UW$ |
| $TRYING-TO$ | $T\ R\ AY\ N\ AX$ |
| $WANT-TO$ | $W\ AA\ N\ AX$ |
| $YOU-CAN$ | $Y\ UW\ N$ |

The use of these compound words serves a dual purpose. Firstly, they enable the modelling of cross-word co-articulation effects. Secondly, 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. We decided to extend the second piece of reasoning above and apply it to model commonly occurring phrases in the voicemail data. Hence, we constructed compound words of the form 'give-me-a-call', 'thank-you', 'thanks-a-lot', 'when-you-get-a-chance' etc. The use of these compound words helped bring down the error rate as shown in the section on experimental results.

## 4.3 Phonological rules

In order to model co-articulation effects in words other than compound words, we used some of the phonological rules described in [6]. Examples of such co-articulation effects are plosive deletion (deletion of word final TD in the word sequence 'excellent point'), palatization (did-you being pronounced as 'D IH JH UW'), etc. Such effects can be accounted for using linguistic rules [6, 7, 8], that specify the conditions under which the boundary phones in a word may be deleted or replaced by other phones.

In our implementation, we assumed that only the final and initial phones of the two words in question would be candidates for modification. Also, the changes to the boundary phones were determined using the last two phones of the previous word and the initial phone of the succeeding word only. Further, any number of words could be combined using these rules to produce one long word; for example, 'what-did-you' is a result of the application of two rules, one at the 'what did' juncture and the other at the 'did you' juncture. Finally, all the phonological rules used were optional, i.e., there were no compulsory replacements.

Some of the rules that we implemented are listed below ($P_{n-1}$ and $P_n$ denote the last two phones of the first word, and $N_1$ denotes the first phone of the next word).

1. Geminate Deletion: If $P_n$ = Consonant and $N_1$ = Same consonant then delete $P_n$ Example: this-street DH IH S T R IY TD
2. Palatization: If $P_n$ =D and $N_1$ = Y then replace $P_n$ with JH and delete $N_1$ Example: did-you D IH JH UW and what-you W AH CH UW
3. Plosive Deletion: If $P_{n-1}$ = N, $P_n$ = plosive and $N_1$ = plosive the delete $P_n$ Example: went-down W EH N D AW N
4. If $P_{n-1}$ = N, $P_n$ = D and $N_1$ = DH then delete $P_n$ Example: and-then AX N DH EH N
5. If $P_{n-1}$ = DH, $P_n$ = AX and $N_1$ = Vowel then replace $P_n$ with IH Example: the-apple DH IX AE P AX L
6. If $P_n$ = S or Z and $N_1$ = SH then delete $P_n$ Example: this-show DH IH SH OW
7. If $P_{n-1}$ = Vowel, $P_n$ =T and $N_1$ = Vowel then replace $P_n$ with DX Example: that-again DH AE DX AX G EH N

These rules helped bring down the error rate as indicated in Table VI. Also, analysis of the decoded output indicated that we did not introduce any new insertions or deletions in the process of combining words.

## 4.4 Model Complexity Adaptation

As mentioned earlier, we model leaves in our system with mixtures of gaussians. In general, ad-hoc rules are used to determine the number of mixture components that will be used to model a particular leaf - for example, the number of components is made proportional to the amount of data, subject to a maximum number. This choice of the number of components may not necessarily provide the best classification performance - consequently, we introduced a discriminant measure to choose the number of mixture components in a more optimal manner. The details of this algorithm are given elsewhere [11], so we will only summarize it briefly here.

The essence of the algorithm is to start with a small baseline system, and evaluate how well the gaussian mixture model for a leaf models the data for that leaf. This is done by computing the posterior probability of correct classification of the data for that leaf. If this probability is low, this implies that the model for the leaf does not match the data for the leaf very well;

hence, the resolution of the model for the leaf is increased by adding more components to its model.

In our implementation, we start with two systems (say S1 and S2), where S2 models each leaf with more gaussians than S1. Subsequently, we find those leaves that are not adequetely modelled by S1 according to our discriminant criterion, and replace the model for that leaf in S1 with the corresponding model from S2.

### 4.5 VTL Adaptation

We implemented the VTL technique described in [12, 13, 14] to obtain speaker-normalized models. The technique of [12] uses a mixture of gaussians to model voiced speech, and tries to warp the frequencies of a speaker such that the likelihood of the warped data is maximized by the voiced speech model. The initial generic voiced speech model (mixture of 512 Gaussians) used to seed the iterative process was obtained from gender-balanced WSJ data (10 male and 10 female speakers). In order to determine the voiced frames of speech, we viterbi aligned the data and picked only the frames corresponding to vowels. We selected 17 discrete warp scales ranging from 0.80 to 1.12, and signal-processed the speaker's data using each of these warp scales, and compute the likelihood of the warped features using the generic voiced speech model. The warp scale that scores best is then selected. We repeated this process a few times, re-estimating the generic voiced-speech model at every iteration. Finally, the gaussians modelling the context-dependent sub-phonetic units were trained on the features corresponding to the best warp scale for each speaker, to obtain speaker-normalized models. Experimental results are tabulated in Table VI.

### 4.6 MLLR Adaptation

Finally, we used MLLR adaptation [15] to adapt the acoustic models. In brief, MLLR tries to compute a linear transform that is applied to the means and variances of the gaussians in order to maximize the likelihood of the adaptation data computed with the transformed model. For this technique, it is necessary to have acoustic adaptation data and the corresponding transcription. We used the test data itself as adaptation data, along with the transcription produced by a speaker-independent system to bootstrap the acoustic models. The acoustic models were adapted independently for every voicemail message in the test set (unsupervised sentence-based adaptation).

## 5 LANGUAGE MODEL

The transcription of the 20 hours of voicemail data contained approximately 220K words. This was adequate to build a bigram/trigram language model for the voicemail task. In addition, we attempted to make use of the 2M words of data from the Switchboard database by constructing a trigram language model from the Switchboard data and using a weighted mixture of the language model probabilities provided by the Voicemail and Switchboard language models in the decoder. Further, these language models were constructed from transcriptions that included compound words (i.e. the original transcriptions had been filtered to replace selected sequences of words with a compound word).

Furthermore, in an attempt to use the small amount of voicemail data parsimoniously, we investigated the use of word-classes. The classes were hand-selected based on semantics and/or transcription inconsistencies, and the trigram model used was :

$$p(w_3|w_2w_1) = p(c_3|c_2c_1)p(w_3|c_3) \qquad (1)$$

where $c_i$ is the class of word $i$ and $p(w_i|c_i)$ is the relative frequency of word $i$ in its class, smoothed against a flat model. Some specimen classes are shown in Table II.

*Table II*

| | |
|---|---|
| $\_BYE$ | $BYE-BYE, BYE-NOW$ *etc.* |
| $\_COUNTRY$ | $CHINA, FRANCE$ *etc.* |
| $\_DIGIT$ | $ONE, TWO$ *etc.* |
| $\_GREETING$ | $HELLO, HI$ |
| $\_LASTNAME$ | $HORN, NAHAMOO$ *etc.* |
| $\_THANKS$ | $GRACIAS, THANK-YOU$ *etc.* |

## 6 EXPERIMENTAL RESULTS

Our first set of experiments were conducted when we had only 10 hours of training data available, and several of these experiments were repeated on 20 hours of training data. We will present experimental results for both these training sets (we will refer to them as Vmail10 and Vmail20), as the difference in performance gives an indication of the effect of increasing the amount of training data on different components of the recognizer (acoustic model, language models, etc.).

### 6.1 Test data

The test data was 43 voicemail messages (picked at random from the collected data, and not included in the training set). The size of the Vmail10 vocabulary was 6K words, and the out-of-vocabulary (o.o.v.) rate of the test data with respect to this vocabulary was 4.6 %. The size of the Vmail20 vocabulary was 10K words, and the oov rate of the test data with respect to this vocabulary was 3.5 %. The results in this paper are reported only on the development test data, as the evaluation set had not yet been defined at the time the paper was written.

**6.1.1 Computation of word error rate**—In computing the word error rate, as disfluencies do not contribute to the semantic meaning of the utterance, we decided to filter out all instances of disfluencies in both the reference transcripts and the decoded transcripts, before computing the word error rate of the decoded transcripts. Consequently, deletions of disfluencies in the original reference transcript would not be interpreted an error, substitutions of disfluencies in the original reference transcript with other disfluencies would again not be interpreted as an error. However, substitutions of disfluencies in the original reference script with words other than disfluencies would be interpreted as insertion errors. Also, as we are primarily concerned with the word error rate and not the compound-word error rate, we replaced all compound words in the reference and decoded transcripts with the corresponding sequence of words before computing the error rate.

**6.1.2 Perplexity of test set**—We computed the word perplexity of the test set using various language models. As mentioned above, the filtered reference transcript did not contain any disfluencies; however, the language model data did contain disfluencies (as they are known to be useful linguistic predictors). Consequently, in our perplexity calculations, we computed the total log probability of the words in the unfiltered reference transcripts using the language model (hence the disfluencies were used to predict the word probabilities, and the log probabilities of the disfluencies was also included in the total), and subsequently subtracted out the log probabilities of all the disfluency words in the original reference transcript, before computing the average log probability per word. Also, this measure of perplexity was computed with compound-words in the reference transcript because the language model data also included compound-words [1]. The word perplexity measure was computed with a bigram and trigram LM constructed from the 220K words of voicemail data, and with a weighted mixture of the voicemail trigram LM and a trigram LM constructed from switchboard data in the proportion 0.8 to 0.2. Also, we present the perplexity numbers both with and without taking into account the log probability of the disfluencies in the reference transcript (see Table III).

Table III (perplexity)

| | Fillers | No fillers |
|---|---|---|
| Bigram LM (Vmail10) | 141.73 | 163.37 |
| Bigram LM (Vmail20) | 122.69 | 140.13 |
| Trigram LM (Vmail20) | 117.55 | 133.64 |
| Trigram (Vmail20 + Swb) | 114.71 | 128.96 |

## 6.2 Switchboard training

As the voicemail data and switchboard data both represent telephone-bandwidth spontaneous speech, we initially decoded the voicemail test data using the models used in the Switchboard '95 evaluation [1] (row 1 of Table IV). Subsequently, we replaced the switchboard language model with a bigram that had been trained on 10 hours of voicemail (row 2 of Table IV). Finally, we re-estimated the parameters of the switchboard acoustic model using the Vmail20 data (row 3 of Table IV). The word error rates are summarized in Table IV. The last row in this table represents a system bootstrapped from a switchboard model and then trained on the voicemail data.

Table IV (word error rate)

| | |
|---|---|
| (1) Acoustic - Swb, Trigram LM - Swb | 87.11 % |
| (2) Bigram LM - Vmail10 | 66.47 % |
| (3) Acoustic - Vmail20 | 55.54 % |

[1] The perplexity is computed as

$$P = 2^{-\frac{1}{N}\sum log_2(p)}$$

where $p$ is a linguistic unit in the script. The linguistic units are in general identical to the words in the vocabulary. In our case, we include compound words in the vocabulary and in the language model; however, before scoring the decoded script for the word error rate, we replace the compound words with their corresponding sequence of words. In this case, the linguistic units are not identical to words in the reference script; consequently, we could compute the total log probability of the linguistic entities in the reference transcript and divide this either by the number of linguistic entities, or by the number of words to get an average log probability per linguistic unit or per word. The numbers presented in the table correspond to the former case.

## 6.3 Vmail10 training set

The results of several experiments are summarized in Table V. For all experiments except the last two, only the Vmail10 training set was used for both the acoustic and language models. Results are presented in an incremental manner i.e. each row of the table represents a single change that was made with respect to the previous row, and the description of this change is indicated in the row of the table. The row numbers referred to in the next paragraph refer to the rows in Table V.

(1) The baseline system corresponded to a system with 83.5K gaussians and a bigram LM (row 1).
(2) Next we added compound words to the vocabulary (see Section. 4.2) and decoded with the same acoustic models as before (row 2).
(3) Next, we cleaned up the transcriptions and retrained the acoustic models (see Section 4.1). The error rate corresponding to this condition is shown in row 3.
(4) Then, we estimated a model-complexity-adapted model by putting together a system with 17K gaussians and 175K gaussians (S1 has 17K gaussians, S2 has 175K gaussians - see Section 4.4). The model-complexity-adapted (MCA) model had 32K gaussians. The parameters of this system were then re-estimated using the Vmail10 training set, and the corresponding error rate is shown in row 4.
(5) Next, we replaced the bigram LM with a trigram and used the new LM in conjunction with the MCA system described above (row 5).
(6) Then we used a class-based trigram LM (see Section 4.6) (row 6).
(7) Finally, the acoustic models were re-estimated using MLLR adaptation in unsupervised mode, and on a per-sentence basis, and the adapted models were used with the class-based trigram language model (see Section. 4.6) (row 7).

Table V (word error rate)

| | |
|---|---|
| (1) Baseline (83K gaussians) | 56.24 |
| (2) Compound-words | 51.46 |
| (3) Clean-up transcriptions | 49.75 |
| (4) Model complexity adaptation | 48.44 |
| (5) Trigram LM | 48.19 |
| (6) Trigram with classes | 46.88 |
| (7) MLLR adaptation | 43.86 |

## 6.4 Vmail20 training set

We conducted a number of incremental experiments to observe the effect of adding additional training data to different components of the recognizer. The word error rates are given in Table VI (any reference to row numbers in the remainder of this section should be interpreted as row of Table VI).
(1) We started with the system corresponding to row 3 of Table V, which gave an error rate of 49.75 %, and simply re-estimated the parameters of this acoustic model using the Vmail20 datbase (LM is a bigram estimated from the Vmail10 data). This dropped the error rate to 46.22 % (row 1).
(2) Subsequently, we re-estimated the bigram LM using the Vmail20 database, and decoded the test data using the same acoustic model as in row 1. This dropped the error rate to 45.12 % (row 2).
(3) Subsequently, we estimated a trigram LM using the Vmail20 database, and used this with the same acoustic model of row 1. This dropped the error rate to 42.7 % (row 3).
(4) Next we used a weighted mixture of the Vmail trigram LM of row 3, and a trigram built off the Switchboard data (in the proportion 0.3 Swb LM probability + 0.7 Vmail20 LM probability). The error rate corresponding to this condition was 42.95 % (row 4).
(5) Next, we estimated a MCA model putting together a system (S1) with 83.5K gaussians, and a system (S2) with 175K gaussians. The resulting MCA model had 78K gaussians. Using the mixture trigram LM of row 4, and the MCA model dropped the error rate to 42.20 % (further details are given in the next section). Further, tuning the mixture weights in the language model reduced the error to 41.94 % (final weights were 0.2 Switchboard trigram and 0.8 Vmail20 trigram).
(6) Next, we used VTL to construct a speaker-normalized equivalent of the MCA model, and decoded using the weighted mixture LM of row 5. The error rate dropped to 40.52 % (row 6).
(7) Next, we used the iterative MLLR technique to adapt the means of the gaussians of row 5, individually for each message, and used these adapted models with the weighted mixture LM of row 5. The error rate dropped to 39.43 % (row 7).
(8) Next, we started with the speaker-normalized VTL models of row 6 and further applied the iterative MLLR technique to further refine the means of the gaussians for each individual message. These adapted models were then used with the weighted mixture LM of row 5. As can be seen from the results (row 8), the effect of VTL and MLLR does appear to be additive. (9) Finally, we applied the phonological rules of Section. 4.3 in the decoding process, and used them with the models of (8). This brought the error rate down to 38.18 % (row 9).

| Table VI (word error rate) | |
|---|---|
| (1) Bigram LM - Vmail10 | 46.22 |
| (2) Bigram LM - Vmail20 | 45.12 |
| (3) Trigram LM - Vmail20 | 42.70 |
| (4) Trigram LM - Vmail20 + Swb | 43.25 |
| (5) MCA model | 42.20 |
| (5b) | 41.94 |
| (6) VTL | 40.52 |
| (7) MLLR | 39.43 |
| (8) VTL+MLLR | 38.92 |
| (9) Phonological rules | 38.18 |

**6.4.1 Model Complexity Adaptation**—We now present some experimental results on model complexity adapation (MCA) (see Section. 4.4) that indicate that the new method of determining the complexity of the model yields consistent gains over standard methods. We constructed five models using the standard ad-hoc method of allocating a fixed number of gaussians for the each leaf. These models respectively had a maximum of 7, 12, 35, 60, and 150 gaussians per mixture (gpm). Subsequently, we used MCA to construct models that replace the gaussian mixtures for some leaves in the 7 gpm model with gaussian mixtures from the 35 gpm model. This model will be referred to as 7x35 in the following table (Table VII). Table VII tabulates the error rates and the size of several models, constructed by conventional means, and using MCA.

Table VII

| | # gaussians | Word error rate |
|---|---|---|
| Conventional models | | |
| 7 gpm | 16.5K | 47.53 |
| 12 gpm | 25.8K | 45.07 |
| 35 gpm | 69.5K | 43.25 |
| 60 gpm | 89.5K | 43.86 |
| 150 gpm | 128.5K | 44.16 |
| MCA models | | |
| 7x35 | 25K | 45.72 |
| 12x35 | 33K | 43.2 |
| 35x150 | 78.2K | 42.2 |
| 60x150 | 96.5K | 42.9 |

The error rate as a function of the number of gaussians in the model is shown plotted in Fig. 1, and it can be seen that the MCA models consistently outperform the conventional models by around 5% (relative). Also, note that due to the limited amount of training data, the error rate starts increasing as the number of parameters increases beyond a certain point.

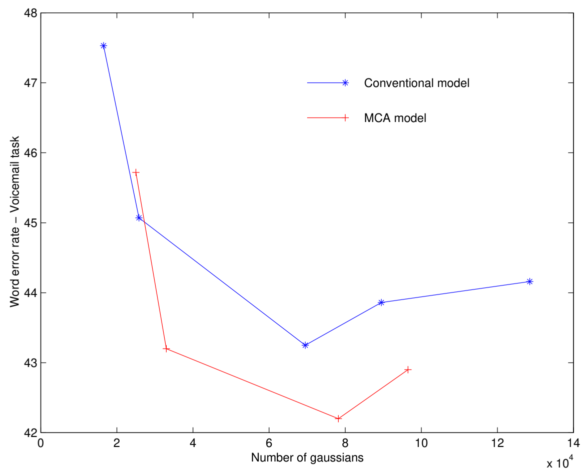


Figure 1:

## 7 CONCLUSION

We reported transcription word error rates on a new testbed representing telephone-bandwidth spontaneous speech i.e., the task of voicemail transcription. We described the process of bootstrapping the models starting from either the Switchboard data, or bandlimited Wall Street Journal data. The results show that better performance was obtained in the latter case. We described several techniques that were used to construct the acoustic models including (i) the use of compound words and linguistically derived phonological rules to model co-articulation effects that occur in spontaneous speech (ii) a new model-complexity adaptation technique that uses a discriminant measure to allocate gaussians to the mixtures modelling allophones. We also investigated the efficacy of some well known acoustic adaptation techniques on this task. We also described experiments related to building language models using the limited amount of training data available in this domain. We then reported experimental results that showed that most of the modelling techniques we investigated were useful in reducing the word error rate. Finally, we reported experimental results on two different sized training sets to show the effect of increasing the training data on different (acoustic and linguistic) components of the recognizer.

## 8 ACKNOWLEDGEMENT

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

## REFERENCES

[1] Proceedings of LVCSR Workshop, Oct 1996, Mar-

itime Institure of Technology.

[2] Proceedings of ARPA Speech and Natural Language Workshop, 1995, Morgan Kaufman Publishers.

[3] M. Padmanabhan, G. Ramaswamy, B. Ramabhadran, P. S. Gopalakrishnan, C. Dunn, "Issues involved in voicemail data collection", elsewhere in these proceedings.

[4] L. R. Bahl et al., "Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task", Proceedings of the ICASSP, pp 41-44, 1995.

[5] K. Martin and M. Padmanabhan, "Resonator-In-A-Loop Filter-Banks based on a Lerner grouping of outputs", Proceedings of the ICASSP, 1992.

[6] E. P. Giachin, A. E. Rosenberg and C. H. Lee, "Word juncture modeling using phonological rules for HMM-based continuous speech recognition", Computer, Speech and Language, pp 155-168, Academic Press, 1991.

[7] P. S. Cohen and R. L. Mercer, "The Phonological Component of an Automatic Speech Recognition System", Speech Recognition, (D. Raj Reddy ed.), Academic Press, pp 275 - 320, 1975.

[8] B. T. Oshika, V. W. Zue, R. V. Weeks, H. New, and J. Aurback, "The Role of Phonological Rules in Speech Undersatnding Research", IEEE Transactions on ASSP, vol. 23, pp 104-112, 1975.

[9] 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.

[10] P. Jeanrenaud, et al., "Reducing word error rate on conversational speech from the Switchboard corpus", Proceedings of ICASSP, 1995, pp 53-56.

[11] L. R. Bahl, M. Padmanabhan, "A discriminant measure for model complexity adaptation", submitted to ICASSP 98.

[12] S. Wegman, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech".

[13] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization", Proceedings of ICASSP, 1996, pp 346-348.

[14] T. Kamm, G. Andreou, and J. Cohen, "Vocal tract normalization in speech recognition: compensating for systematic speaker variability", Proc. 15th Annual Speech Research Symposium, CLSP, Johns Hopkins University, Baltimore, June 1995, pp 175-178.

[15] C. J. Legetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous density HMM's", Computer Speech and Language, vol. 9, no. 2, pp 171-186.