SPEECH RECOGNITION PERFORMANCE ON A VOICEMAIL TRANSCRIPTION TASK

M. Padmanabhan, E. Eide, B. Ramabhadran, G. Ramaswamy, L. R. Bahl *
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 - the ARPA-sponsored VoiceMail transcription task, analogus to other tasks such as the Switchboard, CallHome [1] and the Hub 4 tasks [2] which are currently used by speech recognition researchers. As the name indicates, the task involves the transcription of voicemail conversations. Voicemail represents a very large volume of real-world speech data, which is however not particularly well represented in existing databases. For instance, the Switchboard and CallHome databases contain telephone conversations between two humans, representing telephone-bandwidth spontaneous speech; the Hub 4 database contains radio broadcasts which represents different kinds of speech data such as spontaneous speech from a well-trained speaker, conversations between two humans possibly over the telephone, etc. The Voicemail database on the other hand also represents 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

The data was collected at various IBM sites in the US. The method that was used to collect the data was as follows: volunteers from the site would be asked to donate their non-confidential voicemail messages to the database in return for some incentives. However for privacy reasons, it was necessary to inform the person leaving the message (the caller) that the data could be used for research purposes, so the volunteers were required to add a sentence to their outgoing message of the form 'Your voicemail data may also be used for commercial research purposes in developing speech recognition algorithms. If you do not want your data to be used, please say so in your message.' Subsequently, if the caller did not specify any objection to his/her data being used, and if the volunteer felt that the message did not contain any confidential information, he/she would forward the message to a telephone extension which was set up for the purpose of collecting these messages.

At the time this paper was written, the database comprised of around 10 hours of data collected from volunteers at the IBM T. J. Watson Research Center in Yorktown Heights. Some of the characteristics of the voicemail data that we collected are as follows:

- The data contains both long-distance and local calls.
- Each voicemail message typically has a click at the beginning or end of the message arising from the caller hanging up.
- The data represents extremely spontaneous speech. One of the initial assumptions prior to collecting the data was that as the caller would be leaving a message on a machine, the speech would be relatively well articulated, but more often than not, there are a lot of disfluencies in the speech, and often the rate of speaking is also quite high, leading to cross-word co-articulation effects.

We will next give a detailed description of the speech recognition system, and the new algorithms that were developed to transcribe the voicemail data.

[0] We would like to thank Dr. Salim Roukos, Dr. P. S. Gopalakrishnan, and Dr. David Nahamoo for their useful com-

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 [3]; however, we will summarize the main features here :

The 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 subphonetic 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 [3]. 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 [3] to hypothesize a sequence of words corresponding to the utterance. A simple N-gram (bigram or trigram) 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 in the construction of the acoustic models 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 isolate the environment rather than pronunciation variations. Further, the amount of voicemail data currently available is only around 10 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, 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 10 hours of voicemail data. This represented our baseline system.

Further, in order to model the clicks in the voicemail messages, we decided to augment the phone alphabet by adding a 'click' phone. Further, 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 transcription that we started off for the 10 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 flagged around 1 % of the data, and we then corrected these transcriptions manually. The scheme was as follows: we first viterbi-aligned the voicemail data against the initial transcriptions using the baseline model. Subsequently, we computed the log-likelihood of each instance of a phone in the training data, conditioned on the alignment, and computed the average per-frame likelihood by normalizing by the number of frames that aligned to the phone. Then, we computed a histogram of these per-frame log-likelihood scores for each phone over all the training data. Next we went through the training data again and identified those instances of phones with per-frame likelihoods less than three 3σ below the mean per-frame likelihood for that phone (where σ represents the standard deviation of the score), and tagged the region of the acoustic corresponding to that instance of the phone as a possible transcription error. Finally, we listened to the tagged acoustic segments and manually corrected the transcriptions. Some examples of such corrections were
(i) the baseform for the name IRA was initially incorrectly specified as AY AA R EY (the correct baseform was AY R AX), and this 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

The main objective in attempting this clean-up of the transcriptions was to obtain sharper acoustic models, and as the experimental results will show, this did help the error performance.

4.2 Compound words

An additional observation arising from the tagged segments of the acoustic data was that crossword coarticulation 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 AE

ments and support.

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.). One possibility is to use phonetic rules [4], however, for our initial experiments, we chose to model such effects by constructing compound words. For instance going-to-take would be a compound word, with several possible baseform representations, one of which would be 'G OW N T AE 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 [1].

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 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 eg., 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 [6], so we will only summarize it briefly here.

The essence of the algorithm is to start with a 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. We experimented with two systems that were designed in this manner, the first one with 23K gaussians, and the second one with 32K gaussians. Both these systems were subsequently retrained on the 10 hours of voicemail data. The experimental results show that this manner of adjusting the complexity of the model does provide performance improvements.

4.4 MLLR Adaptation

Finally, we used MLLR adaptation [7] to adapt the acoustic models. We adapted the acoustic models independently for every voicemail message in the test set, starting from the initial transcription of the message produced by the speaker-independent acoustic model (unsupervised sentence-based adaptation).

5 LANGUAGE MODEL

The transcription of the 10 hours of voicemail data contained approximately 100K words. This was adequate to build a bigram language model for the voicemail task. In addition, we attempted to make use of the 2M words of data from the Switchboard database by pooling the voicemail transcription data with the Switchboard data in a proportion of 30:1, and building a trigram from the combined data. Furthermore, in an attempt to use the small amount of voicemail data parsimoniously, we attempted to use 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.

[1] We note that a similar technique was used in [5] to obtain performance improvements on the Switchboard task.

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

The test data was 43 messages also collected from the same IBM site. The size of the vocabulary was 6K words, and the test data had an out-of-vocabulary rate of 4.5%. The perplexity of the test data using the bigram LM was 196. The results of several experiments are summarized in Table II. The conditions corresponding to each experiment are summarized below. Unless otherwise indicated, all experiments used a bigram language model.

(i) The acoustic models corresponded to the baseline with 17K gaussians.

(ii) The acoustic models corresponded to the baseline with 81K gaussians.

(iii) Compound words were added to the vocabulary, and the acoustic model of (i) with 17K gaussians was used.

(iv) Compound words were added to the vocabulary, and the acoustic model of (ii) with 81K gaussians was used.

(v) The 17K baseline acoustic model was retrained with cleaned-up transcriptions, and used along with compound words.

(vi) The 81K baseline acoustic model was retrained with cleaned-up transcriptions, and used along with compound words.

(vii) A model-complexity-adapted acoustic model was designed with 23K gaussians. Compound words were used in the vocabulary.

(viii) A model-complexity-adapted acoustic model was designed with 32K gaussians. Compound words were used in the vocabulary.

(ix) The acoustic model of (viii) was used along with a trigram language model.

(x) The acoustic model of (viii) was used along with a class-based trigram language model.

(xi) The acoustic models of (viii) 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.

Table II

Experiment #	*Word Error rate (%)*
Baseline	
i	57.6
ii	56.24
$Compound-words$	
iii	52.87
iv	51.46
$Cleaned-up$ *transcriptions*	
v	51.66
vi	49.75
$Model-complexity$ *adaptation*	
vii	50.65
viii	48.44
Trigram language model	
ix	48.19
x	46.88
MLLR adapted models	
xi	43.86

REFERENCES

[1] Proceedings of LVCSR Workshop, Oct 1996, Maritime Institute of Tdchnology.

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

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.