Speech Communication 13 (1993) 263–279
North-Holland

Large vocabulary speech recognition using subword units

C.-H. Lee, J.-L. Gauvain [1], R. Pieraccini and L.R. Rabiner

AT&T Bell Laboratories, Murray Hill, NJ 07974, USA

Received 5 February 1993
Revised 31 May 1993

Abstract. Research in large vocabulary speech recognition has been intensively carried out worldwide, in the past several years, spurred on by advances in algorithms, architectures and hardware. In the United States, the DARPA community has focused efforts on studying several continuous speech recognition tasks including Naval Resource Management, a 991 word task, ATIS (Air Travel Information System), a speech understanding task with an open vocabulary (in practice on the order of several thousand words) and a natural language component, and Wall Street Journal, a voice dictation task with a vocabulary on the order of 20,000 words. Although we have learned a great deal about how to build and efficiently implement large vocabulary speech recognition systems, there remain a whole range of fundamental questions for which we have no definitive answers. In this paper we review the basic structure of a large vocabulary speech recognition system, address the basic system design issues, discuss the considerations in the selection of training material, choice of subword unit, method of training and adaptation of models of subword units, integration of language model, and implementation of the overall system, and report on some recent results, obtained at AT&T Bell Laboratories, on the Resource Management task.

Zusammenfassung. Die Forschung im Bereich der Sprachworterkennung bei grossem Wortschatz wurde in den letzten Jahren weltweit intensiv betrieben und durch Fortschritte in den Algorithmen, in den Strukturen und in den Geräten angespornt. In den USA hat die DARPA-Gemeinschaft ihre Anstrengungen auf die Untersuchung verschiedener Methoden zur kontinuierlichen Spracherkennung konzentriert, darunter das "Naval Resource Management", eine Aufgabe mit 991 Worten, ATIS (Air Travel Information System = Informationssystem für Flugreisen), eine Aufgabe für Sprachverständigung mit offenem Vokabular (in der Praxis mit mehreren Tausend Worten) und ein natürlicher Sprachanteil sowie das Wall Street Journal, eine Sprachdiktataufgabe mit einem Vokabular von ca. 20.000 Worten. Obwohl wir viel darüber erfahren haben, wie man Erkennungssysteme für ein breites Sprachvokabular aufbaut und wirksam implementiert, bleiben noch eine ganze Reihe von grundlegenden Fragen offen, für die wir keine definitiven Antworten haben. In diesem Artikel geben wir einen Überblick über die grundlegende Struktur von Erkennungssystemen für ein breites Sprachvokabular, sprechen über die Ziele des Basissystems, diskutieren über die Wahl von Übungsgeräten, sublexikalen Einheiten, einer Methode zur Übung und Anpassung von Modellen von sublexikalen Einheiten, Integration von Sprachmodellen und die Implementierung des Gesamtsystems und berichten über einige neuere Ergebnisse, die im Labor AT&T Bell betreffs der "Resource Management" Aufgabe erzielt wurden.

Résumé. La recherche dans le domaine de la reconnaissance de grands vocabulaires s'est développée de façon intensive, au niveau international ces dernières années, stimulées par les progrès dans les domaines de l'algorithmique, des architectures et des matériels. Aux Etats-Unis, la communauté DARPA a orienté ses efforts sur l'étude de diverses tâches de reconnaissance de parole continue, dont la gestion de ressources navales (Ressource Management: une tâche de 991 mots), un système d'information sur les transports aériens (ATIS, une tâche de compréhension de parole avec un vocabulaire ouvert (en pratique, plusieurs milliers de mots) et une composante de traitement de langage naturel) et le Wall Street Journal (WSJ: une tâche de dictée vocale avec un vocabulaire de l'ordre de 20000 mots). Bien que nous ayons beaucoup appris sur la façon de construire et d'implémenter efficacement des systèmes de reconnaissance de grands vocabulaires, il reste toutes une série de questions fondamentales pour lesquelles nous n'avons pas de réponses définitives. Dans cet article,

[1] Jean-Luc Gauvain is now with the Speech Communication Group at LIMSI/CNRS, Orsay, France.

SSDI 0167-6393(93)E0052-W

nous rappelons la structure de base d'un système de reconnaissance de grands vocabulaires, nous discutons des questions de structure, de sélection du matériau d'apprentissage, de choix des unités sub-lexicales, des méthodes d'apprentissage et des modèles, d'integration d'un modèle de langage et d'implémentation du système global. Nous fournissons des résultats récents obtenus à AT&T Bell Laboratories sur la tâche Ressource Management.

Keywords. Continuous speech recognition; subword HMMs; context dependency; task dependency; speaker dependency; maximum a posteriori estimation.

1. Introduction

In the past few years a significant portion of the research in speech recognition has gone into studying the problem of how to build and implement a large vocabulary, continuous speech recognition system. Much of this effort has been stimulated by DARPA which has funded research on three recognition tasks, namely the Naval Resource Management (RM, see (Price et al., 1988)), the Air Travel Information System (ATIS, see (Hemphill et al., 1990; Hirshman et al., 1992)) and the Wall Street Journal (WSJ, see (Paul and Baker, 1992)) tasks. In addition, there is worldwide interest in large vocabulary speech recognition because of the potential applications for voice database access and management (e.g. ATIS), voice dictation (e.g. (Jelinek, 1985)) and limited-domain spoken language translation (Roe et al., 1992). In Japan, the large vocabulary recognition (LVR) systems are mostly developed around the concept of *interpreting telephony* (Morimoto et al., 1990; Sagayama et al., 1992). In Europe, the Philips SPICOS system (e.g. (Ney and Paeseler, 1988)), the CSELT system (e.g. (Fissore et al., 1989)), the Cambridge University system based on the HKT Toolkit (Woodland and Young, 1992) and the LIMSI effort (Lamel and Gauvain, 1992) are examples of the current activity in large vocabulary recognition research. In Canada, the most notable LVR project is the INRS 86,000 isolated word recognition system (Deng et al., 1990). In the United States, in addition to the LVR research in AT&T (Lee et al., 1990; Ljolje et al., 1992) and IBM (Jelinek et al., 1985), most of the LVR effort is sponsored by DARPA, including the BBN BYBLOS system (Schwartz et al., 1989), the CMU SPHINX system (Lee, 1989) and SPHINX-II system (Huang et al., 1993), the Dragon WSJ system (Baker et al., 1992), the Lincoln Lab. Tied-Mixture System (Paul, 1989), the MIT Summit system (Zue et al., 1989) and the SRI DECIPHER system (Murveit et al., 1989). Although some of the systems have been trained to individual speakers (Averbuch et al., 1987; Roe et al., 1992), most current large vocabulary recognition systems have the goal of continuous speech recognition on fluent input by any talker (speaker independent systems).

Although we have learned a great deal about how to build and efficiently implement large vocabulary speech recognition systems, there remain a whole range of fundamental questions for which we have no definitive answers. For example we do not yet know the best way to build and train the fundamental subword units from which word models are created. We do not yet know the best way to impose language constraints on the recognizer so as to utilize all available knowledge in the most computationally efficient manner. We do not yet understand the best way to implement a recognition system so as to maximize the probability of recognizing the spoken string while minimizing the computation for string comparison and searching through the recognition network. We do not yet know how to integrate suprasegmental information such as prosody and duration into the existing recognition systems which rely mainly on frame-level spectral information. We do not yet know how to extract robust features so that the recognition systems are less vulnerable to acoustic mismatch problems caused by talkers, transducers, channels and speaking environments. We do not yet have satisfactory solutions to address the portability issues so that we can efficiently acquire the knowledge sources needed for solving new applications. Most of the existing systems require acquisition of a large amount of application-specific acoustic and language training data in order to build application-specific systems.

Even though there are still a number of unre-

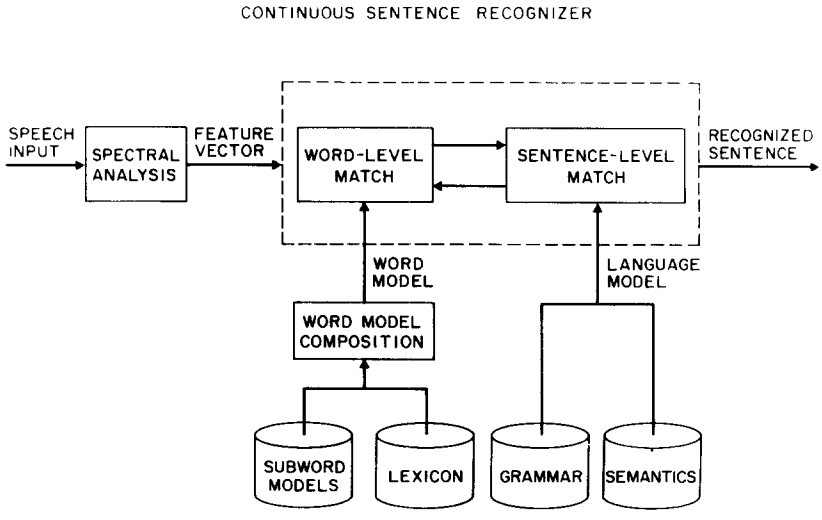

Fig. 1. Block diagram of the continuous speech recognizer.

solved issues, the research community has made significant advances in the state of the large vocabulary recognition technology over the last few years. The approach that is conventionally taken to large vocabulary speech recognition is basically a statistical pattern recognition approach. The fundamental speech units use phonetic labels but are modeled acoustically based on a lexical description of the words in the vocabulary. In general, no assumption is made, a priori, about the mapping between acoustic measurements and subword linguistic units such as phonemes; such a mapping is entirely learned via a finite labeled training set of speech utterances. The resulting speech units, which we call *phone-like units* or PLUs are essentially acoustic descriptions of linguistically-based units as *represented in the words occurring in the given training set*. [2]

A block diagram of a large vocabulary continuous speech recognition system developed at AT&T Bell Laboratories is shown in Figure 1. The system consists of three main modules, namely a feature analysis (or spectral analysis) module, a word-level acoustic match module, and a sentence-level language match module. The feature analysis module provides the acoustic feature vectors used to characterize the spectral properties of the time varying speech signal. The word-level acoustic match module evaluates the similarity between the input feature vector sequence (corresponding to the input speech) and a set of acoustic word models to determine what words were most likely spoken. The sentence-level match module uses a language model (based on a set of syntactic and semantic rules) to determine the word sequence for the most likely sentence.

This paper is organized as follows. In Section 2 we discuss each module of the baseline system of Figure 1 in more detail. We will attempt to explain what is understood about each module, and where active research is ongoing in order to resolve differences of opinion as to the best way to implement the desired processing. In Section 3 we focus our discussion on acoustic modeling issues, including training and adaptation of models for subword units. A brief discussion of the DARPA Naval Resource Management task, databases for training and adaptation and experimental setup is given in Section 4. We then present some recent results on speaker-independent, speaker-dependent and speaker-adaptive recognition of the RM task in Sections 5 and 6.

[2] We will return to this important point later in this paper when we discuss creation of so-called task-independent or vocabulary-independent subword units.

2. The baseline speech recognition system

2.1. Acoustic analysis module

The purpose of the acoustic analysis module is to parameterize the speech into a series of feature vectors that contain the relevant (for recognition) information about the sounds within the utterance. Although there is no consensus as to what constitutes the optimal spectral analysis, there are generally several aspects of the analysis that are common to most recognition systems. For example most systems use LPC spectral analysis methods based on fixed sized frames, e.g., every 10 ms an analysis of a fixed frame of 30 ms of signal is performed. Typically the LPC analysis provides a set of cepstral coefficients for the frame. Sometimes non-uniform frequency scales are used giving the so-called *mel frequency cepstral coefficients* (Davis and Mermelstein, 1980). The rationale here is that since the human ear perceives frequencies on a non-uniform scale, it would be desirable to represent the spectral information of sounds on the same perceptual scale.

In the last few years the spectral feature set for each frame has been extended to include dynamic information about the derivatives (first and second order) of the cepstral vector as well as the static information about the cepstrum (Furui, 1986; Juang et al., 1987; Soong and Rosenberg, 1988; Lee et al., 1992) Also scalars representing frame energy and its derivatives are often used as part of the representation for each frame. For the system implemented at Bell Labs and presented in this study, each 30 ms of speech (8 kHz sampling rate) was analyzed 100 times per second (10 ms shift) to give a spectral vector with 12 LPC-derived cepstral coefficients (on a uniform frequency scale), 12 first order cepstral derivatives, 12 second order cepstral derivatives, and first and second order log energy derivatives. Hence a spectral vector with 38 components was created every 10 msec throughout the signal (Lee et al., 1992).

2.2. Word level match module

The essence of the word level match module is the set of subword models and the lexicon, as seen in Figure 1. The subword models are the representation of the fundamental speech units used as the building blocks for words, phrases and sentences. Probably the most research in large vocabulary speech recognition has gone into defining these subword units in a manner such that they can be easily trained from finite training sets of speech material, such that they are robust to natural variations in accent, word pronunciation and test material, and such that they provide high recognition accuracy for the required speech

Table 1
The 47 context-independent PLUs

Number	Symbol	Word	Number	Symbol	Word	Number	Symbol	Word
1	h#	silence	17	er	b*ir*d	33	p	*p*op
2	aa	f*a*ther	18	ey	b*ai*t	34	r	*r*ed
3	ae	b*a*t	19	f	*f*ief	35	s	*s*is
4	ah	b*u*tt	20	g	*g*ag	36	sh	*sh*oe
5	ao	b*ough*t	21	hh	*h*ag	37	t	*t*ot
6	aw	b*ou*gh	22	ih	b*i*t	38	th	*th*ief
7	ax	*a*gain	23	ix	ros*e*s	39	uh	b*oo*k
8	axr	din*er*	24	iy	b*ea*t	40	uw	b*oo*t
9	ay	b*i*te	25	jh	*j*udge	41	v	*v*ery
10	b	*b*ob	26	k	*k*ick	42	w	*w*et
11	ch	*ch*urch	27	l	*l*ed	43	y	*y*et
12	d	*d*ad	28	m	*m*om	44	z	*z*oo
13	dh	*th*ey	29	n	*n*o	45	zh	mea*s*ure
14	eh	b*e*t	30	ng	si*ng*	46	dx	bu*tt*er
15	el	bott*le*	31	ow	b*oa*t	47	nx	ce*n*ter
16	en	butt*on*	32	oy	b*oy*			

task. To date no one has defined the "ideal" set of subword units. However a great deal of thought has gone into deciding what the real issues are in defining and using various alternatives for the subword units.

Perhaps the simplest set of subword units, which are widely used, is the set of basic phonemes of the language. Although there is no complete agreement as to what sounds are part of this basic set, Table 1 shows one representative set of 47 such phonemes with typical words in which the phonemes appear. These basic units, when trained from real speech material, are called context-independent (CI) phone-like units (PLU) since the sounds are represented independent of the linguistic context in which they occur, and since the spectral properties of the sounds are learned from a training set, rather than postulated on the basis of the linguistic features of the units.

In contrast to the 47 CI-PLUs of Table 1, one could consider subword units which were context dependent (CD). Thus, for example, a separate unit could exist for the sound /ae/ when preceded by /f/ and followed by /t/ (as in fat), then for /ae/ when preceded by /b/ and followed by /t/ (as in bat). In theory there could be as many as $(47)^3$ CD-PLUs when considering all preceding and following sounds; in practice there are on the order of 10,000 such possibilities, a number significantly less than the 100,000 count of $(47)^3$, but significantly more than the 47 CI-PLUs of Table 1. Such CD-PLUs have been extensively used for large vocabulary speech recognition (Lee, 1989; Morimoto et al., 1990), but practical methods are generally used to restrict the number of units to something on the order of 1000–2000 units.

The second basic component of the word-level match module is the lexicon which provides a linguistic description of the words in the task vocabulary in terms of the basic set of subword units. Among the issues in the creation of a suitable word lexicon is the base (or standard) pronunciation of each word and the number of alternative pronunciations provided for each word. The base pronunciation is the equivalent, in some sense, of a pronunciation guide to the word. The number of alternative pronunciations is a measure of word variability across different regional accents and talker populations. Although there have been some very interesting experiments based on multiple word pronunciation lexicons (Weintraub et al., 1989), most large vocabulary speech recognition systems rely on a lexicon with only a single pronunciation provided for each word. This "canonic" representation of each word must be consistent with the subword units; hence its form changes as different sets of CD or CI subword units are used. Also, for function words like "the", "and", "to", etc., it is well known that there is no "canonic" or standard pronunciation. Hence a single representation for such function words will invariantly lead to some problems with recognition. Another issue with the lexical representation of words is that the word pronunciation changes as a result of coarticulation in fluently spoken continuous speech. This can lead to an increase in recognition error rate unless such pronunciation changes are modeled properly. Phonological rules have been proposed to handle intra-word and inter-word coarticulation (Giachin et al., 1991; Lamel and Gauvain, 1992) and shown to be effective in producing a more accurate set subword models and giving better results in recognition.

The word model composition component of the word-level match module is simply the process of retrieving the word pronunciation from the lexicon, and then connecting appropriate subword units to form pronunciation networks according to some phonological rules. The individual word model is created according to the network that corresponds to the word. Word models are then used to match against the spectral vectors of the input speech signal to locate words in continuous speech. In the next section, we will discuss how subword units are modeled and how the models are trained from a finite set of speech utterances.

2.3. Sentence level match module

The sentence level match module uses the constraints imposed by a grammar (a set of syntactic rules on which words are allowed in given contexts) and a set of semantic rules (which eliminate meaningless sentences) to determine the op-

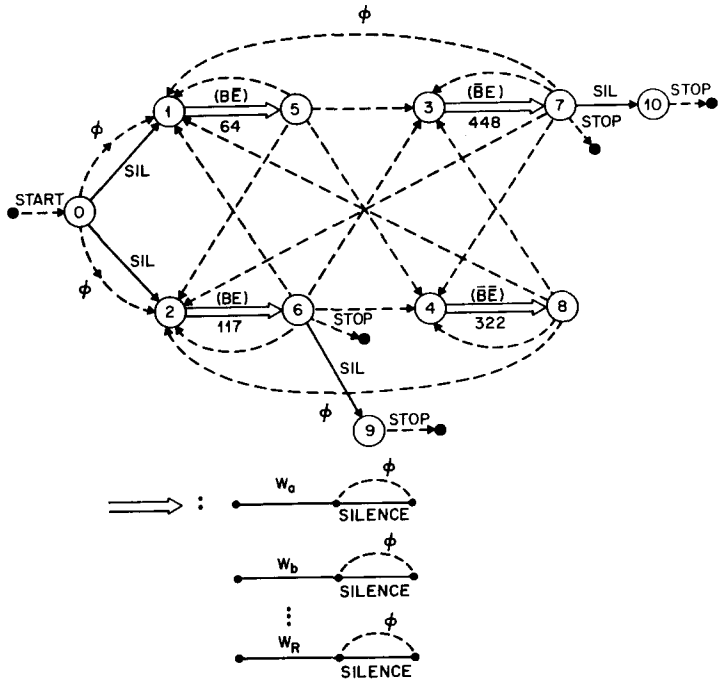

Fig. 2. Network representation of the WP grammar.

timal sentence in the language – i.e. the best word sequence, consistent with the grammar and the semantics, that matches the input speech. Although there have been proposed a number of different forms for the grammar (e.g. formal grammar, N-gram word probabilities, word pair, etc.), we assume a simple grammar that can be represented as a finite state network (FSN). In this manner it is relatively straightforward to implement the grammar directly with the word-level match module. In particular, for the DARPA RM task (991 words), we have used either a word-pair (WP) grammar, which specifies explicitly, for each word in the vocabulary, which words are allowed to follow that word, or a no-grammar (NG) grammar, in which we assume that every word can follow every word in the vocabulary. The perplexities (average word branching factor) of these two grammars is 60 for the WP case and 991 for the NG case. The implementations of these grammars as FSNs is shown in Figures 2 and 3. For the WP case we exploit the fact that only a subset of the vocabulary occurs as the first word in a sentence (condition B for beginning words), and only a subset of the vocabulary occurs as the last word in a sentence (condition E for ending words); hence we can partition the vocabulary into 4 non-overlapping sets of words, namely

$\{BE\}$ = set of words which can either begin or end a sentence, $|BE| = 117$,

$\{B\bar{E}\}$ = set of words which can begin but which cannot end a sentence, $|B\bar{E}| = 64$,

$\{\bar{B}E\}$ = set of words which cannot begin but can end a sentence, $|\bar{B}E| = 488$,

$\{\bar{B}\bar{E}\}$ = set of words which cannot begin or end a sentence, $|\bar{B}\bar{E}| = 322$.

The resulting FSN of Figure 2 has 995 real arcs and 18 null arcs. To account for silence between words (which is optional) each word arc bundle

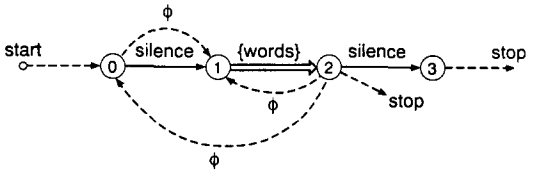

Fig. 3. Network representation of the NG grammar.

(nodes 1 to 4) is expanded to individual words followed by optional silence, as shown at the bottom of Figure 2. Hence the overall FSN allows recognition of sentences of the form

S: (silence)–$\{B\bar{E}, BE\}$–(silence)–$(\{W\})\cdots(\{W\})$–(silence)–$\{\bar{B}E, BE\}$–(silence),

where $\{W\}$ is any word which is allowed to follow the previous word and includes optional silence.

The FSN for the NG case, as shown in Figure 3, is considerably simpler than the FSN for the WP case. The sentences allowed by this grammar are of the form

S: (silence)–$(\{W\})\cdots(\{W\})$–(silence),

where $\{W\}$ is now any word in the task vocabulary.

The grammar FSNs of Figures 2 and 3 have the property that they can produce any valid sentence in the task language. Unfortunately they also have the property that they can produce a large number of sentences which are not valid in the task language, e.g. the sentence S: "and" "and" "and" is valid for the NG network but not for the RM task. The overcoverage (ratio of sentences generated by the FSN to sentences valid in the task language) of the FSNs is often extremely large and this is a negative feature of using these simple networks as the grammar network. On the other hand, using a full grammar (i.e. no overcoverage) is generally prohibitive from a computational point of view.

One way to compensate for the overcoverage of the FSN grammar implementations is to use a semantic processor to detect and correct invalid sentences. In a sense the semantic processor exploits the fact that the syntax used in recognition has a great deal of overcoverage, i.e. it allows meaningless sentences to be passed to the semantic module. The semantic processor can use the actual task perplexity (generally much lower than the perplexity of the syntax) to convert the recognized output to a semantically valid string (Pieraccini and Lee, 1991).

In theory, the semantic processor should be able to communicate back to the recognizer to request a new string whenever the resulting string from the syntactic FSN is deemed invalid. In practice, one of two simple strategies can be used; either the recognizer can generate a list of the best N sentences ($N = 10$–1000) that a language processor can search until a semantically valid string is found (Schwartz et al., 1992), or it can assume that the best (recognized) string is semantically "close" to the correct string, and therefore process it appropriately to determine a semantically valid approximation (e.g. (Pieraccini and Lee, 1992)).

3. Training of subword units

3.1. Subword unit models

A key to the success of modern speech recognition systems is the use of statistical modeling techniques (e.g. hidden Markov models – HMMs) to represent the basic subword units (e.g. (Rabiner, 1989)). Although many variants exist, perhaps the simplest way subword units are modeled is as a left-to-right HMM, of the type shown in Figure 4. Each unit is represented by a simple first-order, left-to-right HMM having N states, $S_1, S_2, \ldots, S_N$, with only *self* and *forward* transitions.

Within each state of the model there is an observation density which specifies the likelihood (or probability) of a spectral vector from the speech signal occurring within the model state. This observation density can either be a discrete density (implying the use of a common codebook to discretize the input spectral vector), or a continuous density, or even what is called a semi-continuous density (Huang et al., 1990) or a tied-mixture density (Bellegarda et al., 1990) which is a codebook of continuous densities whose weights are chosen according to the model state. Although continuous density modeling usually provides the highest performance recognition systems, it requires the most computation to imple-

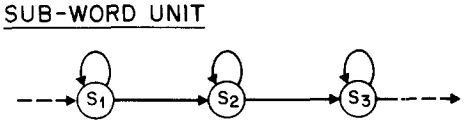

Fig. 4. HMM representation of subword model.

ment. The performance obtained with discrete or semi-continuous densities is often comparable to or only slightly lower than that obtained with continuous densities; often at significantly reduced computation rates.

For continuous density modeling in the Bell Labs system described in Section 2 uses both an observation probability density function (for each state) represented by a weighted sum of M multivariate Gaussian density functions with a diagonal covariance matrix, and an energy histogram representing the log probability of observing a frame with a given log-energy. All subword unit models have three states except the model for silence which has only one state. Furthermore no transition probabilities are used.

3.2. Training of subword unit models

Training of subword unit models consists of estimating the HMM parameters from a labeled training set of continuous speech utterances where all of the relevant subword units are known to occur "sufficiently" often. The training problem is another key aspect of the system, as the way in which training is performed affects greatly the overall recognition system performance.

The first issue of note is the size of the training set. The optimal training set size is infinity – i.e. the more training material that is used, the higher the reliability of the resulting speech models. Since infinite size training sets are impossible to obtain (and computationally unmanageable), we must use a finite size training set. This immediately implies that some subword units will occur much less often than others (at least in any natural recording this will be the case). Hence we immediately see a tradeoff between using fewer subword units (where we get better coverage of individual units, but poor resolution as to linguistic context), and more subword units (where we get poor coverage of the infrequently occurring units, but improved resolution of linguistic context).

A second issue is the choice of training material. For a given amount of training material, the best coverage is obtained when the statistics of occurrence of the training set units match those of the recognition task; i.e. the training set sentences should come from the same linguistic material as the recognition task (i.e. same vocabulary, same language model). However, in such a case, the universality of the resulting speech models is poor; i.e. the same models may perform poorly on a totally different recognition task because of poor coverage of subword units for the new task. Hence the issue of "task dependent" training, which attempts to maximize performance for a given task, versus "task independent" training, which maximizes performance for any task has to be addressed. Most systems use task dependent training – we will present results on both types of training in this paper.

3.3. Adaptation of subword unit models

An alternative to using a large training set is to use some initial set of subword unit models and adapt them over time (with new training material, possibly derived from actual test utterances) to the speaker and speaking environment. In principle adaptation can be performed on lexical, syntactical and semantic models. We will focus our discussion only on adaptive training of subword acoustic models. Such methods of adaptive training are reasonable for new speakers, vocabularies, transducer or environments, and will be shown later to be an effective way of bootstrapping a good set of specific models from a more general set of models.

For adaptive training of subword unit models, we assume a set of initial models, called *seed models*, is available. The seed model can be a set of speaker-independent subword models or a set of gender-dependent subword models. Based on a small number of adaptation utterances, the adaptation algorithm attempts to combine the seed models with the adaptation data and generate a speaker adaptive model. In doing so, the dispersed seed models (e.g. the speaker independent models), which were designed to cover a wide range of speaking environments and a large number of speakers in the test population, are modified by the adaptation data (e.g. speaker-specific, application-specific utterances) so that a more focused set of models is created. The adaptive models are therefore useful in a more specific environment for a smaller number

of speakers whose speech has similar acoustic characteristics to those of the adaptation data. In our HMM-based system, we used the *segmental MAP algorithm* (Gauvain and Lee, 1992a, 1992b) to perform adaptive training. Given the acoustic data and the word transcriptions of the adaptation utterances, the Viterbi algorithm is first used to segment the adaptation utterances into subword segments using the seed models. The MAP estimation algorithm is then used to obtain the parameters of the adapted subword models. In essence, the MAP estimate is a weighted sum of the prior parameters and the statistics of the adaptation data (Lee and Gauvain, 1992). The weights are functions of both the prior parameters and the adaptation data, and are recomputed in a nonlinear manner using the *expectation-maximization* (EM) algorithm.

4. RM task and experimental setup

The DARPA Resource Management (RM) task (Price et al., 1988) is a database access and retrieval task based on information about properties of battleships throughout the world. The task vocabulary contains 991 words and the language perplexity is about 9 (i.e. considerably smaller than that of the WP or NG FSNs). Three training sets were used in this study to create subword models. The first, referred to as SI-109, consisted of approximately 3990 read sentences from 109 talkers (30–40 sentences per talker) as provided by NIST. It was used for speaker-independent model training. The second training set, referred to as SD-600, was designed for speaker dependent model training and can be used for speaker adaptation experiments. It consisted of 600 sentences from each of 12 talkers, 6 females and 6 males, as provided by NIST. None of the 12 talkers in the SD-600 set was in the SI-109 training set. The third training set, referred to as GE-10000, was obtained from a set of 10,000 general English utterances recorded at CMU. It was designed for creating subword unit models from non-task specific sentences (Hon, 1992). [3]

[3] These sentences were recorded at CMU and graciously provided to AT&T by the speech group at CMU.

The original data provided by NIST and CMU were sampled at 16 kHz. For our experiments, we bandpass filtered the speech data and down-sampled them to 8 kHz. This makes the speech data compatible with a standard telephone bandwidth signal however this down-sampling procedure has the potential of lowering overall system performance because the information in high band was discarded. In a recent study (Lamel and Gauvain, 1992), it was found that recognition, based on mel frequency cepstral coefficients generated from 16 kHz speech data, gave significantly better performance than that obtained with both mel frequency and LPC-derived cepstral coefficients generated from the downsampled 8 kHz data. However, there was no observed performance difference between the system using either the mel frequency cepstral coefficients or LPC-derived cepstral coefficients generated from the 8 kHz data. We also conducted a similar study and found that using the 16 kHz signal resulted in 30% fewer word errors than those obtained from using the 8 kHz signal. In this study, we only present results obtained with the LPC-derived cepstral coefficients using the 8 kHz signal.

The recognizer was implemented as a large FSN with something on the order of 20,000 HMM states and word junction nodes to keep track of at each frame of the input. To reduce computation, a frame synchronous beam search algorithm was used (Lowerre and Reddy, 1980, Lee et al., 1990), in which the best accumulated likelihood score, L^*, was determined at each frame, and based on the beam width Δ, all nodes whose accumulated likelihoods were less than $(L^* - \Delta)$ were eliminated from a list of active nodes (i.e. these paths were no longer followed). In order to prevent an excessive number of insertions of short function words, a word insertion penalty was used in the Viterbi decoding at the end of each word arc. By adjusting the value of the word penalty we can balance the word insertion and word deletion error rates. Appropriate values of word penalty are determined experimentally.

In the following two sections we present some recent experimental results on the RM task. In Section 5, we report on speaker-independent (SI) recognition results obtained with both CI and CD unit models. We also evaluate system perform-

ance based on both task-independent (TI) and task-dependent (TD) training materials. By TD training material, we mean the training sentences are designed to cover typical sentences used for a particular task. While the TI training material are sentences with no task specification. It is generally agreed that TD training usually outperforms TI training in speech recognition under normal testing conditions. The TD training material used in this study is the commonly known SI-109 RM training set and the TI training material is the CMU GE-10000 training set. Comparisons of speaker-independent, speaker-dependent (SD) and speaker-adaptive (SA) recognition results will be given in Section 6.

5. Speaker-independent recognition

We used 4 different sets of test data to evaluate speaker-independent recognition performance. The test set names reflect the date that DARPA distributed the testing material. The first three sets are referred to as feb89, oct89 and feb91. Each of the sets consisted of 30 test utterances from non-overlapping sets of 10 talkers, none of whom was in the 109 talker training set. The fourth set, referred to as jun90, consisted of a set of 120 sentences from each of 4 new talkers, none of whom was in the 109 talker training set.

We contrast speech recognition performance based on choice of subword units (context dependency) and selection of training material (task dependency). The combination of the two factors gives four sets of experimental conditions. The recognition results are reported in the following four subsections, respectively.

5.1. Context-independent, task-dependent units

The set of 47 CI units were trained from the 3990 training sentences for the DARPA task (SI-109 set). Hence these units are task dependent units. HMMs were built with continuous mixture densities with up to 256 mixtures per state. Recognition tests were then performed using each of the four independent test sets with both the WP and NG grammars with no semantic processing. The results of these baseline tests, in terms

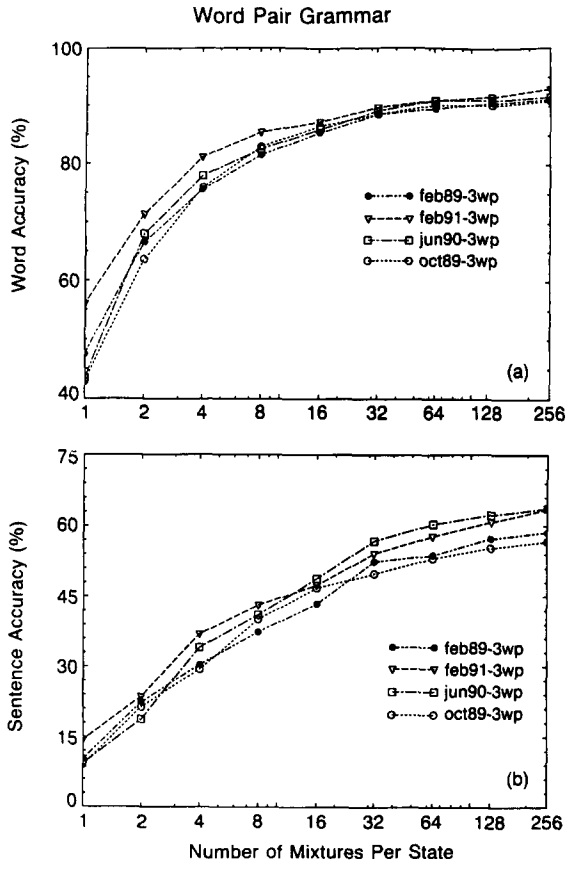

Fig. 5. Word and sentence accuracy versus the maximum number of mixture components per state using CI/TD units for the WP grammar.

of word and sentence accuracies, are given in Figures 5 and 6. It can be seen that, although there are detailed differences in performance among the different test sets (especially for small numbers of mixtures per state), the performance trends are essentially the same for all test sets. In particular, we see that with the WP grammar, the range of word accuracies for 1 mixture (Gaussian) per state is 42.9% (feb89) to 56.0% (jun90), whereas for 256 mixtures per state the range is 90.9% (feb89) to 93.0% (jun90). For the NG grammar, the range of word accuracies for 1 mixture per state is 20.1% (feb91) to 28.5% (jun90) and for 256 mixtures per state it is 68.5% (oct89) to 70.0% (feb91). Therefore, it is clear that high recognition accuracy can be achieved with simply the CI unit models provided each

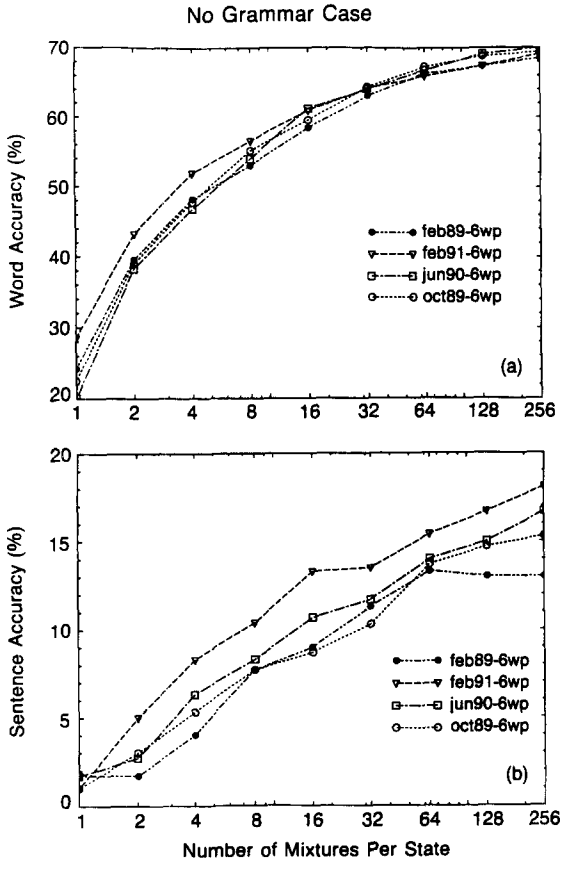

Fig. 6. Word and sentence accuracy versus the maximum number of mixture components per state using CI/TD units for the NG grammar.

acoustic HMM state is characterized with an enough number of Gaussian mixture components.

5.2 Context-independent, task-independent units

The same set of 47 CI units were trained from the set of GE-10000 sentences which were totally unrelated to the RM task (i.e. different vocabulary, different syntax, etc.). Tests were run using HMMs with a maximum number of 64 and 256 mixture components per state and the results are shown in Table 2.

By comparing the results in Table 2 with those in Figures 5 and 6 it can be seen that the word accuracy falls from about 92% with TD units to about 83% with TI units for the WP case, and from about 69% with TD units to 56% with TI units for the NG case when using models with 256 mixtures per state. Hence there is a significant loss of performance – even for the case of 47 CI units. One reason for this loss of performance is due to possible differences in recording environments between the CMU data and the RM testing data recorded at Texas Instruments. The second reason for the loss of performance is due to the word context mismatch, as discussed previously. Another reason for the loss of performance is the fact that it is considerably more difficult to define the linguistic content of the GE-10000 sentence training set because of the generality of the sentences. Hence we used a set of isolated word pronunciations, of the words in the 10,000 sentence training set. For many, if not most, of the sentences, this formal pronunciation is grossly inadequate. We did try to modify the word pronunciations based on rules appropriate for a speech synthesizer – but did not find any performance improvement for the recognition tests.

5.3. Context-dependent, task-dependent units

Based on the DARPA RM training set we created several sets of subword units that were context-dependent. Each set contained right-and-left context units, right-context units, left-context units and context-independent units. The

Table 2
Word accuracies using CI/TI units

Number of mixtures per state	Grammar	Test set				Average
		feb89	oct89	jun90	feb91	
64	WP	79.8	80.3	87.7	82.0	82.5
256	WP	81.5	80.1	88.7	82.0	83.1
64	NG	52.8	51.3	61.6	53.2	54.7
256	NG	54.1	53.2	64.2	53.6	56.3

Table 3
Word accuracies using CD/TD units

Grammar	Number of interword units	Test set				Average
		feb89	oct89	jun90	feb91	
WP	1769	96.3	95.8	95.6	96.5	96.0
WP	2421	95.3	95.9	95.7	96.7	95.9
NG	1769	81.6	79.3	80.4	81.1	80.6
NG	2421	80.5	79.9	80.4	81.8	80.7

criterion for including a CD unit was that it occurred sufficiently often in the training set so that enough data can be used to model the unit. We found, experimentally, that a particular unit should appear at least between 20 and 30 times in the training data in order to be reliably modeled. We call this criterion the *unit selection rule* (Lee et al., 1990) and the occurrence count in the rule the *unit selection threshold*. We designed CD unit sets based both on intraword and interword units. We also constructed male and female models for each unit set using the segmental MAP algorithm (Gauvain and Lee, 1992a, 1992b). The results on the 4 test sets are shown in Table 3.

The results in Table 3 should be compared with the results in Figures 5 and 6. It can be seen that word accuracies improve significantly using CD/TD units when context-dependent units and interword units are used and gender dependency is also incorporated in modeling. Compared with the CI/TD results, the gender-dependent CD/TD model sets reduce the word error rate by about 50% for the WP case and by about 34% for the NG case. Detailed comparison of results obtained with intraword CI units and results obtained with intraword CD units, interword CD units and gender dependent units can be found elsewhere (Lee et al., 1990; Gauvain and Lee, 1992b). It is also noted that the two sets of units in Table 3 give virtually similar recognition results. Since no smoothing is used in our modeling framework, the unit model reliability depends on getting enough observations for the particular unit. When the number of units increases, the number of training vectors for each unit reduces accordingly. There is a tradeoff between model reliability and unit coverage. This is still a research issue yet to be solved.

5.4. Context-dependent, task-independent units

Using the GE-10000 sentence training set, we again created several sets of subword units that were context dependent, in a manner similar to that used for TD units. However, only intraword units were considered. A major problem associated with creating CD/TI units is that many of the units never occur in the RM task vocabulary. Hence, whenever training data is used to create a CD/TI unit that is not used in the RM task, you essentially are reducing the size of the relevant training set. To alleviate this problem, partially, it is possible to post-process the set of CD/TI units to remove all such units that do not occur in the

Table 4
Word accuracies using CD/TI units

Grammar	Number of intraword units	Test set				Average
		feb89	oct89	jun90	feb91	
WP	1030/modified	85.6	85.5	90.7	88.0	87.5
WP	1418/original	85.4	85.6	90.8	87.5	87.3
WP	1136/original	83.4	83.8	91.3	87.1	86.9
NG	1418/original	60.0	58.8	68.3	60.5	61.9
NG	1030/modified	60.0	58.3	68.4	59.9	61.7
NG	1136/original	60.0	58.2	68.1	60.1	61.6

RM task, and to reassign those units to the "equivalent" CI/TI unit. We called this set of reduced units the *modified* unit set. In this manner all the training data is used to create the CD/TI units.

Based on the same unit selection rule (Lee et al. 1990), using unit selection thresholds of 75 and 100 occurrences of each unit were found to provide the best recognition performance for this task. Using a unit selection threshold of 75 gave a set of 1418 original units and 1030 modified units; a unit selection threshold of 100 gave a set of 1136 units. The recognition performance, on all 4 test sets, using the CD/TI units (both the original set and the modified set of units), is shown in Table 4.

It can be seen from Table 4 that the difference in performance between CD/TI units modified by task information and CD/TI units without task information is essentially negligible, i.e. on the order of 0.2% for both the WP and NG cases. It is also seen that the improvement in performance is about 4.4% (26% reduction in error rate) for the WP case, and 5.6% (13% reduction in error rate) for the NG case, as compared to the results using CI/TI units in Table 2. Considering that no interword units were modeled and no gender-dependent models were used the CD/TI results shown in Table 4 are still reasonable when compared with the CD/TD results shown in Table 3.

6. Speaker-adaptive recognition

Perhaps the ultimate way to create (train) subword units is to adapt them both to the task and to the speaker. In cases where an individual speaker is able to provide sufficient training, this type of subword unit learning is capable of providing the highest performance scores.

In order to incorporate speaker adaptation you need both an initial set of models (seed models) and an adaptive learning algorithm which can incorporate prior knowledge (as embodied in both the initial models and in assumptions about the distributions of the model parameters being adapted), and can perform parameter smoothing and interpolation. The segmental MAP algorithm proposed by Gauvain and Lee (1992a) is an ideal

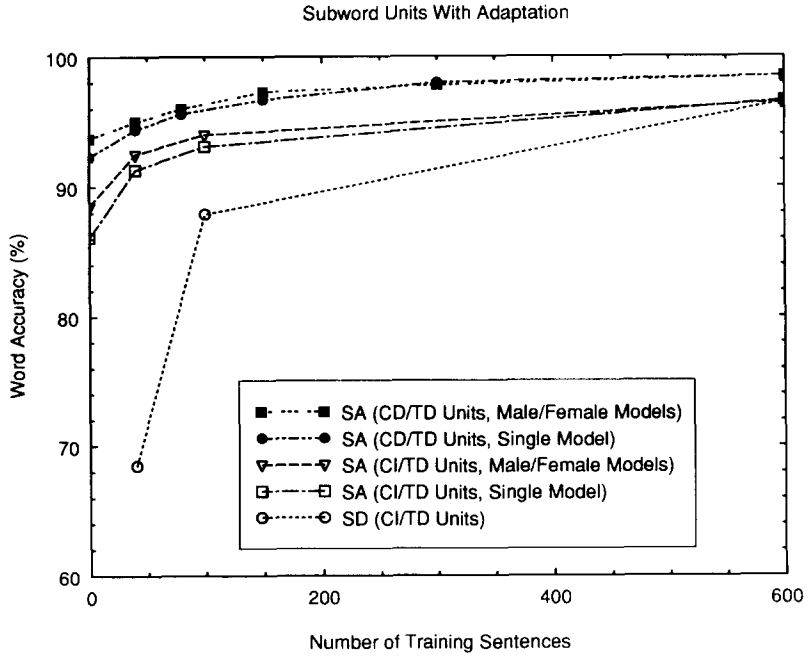

Fig. 7. Word accuracy versus number of training/adaptation sentences for different initial subword unit models.

candidate for performing speaker adaptation. It was used on the RM task based on the separate training set of 600 sentences (about 30 minutes of training material) by each of 12 talkers in the speaker dependent part of the RM database. The testing data used in the experiments discussed in this section is an independent test set of 25 sentences by each of these 12 talkers. It was distributed for RM evaluation in February 1991. We refer to this set as the feb91-sd set. Six initial sets of seed models were used, including

1. a single set of CI/TD models with 47 subword units;
2. a pair of CI/TD models, one for male speakers and one for female speakers, each with 47 subword units;
3. a single set of CD/TD models with 1769 subword units (both intraword and interword units were used);
4. a pair of CD/TD models, one for male speakers, and one for female speakers, each with 1769 subword units;
5. a single set of CI/TD models with no bootstrapping, i.e. trained entirely on the speaker dependent sentences;
6. a single set of CI/TI models with 47 subword units.

Details of the segmental MAP algorithm and issues related to speaker adaptation based on the segmental MAP algorithm are given in (Lee and Gauvain, 1992). The TD seed models were obtained with the RM SI-109 speaker independent training set and the TI seed model was obtained from the GE-10000 training set. For each of these models, adaptation (or initial learning in the case of the fifth model) was performed using 40, 80 (models 4 and 5), 100 (models 1–3), 150 (models 4 and 5), 300 (models 4 and 5) and 600 sentences. For each adapted model we measured word accuracy on the independent test set and the results using the WP grammar are shown in Figure 7. It can be seen that for models 1, 2 and 5, where 47 CI/TD units were used, the adapted models all converged to a 96.5% word accuracy when all 600 training sentences were used in the adaptation. Model 5, which did not use a bootstrap model, converged at the fastest rate and had word accuracies significantly lower than models 1 and 2 until all 600 training sentences were used. The differences in word accuracy resulting from single models and male/female models were small, but not insignificant for short adaptation sets.

Table 5
Comparison of SI/SD/SA word accuracies

Training	0	40	100	600
SD	–	68.5	87.9	96.5
SA (SI)	86.1	91.3	93.1	96.6
SA (M/F)	88.5	92.5	94.0	96.5
SA (TI)	74.0	89.1	92.4	95.9

The recognition results based on speaker independent (SI), speaker dependent (SD) and speaker adaptive (SA) CI models are compared in Table 5. Three different training/adaptation sets, 40, 100 and 600 utterances, respectively, were tested. The case with no training data is labeled 0 in the first row of Table 5. No performance result is reported for the SD case with no training data because obviously no SD model can be created in this case. The results given for SA models without any adaptation data are simply the results obtained with the seed models.

The word accuracy for 40 utterances of SD training was 68.5% which is not acceptable for any reasonable application. The SI word accuracy (0 minutes of adaptation data) was 86.1%, comparable to the SD results with 100 utterances of SD training. The SA models perform better than SD models when a relatively small amount of data was used for training or adaptation. When all the available training data were used, the SA and SD results were comparable, consistent with the MAP adaptation formulation where that the MAP and the MLE estimates are *asymptotically similar* (Gauvain and Lee, 1992a). With 40 utterances of adaptation data, the SA results gave a 37% word error reduction over the SI results. It was also noted that a larger improvement was observed for the female speakers (51% word error reduction) than for the male speakers (22% word error reduction).

Speaker adaptation can also be performed starting from gender-dependent models (fourth row of Table 5). The word accuracy with no speaker adaptation was 88.5%. The accuracy rates were increased to 92.5% and 94.0% with 40 and

100 utterances of adaptation data, respectively. Comparing the third and the fourth rows in Table 5 it can be seen that when only a small amount of adaptation data is used, the best results were obtained with gender-dependent seed models. The word error reduction with 40 adaptation utterances was 35% compared to the no adaptation results with gender-dependent models. Moreover, the improvement was 46% compared to the SI recognition results.

The adaptation results shown in the last row were obtained with the GE-10000 TI seed model. It can be seen that the results were inferior to those obtained with the SI-109 TD seed models. The difference in performance may arise from differences in the recording environments for the two databases as well as from different lexical representations for the words in the RM and the GE databases. The performance difference was the greatest when no adaptation data were used. Using more adaptation data reduced the difference in performance. Even though some performance degradation was observed, the advantage of using a *universal acoustic model* generated from a large speech database cannot be overlooked. It is unlikely that one can collect enough training material for every conceivable recognition application so that the trained models can handle any speaker in any speaking environment. A more attractive approach is to start with a universal acoustic model. For a given application, *vocabulary learning* is first performed to extract "relevant" subword units for the particular application vocabulary. Then a small number of adaptation sentences is collected from the user and these data are used to construct speaker adaptive models for the speaker in the particular environment for that specific application. By doing so, the acoustic mismatch problems between training and testing, including speaker mismatch, transducer mismatch and channel mismatch, can generally be minimized. Once an initial speaker adaptation model is obtained for a user, the model can be continuously adapted using *sequential* and *on-line adaptation* schemes. We believe that it is possible to construct a good universal acoustic model from a large pool of training data. How to design such a universal acoustic database is still an open research topic.

The word accuracies for models 3 and 4, where CD/TD units were used, were significantly higher than those obtained using only CI/TD units. Again both models converged to a 98.5% word accuracy when all 600 training sentences were used for adaptation. Also the adapted models, based on using separate male/female models, gave better performance than the adapted models based on a single set of gender-independent models until about 300 training sentences were used for adaptation.

7. Summary

In this paper we have described one of the Bell Labs systems for large vocabulary continuous speech recognition, and discussed the key issues in design and implementation of the system. We have shown that the choice and method of training of the basic subword units is critical, and that a wide range of options exist. We have only presented results comparing the use of CI and CD units based on their frequencies of occurrence in the training data. We have also studied two methods of parameter estimation, namely the maximum likelihood and maximum a posteriori methods. Each choice of subword units, in combination with each method of parameter estimation, has distinct advantages and disadvantages; hence there is no "ideal" or "optimal" set of units, but instead one must consider a wide range of possibilities.

We have explored different ways of incorporating context and task dependency in acoustic modeling. It is concluded that the recognition accuracy of a task can be increased when context dependency is properly incorporated to reduce the acoustic variability of the speech units to be modeled. We have found that context-dependent units provide better recognition performance than context-independent units. We have also shown that interword units take into account cross-word coarticulation and therefore provide more accurate modeling of speech units than intraword units in fluently spoken continuous speech. Similarly, the acoustic variability of speech units can further be reduced when gender dependency is considered in the design of acoustic models for

Table 6
Word accuracies (%) summary for CI/CD and TD/TI units

Grammar	CI/TD	CI/TI	CD/TD	CD/TI
WP	91.7	83.1	96.0	87.3
NG	69.3	56.3	80.6	61.9

the set of speech units. Gender-dependent models usually give better performance than that obtained with gender-independent models at a slightly higher computational cost. We have also shown that, for a given task, speech unit models trained based on task-dependent training data always outperform models trained with task-independent training data. A comparison of the performance for speaker independent recognition of the Resource Management task is shown in Table 6. The word accuracies given are based on testing 1380 utterances from 34 new speakers not contained in the training set.

Most of the results presented in this study are obtained with the Resource Management task using the WP and NG covering grammars. We have also experimented with the perplexity 9 full grammar by performing speech recognition first with a covering grammar then followed by a semantic post-processor (Pieraccini and Lee, 1991) to correct obvious word errors. By incorporating the simple set of semantic and syntactic rules for the RM task in this two-pass recognition, we have achieved over 99% word accuracy and 92% string accuracy on a random subset of 300 testing utterances. In addition to the RM task, many other subword-based studies have also been carried out. For example, we have applied the same subword-based approach to the problem of connected digit recognition and obtained very high performance on the TI connected digit database (Gauvain and Lee, 1992c). We have implemented the ATIS speech understanding task and good performance has been obtained (Pieraccini et al., 1992) for recognition of spontaneously spoken utterances. Finally, the same subword-based approach has also been applied to speaker verification to enhance flexibility of verification systems (Rosenberg et al., 1990).

The problems of large vocabulary continuous speech recognition are far from solved. Key issues include the need to eliminate specification of a finite task vocabulary, and a rigid task syntax. As a result, modern systems attempt to use natural language front ends with essentially unlimited vocabulary and syntax. This type of system implies an entirely different system implementation with a completely new set of problems associated with unknown words, non-grammatical constructions, extraneous speech, etc. On top of this, the "traditional" problems associated with noisy environments, speaker variability, transmission system variability, etc. remain, along with the need to improve the acoustic front end signal processing, and to provide efficient search strategies for large applications.

References

A. Averbuch et al. (1987), "Experiments with the Tangora 20,000 Word Speech Recognizer", *Proc. Internat. Conf. Acoust. Speech Signal Process., Dallas, TX*, pp. 701–704.

J. Baker et al. (1992), "Large vocabulary recognition of Wall Street Journal sentences at Dragon system", *Proc. DARPA Speech and Natural Language Workshop, Harriman, NY*, pp. 387–392.

J.R. Bellegarda and D. Nahamoo (1990), "Tied mixture continuous parameter modeling for speech recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-38, No. 12, pp. 2033–2045.

S.B. Davis and P. Mermelstein (1980), "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-28, No. 4, pp. 357–366.

L. Deng, V. Gupta, M. Lennig, P. Kenny and P. Mermelstein (1990), "Acoustic recognition component of an 86,000-word speech recognizer", *Proc. Internat. Conf. Acoust. Speech Signal Process., Albuquerque, NM*, pp. 741–744.

F. Fissore, P. Laface, G. Micca and R. Pieraccini (1989), "Lexical access to very large vocabulary", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-37, No. 8, pp. 1197–1213.

S. Furui (1986), "Speaker-independent isolated word recognition using dynamic features of speech spectrum", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-34, No. 1, pp. 52–59.

J.-L. Gauvain and C.-H. Lee (1992a), "MAP estimation of continuous density HMM: Theory and applications", *Proc. DARPA Speech and Natural Language Workshop, Harriman, NY*, pp. 185–190.

J.-L. Gauvain and C.-H. Lee (1992b), "Bayesian learning for hidden Markov model with Gaussian mixture state observation densities", *Speech Communication*, Vol. 11, Nos. 2–3, pp. 205–213.

J.-L. Gauvain and C.-H. Lee (1992c), "Improved acoustic

modeling with Bayesian learning", *Proc. Internat. Conf. Acoust. Speech Signal Process., San Francisco*, pp. 481–484.

E. Giachin, C.-H. Lee and A.E. Rosenberg (1991), "Word juncture modeling using phonological rules for HMM-based continuous speech recognition", *Comput. Speech Language*, Vol. 5, No. 2, pp. 155–168.

C.T. Hemphill, J.J. Godfrey and G.D. Doddington (1990), "The ATIS spoken language system pilot corpus", *Proc. DARPA Speech and Natural Language Workshop, Hidden Valley, PA*, pp. 96–101.

L. Hirshman and MADCOW Group (1992), "Multi-site data collection for a spoken language corpus", *Proc. ICSLP-92, Banff, Canada*, pp. 903–906.

H.-W. Hon (1992), Vocabulary-independent speech recognition: The VOCIND system, Ph.D. Thesis, School of Computer Science, Carnegie Mellon University.

X.D. Huang, Y. Ariki and M.A. Jack (1990), *Hidden Markov Models for Speech Recognition* (Edinburgh Univ. Press, Edinburgh).

X.D. Huang et al. (1993), "The SPHINX-II speech recognition system: An overview", *Comput. Speech Language*, Vol. 7, No. 2, pp. 137–148.

F. Jelinek (1985), "The development of an experimental discrete dictation recognizer", *Proc. IEEE*. Vol. 73, pp. 1616–1624.

L.F. Lamel and J.-L. Gauvain (1992), "Continuous speech recognition at LIMSI", *Proc. DARPA Continuous Speech Recognition Workshop, Stanford,* pp. 77–82.

K.-F. Lee (1989), *Automatic Speech Recognition – The Development of the SPHINX-System* (Kluwer Academic Publishers, Boston).

C.-H. Lee and J.-L. Gauvain (1992), "A study on speaker adaptation for continuous speech recognition", *Proc. DARPA Continuous Speech Recognition Workshop, Stanford,* pp. 59–64.

C.-H. Lee, L.R. Rabiner, R. Pieraccini and J.G. Wilpon (1990), "Acoustic modeling for large vocabulary speech recognition", *Comput. Speech Language*, Vol. 4, No. 2, pp. 127–165.

C.-H. Lee, E. Giachin, L.R. Rabiner, R. Pieraccini and A.E. Rosenberg (1992), "Improved acoustic modeling for large vocabulary continuous speech recognition", *Comput. Speech Language*, Vol. 6, No. 2, pp. 103–127.

A. Ljolje and M.D. Riley (1992), "Optimal speech recognition using phone recognition and lexical access", *Proc. ICSLP-92, Banff, Canada*, pp. 313–316.

B. Lowerre and D.R. Reddy (1980), "The HARPY speech understanding system", *Trends in Speech Recognition*, ed. by W. Lea (Prentice Hall, Englewood Cliffs, NJ), pp. 340–346.

T. Morimoto, H. Iida, A. Kurematsu, K. Shikano and T. Aizawa (1990), "Spoken language: Towards realizing an automatic telephone interpretation", *Proc. INFO Japan 90*, pp. 553–559.

H. Murveit et al. (1989), "SRI's DECIPHER system", *Proc. DARPA Speech and Natural Language Workshop, Philadelphia, PA*, pp. 238–242.

H. Ney and A. Paeseler (1988), "Phoneme-based continuous speech recognition results for different language models in a 1000-word SPICOS system", *Speech Communication*, Vol. 7, No. 4, pp. 367–374.

D.B. Paul (1989), "The Lincoln robust continuous speech recognizer", *Proc. Internat. Conf. Acoust. Speech Signal Process., Glasgow*, pp. 449–452.

D.B. Paul and J.M. Baker (1992), "The design for the Wall Street Journal-based CSR corpus", *Proc. ICSLP-92, Banff, Canada*, pp. 899–902.

R. Pieraccini and C.-H. Lee (1991), "Factorization of language constraints in speech recognition", *Proc. ACL-91*, Berkeley, CA.

R. Pieraccini et al. (1992) "A speech understanding system based on statistical representation of semantics", *Proc. Internat. Conf. Acoust. Speech Signal Process., San Francisco*, pp. 193–196.

P.J. Price, W. Fisher, J. Bernstein and D. Pallett (1988), "A database for continuous speech recognition in a 1000-word domain", *Proc. Internat. Conf. Acoust. Speech Signal Process.*, New York, pp. 651–654.

L.R. Rabiner (1989), "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE*, Vol. 77, pp. 257–286.

D.B. Roe, F.C. Pereira, R.W. Sproat and M.D. Riley (1992), "Efficient grammar processing for a spoken language translation system", *Proc. Internat. Conf. Acoust. Speech Signal Process., San Francisco*, pp. 213–216.

A.E. Rosenberg, C.-H. Lee, F.K. Soong and M.A. McGee (1990), "Experiments in automatic talker verification using sub-word unit hidden Markov models", *Proc. ICSLP-90*, Kobe, Japan.

S. Sagayama et al. (1992), "ATREUS: Continuous speech recognition systems at ATR interpreting telephony research laboratories", *Proc. Australian Speech Science & Technology Conf. SST-92*, Brisbane, Australia, pp. 324–329.

R. Schwartz et al. (1989), "The BBN BYBLOS continuous speech recognition system", *Proc. DARPA Speech and Natural Language Workshop, Philadelphia*, pp. 94–99.

R. Schwartz et al. (1992), "New uses for the *N*-best sentence hypotheses within the BYBLOS speech recognition system", *Proc. Internat. Conf. Acoust. Speech Signal Process., San Francisco*, pp. 1–4.

F.K. Soong and A.E. Rosenberg (1988), "On the use of instantaneous, and transitional spectral information in speaker recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 36, No. 6, pp. 871–879.

M. Weintraub et al. (1989), "Linguistic constraints in hidden Markov model based speech recognition", *Proc. Internat. Conf. Acoust. Speech Signal Process., Glasgow*, pp. 699–702.

P.C. Woodland and S.J. Young (1992), "Benchmark DARPA RM results with the HTK portable HMM ToolKit", *Proc. DARPA Continuous Speech Recognition Workshop, Stanford,* pp. 47–52.

V. Zue, J. Glass, M. Phillips and S. Seneff (1989), "The MIT summit speech recognition system: A progress report", *Proc. DARPA Speech and Natural Language Workshop, Philadelphia, PA*, pp. 179–189.