Subword-Based Large-Vocabulary Speech Recognition

Chin-Hui Lee
Jean-Luc Gauvain
Roberto Pieraccini
Lawrence R. Rabiner

During the past several years, research in large-vocabulary speech recognition has been intensively carried out worldwide, encouraged by advances in algorithms, architecture, and hardware. In the United States, the defense advanced-research projects agency (DARPA) spoken-language-processing community has focused its efforts on studying several systems. These include the 991-word naval resource management (RM) speech-recognition task, the open-vocabulary, spontaneous-speech, air-travel information system (ATIS) speech-understanding task, and the 20,000-word *Wall Street Journal* (WSJ) dictation task. Although researchers have learned a great deal about how to build and efficiently implement large-vocabulary speech-recognition systems, many fundamental questions remain, for which there are no definitive answers. This paper focuses on the basic structure of a large-vocabulary speech-recognition system, considerations in choosing a set of subword units, method of "training," integration of a language model, and implementation of a complete system. The paper also reports on some recent results, obtained at AT&T Bell Laboratories, on the DARPA RM task.

Introduction

In the past few years, a significant portion of speech-recognition research has been devoted to studying the problems of building and implementing a large-vocabulary, continuous, speech-recognition system. For the most part, DARPA has been responsible for stimulating this effort, and it has funded research on three large-vocabulary recognition (LVR) tasks—namely, the naval RM task[1], the ATIS task[2,3], and the WSJ[4] task. In addition, there is worldwide interest in large-vocabulary speech recognition. This is due to potential applications in voice data-base access and management, voice dictation[5], and limited-domain spoken-language translation[6].

In Japan, large-vocabulary speech-recognition systems are mostly developed around the concept of interpreting telephony.[7,8] In Europe, the Philips SPICOS system[9], the CSELT system[10], and the LIMSI effort[11] are examples of current activity in large-vocabulary speech-recognition research. In Canada, the most notable LVR project is the INRS 86,000 isolated-word recognition system.[12] In the United States, in addition to LVR research at AT&T[13,14] and IBM[5] (IBM is a registered trademark of International Business Machines Corp.), most LVR effort is sponsored by DARPA. Projects include the Bolt, Beranek and Newman BYBLOS system[15], the Carnegie Mellon University SPHINX system[16], the Dragon WSJ system[17], the Lincoln Lab Tied-Mixture system[18], the Massachusetts Institute of Technology Summit system[19], and the Stanford Research Institute DECIPHER system[20].

Although some of the systems have been "trained" to individual speakers[6,21], most current large-vocabulary speech-recognition systems have the goal of continuous speech recognition of fluent input by any individual (speaker-independent systems).

The conventional approach to large-vocabulary speech recognition is basically one of statistical pattern recognition. The fundamental speech units use phonetic labels,

Panel 1. Abbreviations, Acronyms, and Terms

ATIS—air-travel information system

cepstrum—Fourier transform of the log spectrum of a signal

DARPA—defense advanced-research project agency

FSN—finite-state network

grammar—the set of rules governing word order in a sentence

HMM—hidden Markov model, a statistical modeling technique

LVR—large-vocabulary recognition

mel-frequency scale—a psychological scale of perceived tonal frequencies

NG—no grammar

phoneme—one of the set of the smallest units of speech that distinguish one utterance or word from another

PLU—phone-like unit, an acoustic description of a linguistically based unit, as represented in words occurring in a given "training" set

prosody—elements of meaning that are conveyed by features, such as stress and pitch, that span boundaries of sound segments. For example, in the word "window," the sound segments comprise /w/I/n/d/o/. The prosodic feature of stress is present across the first three segments.

RM—resource management, used in conjunction with the 991-word naval speech-recognition task

scalar—a quantity, such as time, completely specified by a number on an appropriate scale

semantics—a set of task-dependent rules that help determine the optimal sequence of words in a sentence

subword unit—a unit smaller than a word, such as a phoneme

task—a speech-recognition problem or application

"training"—creation of speech models, which are suitable for use in a recognition system

but are modeled acoustically, based on a lexical description of words in a vocabulary. In general, no assumption is made, *a priori*, about the mapping between acoustic measurements and subword linguistic units, such as *phonemes*. A system learns this mapping entirely by means of a finite, labeled, "training" set of utterances. The resulting speech units, which are known as *phone-like units* (PLUs), are essentially acoustic descriptions of linguistically based units, as represented in the words occurring in the given "training" set. This pattern-recognition approach offers the potential of modeling virtually all words and word sequences in a language, because the set of PLUs is usually chosen and designed to cover all phonetic labels of a particular language, and words in a language can usually be pronounced based on this set of fundamental speech units.

This subword-based modeling approach has two important advantages. First, it facilitates design of flexible vocabulary applications. An application's vocabulary can be reconfigured dynamically according to specific requirements, because all the words can be modeled approximately by a set of fundamental speech units. Second, subword-based modeling eases the collection difficulty of application-specific "training" data. In contrast to whole-word-based modeling, which requires "training" data containing a large number of occurrences for all vocabulary words in a particular application, the subword-based approach allows modeling of vocabulary words in cases where some—if not all—of the words do not appear in "training" data even once.

Although a great deal has been learned about how to build large-vocabulary speech-recognition systems, and how to implement them efficiently, there remains a whole range of fundamental questions for which there are no definitive answers. For example, the best ways to build and "train" fundamental subword units, from which word models are created, are not yet known. The best way to impose language constraints on a recognizer—to utilize all available knowledge—in the most computationally efficient manner, is also not yet known. The best way to implement a recognition system to maximize the probability of recognizing a spoken string—while minimizing the computation for string comparison—and searching through the recognition network, is not well understood. Ways to integrate suprasegmental information, such as *prosody* and duration, into existing recognition systems, which rely mainly on frame-level spectral information, are not understood. Ways to extract robust (relatively insensitive) features, so that recognition systems are less vulnerable to acoustic mismatch problems caused by talkers, transducers,

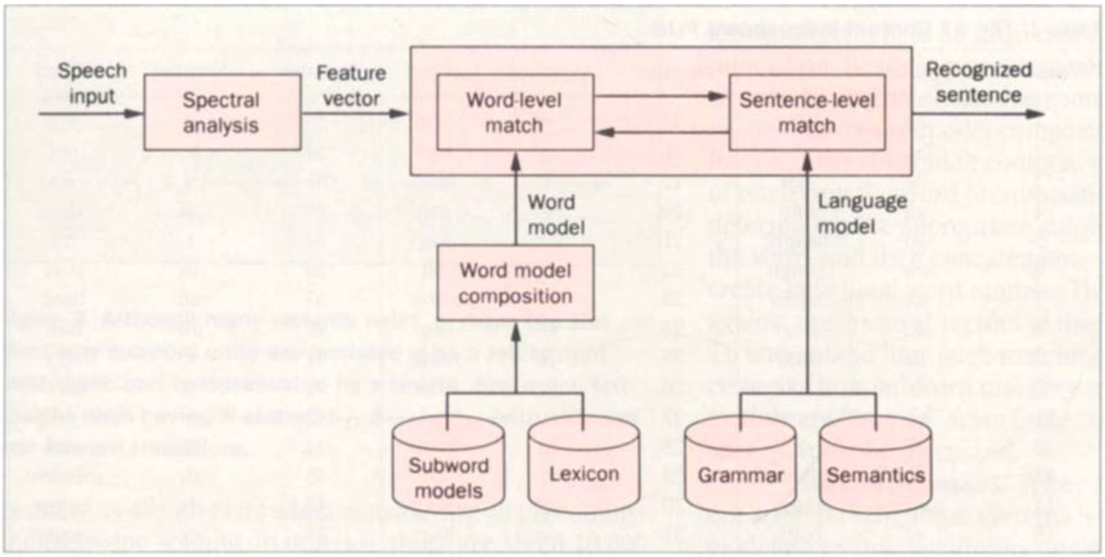

Figure 1. Block diagram of the continuous speech recognizer. The system consists of three main components: feature (or spectral) analysis, word-level acoustic matching, and sentence-level language matching.

channels, and speaking environments, are not yet known. Satisfactory solutions to portability problems, so that knowledge sources needed for creating new applications can be efficiently acquired, are also not yet understood. Most existing systems require acquisition of large amounts of application-specific acoustic and language "training" data to build application-specific systems. These requirements often limit development and performance of cross-domain applications.

Even though there is still a number of unresolved issues, the research community has made significant advances in large-vocabulary speech-recognition technology over the last few years. In the following sections, one particular version of a subword-based speech-recognition system is discussed in more detail. Also discussed is the current understanding about each system component, and where active research is attempting to learn the best way to implement the system component. Additionally, some recent results are presented on performance of the DARPA RM task, based on using so-called task-independent (or vocabulary-independent) units and adaptive "training" methods.

Baseline Speech-Recognition System

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 components: feature (or spectral) analysis, word-level acoustic matching, and sentence-level language matching. Feature analysis provides the acoustic feature vectors used to characterize the spectral properties of a time-varying speech signal. Word-level acoustic matching 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. Sentence-level matching uses a language model (based on a set of syntactic and semantic rules) to determine the most likely word sequence corresponding to a valid sentence in the task language.

Feature Analysis. The purpose of this component is to compute a set of spectral vectors, over time, that contain the relevant information—for recognition purposes—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 speech-recognition systems. Most systems use linear predictive coding (LPC) spectral analysis methods, based on fixed-size frames (for example, every 10-ms interval an analysis of a fixed frame of 30 ms of signal is performed). Typically, the LPC analysis provides a set of cepstral (Fourier transform of the log spectrum) coefficients for the frame. Sometimes, non-uniform frequency scales are used, giving the so-called *mel-frequency scale* cepstral coefficients.[22] (The mel-frequency scale is a psychologically based frequency scale, which is quasi-linear until about 1kHz and quasi-logarithmic above 1 kHz.) The rationale

Table I. The 47 Context-Independent PLUs.

Number	Symbol	Word	Number	Symbol	Word	Number	Symbol	Word
1	h#	silence	17	er	b*i*rd	33	p	*p*op
2	aa	f*a*ther	18	ey	b*a*it	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*ou*ght	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*o*ok
8	axr	din*er*	24	iy	b*e*at	40	uw	b*o*ot
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*nt*er
16	en	butt*on*	32	oy	b*oy*			

is that, because the human ear perceives frequencies on a non-uniform scale, it is 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 derivatives (first and second order) of the cepstral vector, as well as static information about the cepstrum.[23-26] Also, *scalars* representing frame energy and its derivatives are often used as part of the representation for each frame. For the system discussed in this paper, each 30-ms duration of speech (at an 8-kHz sampling rate) was analyzed 100 times a second (10-ms shift). This gives a spectral vector with 12 cepstral coefficients (on a uniform frequency scale), 12 first-order cepstral derivatives, 12 second-order cepstral derivatives, and first-order and second-order log energy derivatives. Hence, a spectral vector with 38 components was created every 10 ms throughout the signal.[27]

Word-Level Acoustic Matching. The basic element of this component is the set of subword models and the lexicon, as shown in Figure 1. Subword models are the representations of the fundamental speech units used as building blocks for words, phrases, and sentences. A great deal of research in large-vocabulary speech recognition has gone into defining these subword units in such a way that:

- They can be easily "trained" from finite "training" sets of speech material,
- They are robust (relatively insensitive) to natural variations in accents, word pronunciations, and test materials, and
- They provide high recognition accuracy for the required speech 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, modeling, and using various alternatives for subword units.

Perhaps the simplest set of subword units, and one which is widely used, is the set of basic phonemes of a given language. Although there is no complete agreement as to what sounds are part of this basic set—even for English—Table I 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 phone-like units (CI-PLUs). This is because the sounds are represented independently of the linguistic context in which they occur, and because the spectral properties of the sounds are "learned" from a "training" set, rather than postulated on the basis of linguistic unit features.

In contrast to the 47 CI-PLUs of Table I, one could consider subword units that are context dependent (CD). For example, two separate units could be defined, one for the sound /*ae*/ when preceded by /*f*/ and followed by /*t*/ (as in *fat*), and the other when preceded by /*b*/ and followed by /*t*/ (as in *bat*). In theory, there could be

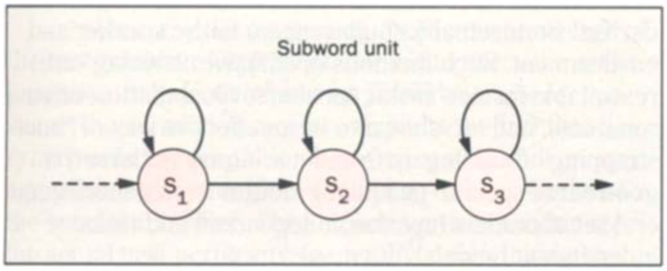

Figure 2. Although many variants exist, perhaps the simplest way subword units are modeled is as a left-to-right HMM. Each unit is represented by a simple, first-order, left-to-right HMM having *N* states, S_1, S_2, ..., S_N, with only *self* and *forward* transitions.

as many as 47^3 CD-PLUs when considering all preceding and following sounds. In practice, there are about 10,000 such possibilities—significantly less than the 100,000-plus count of 47^3, but substantially more than the 47 CI-PLU of Table I. Such CD-PLUs have been extensively used for large-vocabulary speech recognition.[7,28] For a given LVR task, practical methods are generally used to restrict the number of units to something on the order of 1,000 to 2,000.[13] The number of CD-PLUs required to achieve good performance depends on the vocabulary of a particular application and the speech material used in the "training" set.

The second basic element of the word-level acoustic matching component 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 creating a suitable word lexicon is the base (or standard) pronunciation of each word, as well as 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 speaker populations. Although there have been some very interesting experiments based on multiple-word-pronunciation lexicons[29], most large-vocabulary speech-recognition systems rely on a lexicon having only a single pronunciation for each word. This canonical representation of each word must be consistent with the subword units. Therefore, the representation's form changes as different sets of CD or CI subword units are used. Also, for function words, such as "the," "and," and "to," it is well known that there is no ideal canonical or standard pronunciation. A single representation for such function words will invariantly lead to some recognition problems.

The word-model composition part of the word-level acoustic matching component is simply the process of retrieving the word pronunciation from the lexicon, determining the appropriate subword units that make up the word, and then concatenating these subword units to create individual word models. These are then matched against the spectral vectors of the input speech signal. To understand how such matching takes place, the processes of how subword units are modeled, and how the models are "trained" from finite "training" sets of speech, must be discussed.

Subword unit models. A key to the success of modern speech-recognition systems is the use of statistical modeling techniques (for example, hidden Markov models [HMMs]) to represent basic subword units.[30] 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 2. In this case, each unit is represented by a simple, first-order, left-to-right HMM having N states, $S_1, S_2, ..., S_N$, with only *self* and *forward* transitions.

Within each state of the model, there is an observation-probability density function. This function specifies the likelihood (probability) of a spectral vector from the speech signal occurring within the model state. This observation density can be either a discrete density (implying the use of a common code book to quantize the input spectral vector[28]), a continuous density[13], or even what is called a semi-continuous density[31], or a tied-mixture density[32], which is a code book of continuous densities whose weights are chosen according to model state. Although continuous-density modeling usually provides the highest performance recognition systems, it requires the most computation to implement. Performance obtained with discrete or semi-continuous densities is often comparable to—or only slightly lower than—performance obtained with continuous densities, often at significantly reduced computation rates.

For continuous-density modeling, the baseline system in this study 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 are three-state, left-to-right models with no state skip. An exception is the model for silence, which has only one state. Furthermore, no transition probabilities are used, and self transitions and forward transitions from a state are assumed to be equally likely.

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

The first issue to note is "training" set size. Optimal "training" set size is infinity—that is, the more "training" material used, the higher the reliability of the resulting speech models. A finite-size "training" set must be used, because infinite-size "training" sets are impossible to obtain (and computationally unmanageable). This implies that some subword units will occur much less frequently than others (at least this will be the case in any natural recording). Therefore, a tradeoff exists between using fewer subword units (with better coverage of individual units, but poorer resolution as to linguistic context), and more subword units (with poorer 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 occurrence statistics of the "training"-set units match those of the recognition task. That is, the "training"-set sentences should come from the same linguistic material as the recognition task (same vocabulary and same language model). However, in such a case, the universality of the resulting speech models is poor; the same models may perform poorly on a totally different recognition task because of poor coverage of subword units for the new task. As a result, two types of "training" were used:

- Task-dependent (TD) "training," which attempts to maximize performance for a given task, and
- Task-independent (TI) "training," which maximizes performance for any task.

Most systems use TD "training." However, results of both "training" types are presented in this paper.

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 environment. Such methods of adaptive "training" are reasonable for new tasks, speakers, vocabularies, or environments, and are shown to be an effective way of "bootstrapping" (building up from an arbitrary initial set) a good set of specific (adaptive) models from a more general set of models (speaker-independent and task-independent models).

Sentence-Level Language Matching. This component 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 optimal sentence in the language—that is, the best word sequence, consistent with the grammar and semantics, which matches the input speech. The set of semantic and syntactic rules is usually specified by task requirements.

A number of different forms for the grammar has been proposed. These include context-free grammar, N-gram word probabilities, and word pair. Still, a simple grammar is assumed to exist, which can be represented as a finite-state network (FSN). This way, it is relatively straightforward to implement the grammar directly within the word-level acoustic matching component. In particular, for the DARPA RM task (991 words), either a word-pair (WP) grammar, which specifies explicitly—for each word in the vocabulary—what words are allowed to follow a vocabulary word, or a no-grammar (NG) grammar, in which it is assumed that every word in the vocabulary can follow every other word, has been used. The perplexities (average word-branching factor) of these two grammars are 60 for the WP case and 991 for the NG case. Thus, on average, there are about 60 possibilities following each word in the WP grammar, and 991 in the NG grammar.

The WP and NG grammar FSNs can produce any valid sentence in the task language. Unfortunately, they also can produce a large number of sentences that are not valid in the task language. The sentence *S*: "and and and," is valid for the NG network, but it is not valid for the RM task. Overcoverage (the ratio of sentences generated by the FSN to sentences valid in the task language) of FSNs is often extremely large. This is a negative feature of using these simple networks as the grammar network. On the other hand, using a full grammar (that is, no overcoverage) generally demands more computation to solve for the recognized sentence.

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, allowing meaningless sentences to be passed on to the semantic verifier. The semantic processor can use the actual task perplexity (generally, much lower than the perplexity of the artificial syntax network) to convert the recognized output to a semantically valid string.[33] 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 invalid. In practice, one of two simple strategies can be used: either the recognizer can generate a list of the best N sentences ($N = 500 - 1{,}000$) that the semantic processor can search until a semantically valid string is found; 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.

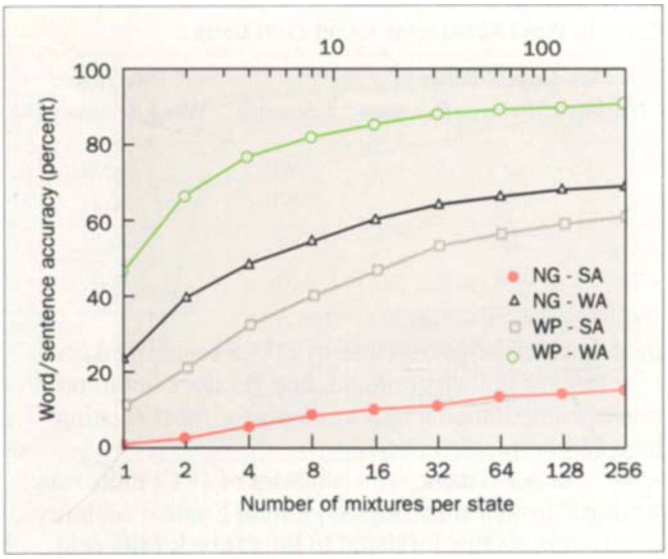

Figure 3. Average word and sentence accuracy rates versus the number of Gaussian mixture components, per state, using CI/TD units for the WP and NG grammars.

Experimental Results

The DARPA naval RM task is a data-base access and retrieval task based on information about battleships throughout the world. As previously mentioned, the task vocabulary contains 991 words; the language perplexity is about nine (that is, considerably less than that of the WP or the NG FSNs). There is a "training" set (referred to as SI-109) consisting of 3,990 read sentences from 109 speakers (30 to 40 sentences per speaker). Four different sets of test data were used, containing a total of 1,380 utterances from 34 new speakers, to evaluate the system's performance.

The recognizer was set up as a large FSN, with about 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[13,34] in which the best accumulated likelihood score, L^*, was determined at each frame. Based on the beam width delta (Δ), all nodes whose accumulated likelihood was less than ($L^* - \Delta$) were eliminated from a list of active nodes (that is, these paths were no longer followed). To prevent an excessive number of short-function-word insertions, a word-insertion penalty was used in the decoding algorithm at the end of each word arc. By adjusting the value of the word-insertion penalty, the rate of word insertion (and word deletion) could be controlled. Appropriate values of the word-insertion penalty were determined experimentally.

CI and TD Units. The set of 47 CI units was "trained" from the 3,990 "training" sentences of the SI-109 "training" set for the DARPA task. Therefore, these units are TD units. HMMs were built with continuous mixture densities with up to 256 Gaussian mixture components per state. Recognition tests were then performed, using all four independent test sets, with both the WP and NG grammars and no semantic processing. The results of these baseline tests, in terms of word and sentence accuracy, are summarized in Figure 3. For the WP syntax, the average word accuracy rate for one mixture (single Gaussian) per state is 47.5 percent; for 256 mixtures per state, the average word accuracy rate is 91.7 percent. For NG syntax, the average word accuracy rate for one mixture per state is 23.8 percent; for 256 mixtures per state, the average word accuracy rate is 69.3 percent. Choice of an appropriate model (the number of mixture components) depends on the application requirements, such as the recognition accuracy, the computation demand, and the system architecture. Although performance improves as the number of mixture components per state increases, a good compromise is the 32-mixture

Table II. Word Accuracies Using CI/TI Units.

Maximum Number of Gaussian Mixtures Per State	Grammar	Average Word Accuracy (%)
64	WP	82.5
256	WP	83.1
64	NG	54.7
256	NG	56.3

model, which achieves close to a 90-percent word accuracy rate for the WP grammar, and yet does not impose severe computational requirements for most existing hardware implementations.

CI and TI Units. The same set of 47 CI units was "trained" from a set of 10,000 general English sentences, which were totally unrelated to the RM task (different vocabulary, different syntax, and so on). These sentences were recorded at CMU and provided to AT&T by the CMU speech group. This set is referred to as the GE-10,000 "training" set ("GE" refers to "general English"). Tests were run using HMMs with up to 64 and 256 Gaussian mixture components per state. Test results are shown in Table II.

A comparison of the results in Table II with those in Figure 3 reveals that the word accuracy rate falls from about 92 percent, with TD units, to about 83 percent with TI units, for the WP case. It falls from about 69 percent with TD units, to 56 percent with TI units, for the NG case, when using models with a maximum of 256 Gaussian mixture components per state. As a result, there is a significant loss of performance—even for the case of 47 CI units. One reason for this performance loss is the word-to-context mismatch, as discussed previously.

Another reason is that it is considerably more difficult to define linguistic content of the GE-10,000 sentence "training" set, because of the sentence generality. For "training," a set of isolated word pronunciations for words in the 10,000-sentence "training" set was used. For many sentences, if not most, this formal pronunciation is grossly inadequate. This was not a problem for sentences from the RM task, however, as most pronunciations (especially with real speech) closely followed isolated word pronunciations in the lexicon. Attempts to modify the word pronunciations, based on rules appropriate for a speech synthesizer, were made. However, the modifications did not yield any performance improvement in the recognition tests.

CD and TD Units. Based on the DARPA RM "training" set, several sets of subword units that were context dependent were created. The criterion for including a CD unit was that it occurred sufficiently often in a "training" set that a reliable model could be designed. It was found, experimentally, that thresholds from 20 to 30 were required. CD-unit sets were designed, based on both intraword and interword units. In the case of using SI-109 "training" sentences, a threshold of 30 resulted in a set of 1,769 intraword and interword units. The word accuracy rates on the test data, obtained using a pair of models for a CD unit set—one for female speakers and one for male speakers—were 96.0 percent and 80.6 percent, using the WP and NG grammars, respectively.

Word accuracy improves significantly using CD/TD units, when units for modeling interword coarticulation are incorporated. One reason for this is that interword units are implicitly providing extra grammatical information, effectively lowering the perplexity of the language model, thereby improving performance. Gender-dependent models of subword units also provide significant improvement in word accuracy, because the models are designed to cope with the acoustic difference that is inherent in the speech signal between female and male speakers.

CD/TI Units. Using the GE-10,000 sentence "training" set, several sets of subword units that were context dependent were created, 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 do not occur even once in the RM task vocabulary. Whenever "training" data is used to create a CD/TI unit that is not used in the RM task, the size of the relevant "training" set is essentially reduced. To partially alleviate this problem, it is possible to post-process the set of CD/TI units, to remove all such units that do not occur in the RM task, and to reassign those units to the equivalent CI/TI unit.[35] In this manner, all of the "training" data is used to create the CD/TI units.

Thresholds of 75 and 100 occurrences of each unit were found to provide the best recognition performance for this task. Using a set of 1,418 units, average word accuracy rates for all testing data were 87.3 percent for the WP grammar and 61.9 percent for the NG grammar. The improvement in performance is about 4.2

percent (25-percent reduction in error rate) for the WP case, and 5.3 percent (12-percent reduction in error rate) for the NG case, as compared to results using the CI/TI units in Table II.

Speaker Adaptation (SA). Perhaps the ultimate way to create ("train") subword units is to adapt them to the task, to the speaking environment, 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 generating the highest performance scores.

One way to accomplish speaker adaptation is through Bayesian learning, in which an initial set of seed models (for example, SI models) is combined with speaker-specific adaptation data to adjust the model parameters. The result is that the set of subword models matches well with the acoustic properties of the adaptation data. This learning scheme has been proposed.[36,37] 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 speakers, with an independent test set of 25 sentences by each of the same speakers. Five initial sets of models were used, and the sets were "trained" from the SI-109 "training" set, which included:

1. A single set of CI/TD models with 47 subword units,
2. A pair of gender-dependent CI/TD models, each with 47 subword units,
3. A single set of CD/TD models, with 1,769 subword units,
4. A pair of gender-dependent CD/TD models, each with 1,769 subword units, and
5. A single set of CI/TD models, "trained" entirely on speaker-dependent sentences.

For each of these models, adaptation (or initial learning in the case of the fifth model) was performed using 40 and 80 sentences (models 4 and 5), 100 sentences (models 1 through 3), 150 sentences (models 4 and 5), 300 sentences (models 4 and 5), and 600 sentences. For each adapted model, word accuracy rates on the independent test set were measured. The results are shown in Figure 4. For models 1, 2, and 5, where 47 CI/TD units were used, the adapted models all converged to a 96.5-percent word accuracy rate when all 600 "training" sentences were used in the adaptation. Model 5, which did not use an initial, well-trained model, converged at the fastest rate, and had a word accuracy rate significantly lower than models 1 and 2 until all 600 "training" sentences were used. Differences in word accuracy rates, resulting from single models and male/female models, were small, but not insignificant for short adaptation sets.

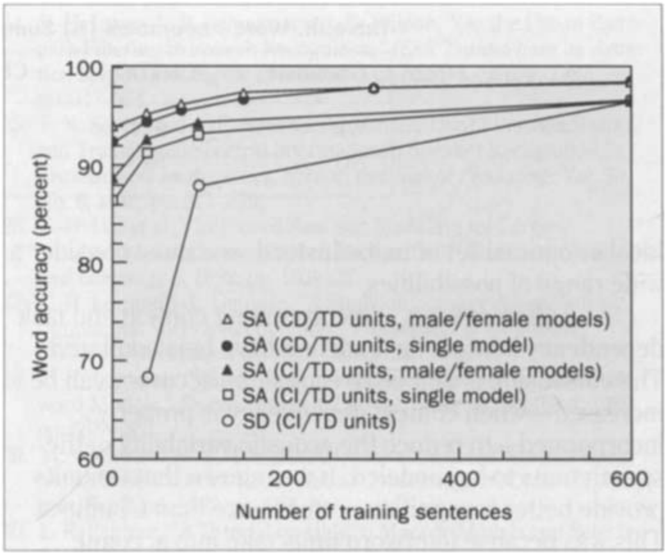

Figure 4. Word accuracy rates versus number of "training" sentences, based on adaptive "training" for different initial subword-unit models.

Word error rates for models 3 and 4, in which CD/TD units were used, were significantly lower than those obtained using only CI/TD units. Again, both of the models converged to a 98.5-percent word accuracy rate when all 600 "training" sentences were used for adaptation.

Summary

This paper describes a system for large-vocabulary, continuous speech recognition, and discusses the key issues in design and implementation of the system. Research has shown that the choice and method of "training" of basic subword units are critical, and that a wide range of options exists. Results were presented comparing the use of CI and CD units, based on their frequencies of occurrence in the "training" data. Two methods of parameter estimation were discussed—namely, the standard maximum likelihood and adaptive training methods. Each choice of subword units, in combination with each method of parameter estimation, has distinct advantages and disadvantages, and there is no

Table III. Word Accuracies (%) Summary.

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

ideal or optimal set of units. Instead, one must consider a wide range of possibilities.

Different ways of incorporating context and task dependency in acoustic modeling have been explored. The conclusion is that task-recognition accuracy can be increased—when context dependency is properly incorporated—to reduce the acoustic variability of the speech units to be modeled. It was shown that CD units provide better recognition performance than CI units. This was because interword units take into account cross-word co-articulation, and, therefore, provide more accurate modeling of speech units than intraword units in fluently spoken, continuous speech.

Similarly, acoustic variability of speech units can be reduced further when gender dependency is considered in the design of the acoustic models for the set of speech units. Gender-dependent models usually give better performance than gender-independent models, but at a slightly higher computational cost. It was also shown that, for a given task, speech-unit models—based on TD "training" data—always outperform models "trained" with TI "training" data. A comparison of performance for speaker-independent recognition of the DARPA RM task is shown in Table III. The word accuracy rates are based on testing 1,380 utterances from 34 new speakers not contained in the "training" set.

Most of the results presented in this paper were obtained with the RM task using the WP and the NG covering grammars. Experiments with the lower-perplexity, full grammar were conducted by performing speech recognition, first with a covering grammar, then using a semantic post-processor[33], to correct obvious word errors. By incorporating a simple set of semantic and syntactic rules for the RM task in this two-pass recognition, a word accuracy rate of over 99 percent, and a 92-percent string accuracy rate on a random subset of 300 testing utterances, were achieved. In addition to the RM task, many other subword-based studies have been carried out. For example, the same subword-based approach has been applied to the problem of connected-digit recognition with very high performance on the TI connected-digit data base.[38] The ATIS speech-understanding task was also implemented—with a word accuracy rate of close to 90 percent[39]—for recognition of spontaneously spoken utterances. Subword models have been used to derive a spoken lexicon for large-vocabulary, isolated word recognition.[28,40] Finally, the same subword-based approach has also been applied to speaker verification, to enhance the flexibility of verification systems.[41]

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 queries (speech input), with essentially unlimited vocabulary and syntax. This type of system suggests implementation of an entirely different process with a completely new set of problems associated with unknown words, non-grammatical constructions, extraneous speech, and so on. In addition, the traditional problems associated with noisy environments, speaker variability, transmission system variability, and other issues remain—along with the need to improve the acoustic front-end signal processing, and to provide efficient search strategies for large task networks.

References

1. P. J. Price et al., "A Database for Continuous Speech Recognition in a 1,000-Word Domain," *Proceedings of ICASSP-88*, New York City, 1988, pp. 651–654.
2. C. T. Hemphill, J. J. Godgrey, and G. D. Doddington, "The ATIS Spoken-Language-System Pilot Corpus," *Proceedings of the DARPA Speech and Natural-Language Workshop*, Hidden Valley, Pennsylvania, 1990, pp. 96–101.
3. L. Hirshman and MADCOW Group, "Multi-Site Data Collection for a Spoken-Language Corpus," *ICSLP-92*, Banff, Alberta, Canada, 1992, pp. 903–906.
4. D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *ICSLP-92*, Banff, Alberta, Canada, 1992, pp. 899–902.

5. F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer," *Proceedings of IEEE 73*, IEEE Publishing Company, 1985, pp. 1616–1624.
6. D. B. Roe et al., "Efficient Grammar Processing for a Spoken-Language Translation System," *Proceedings of ICASSP-92*, San Francisco, California, 1992, pp. 213–216.
7. T. Morimoto et al., "Spoken Language: Toward Realizing An Automatic Telephone Interpretation," *Proceedings of INFO Japan 90*, 1990, pp. 553-559.
8. S. Sagayama et al., "ATREUS: Continuous Speech-Recognition Systems at ATR Interpreting Telephony Research Laboratories," *Proceedings of SST-92*, Brisbane, Australia, 1992, pp. 324–329.
9. H. Ney and A. Paeseler, "Phoneme-Based Continuous Speech-Recognition Results for Different Language Models in a 1,000-Word SPICOS System," *Speech Communication*, Vol. 7, No. 4, 1988, pp. 367–374.
10. F. Fissore et al., "Lexical Access to Very Large Vocabulary," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-37, No. 8, 1989, pp. 1197–1213.
11. L. F. Lamel and J.-L. Gauvain, "Continuous Speech Recognition at LIMSI," *Proceedings of the DARPA CSR/MTO Workshop*, 1992, Palo Alto, California.
12. L. Deng et al., *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-37, No. 8, 1990, pp. 1197–1213.
13. C.-H. Lee et al., "Acoustic Modeling for Large-Vocabulary Speech Recognition," *Computer Speech and Language 4*, 1990, pp. 127–165.
14. A. Ljolje and M. D. Riley, "Optimal Speech Recognition Using Phone Recognition and Lexical Access," *Proceedings of ICSLP-92*, Banff, Alberta, Canada, 1992, pp. 313–316.
15. R. Schwartz et al., "The BBN BYBLOS Continuous Speech-Recognition System," *Proceedings of the DARPA Speech and Natural-Language Workshop*, Philadelphia, Pennsylvania, 1989, pp. 94–99.
16. K.-F. Lee, *Automatic Speech Recognition—The Development of the SPHINX System*, Kluwer Academic Publishers, Boston, Massachusetts, 1989.
17. J. Baker et al., "Large-Vocabulary Recognition of *Wall Street Journal* Sentences at Dragon System," *Proceedings of the DARPA Speech and Natural-Language Workshop*, Harriman, New York, 1992, pp. 387–392.
18. D. B. Paul, "The Lincoln Robust Continuous Speech Recognizer," *Proceedings of ICASSP-89*, Glasgow, Scotland, 1989, pp. 449–452.
19. V. Zue et al., "The MIT Summit Speech-Recognition System: A Progress Report," *Proceedings of the DARPA Speech and Natural-Language Workshop*, Philadelphia, Pennsylvania, 1989, pp. 179–189.
20. H. Murveit et al., "SRI's DECIPHER System," *Proceedings of the DARPA Speech and Natural-Language Workshop*, Philadelphia, Pennsylvania, 1989, pp. 238–242.
21. A. Averbuch et al., "Experiments with the Tangora 20,000-Word Speech Recognizer," *Proceedings of ICASSP-87*, Dallas, Texas, 1987, pp. 701–704.
22. S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-28, No. 4, 1980, pp. 357–366.
23. S. Furui, "Speaker-Independent Isolated-Word Recognition Using Dynamic Features of the Speech Spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34, No. 1, 1986, pp. 52–59.
24. B.-H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the Use of Bandpass Filtering in Speech Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-35, No. 7, 1987, pp. 947–954.
25. F. K. Soong and A. E. Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," *Transactions on Acoustics, Speech, and Signal Processing*, Vol. 36, No. 6, 1988, pp. 871–879.
26. C.-H. Lee et al., "Improved Acoustic Modeling for Large-Vocabulary, Continuous Speech Recognition," *Computer Speech and Language 6*, 1992, pp. 103–127.
27. C.-H. Lee and J.-L. Gauvain, "A Study on Speaker Adaptation for Continuous Speech Recognition," *Proceedings of the DARPA CSR/MTO Workshop*, Palo Alto, California, 1992.
28. C.-H. Lee et al., "Word Recognition Using Whole-Word and Subword Models," *Proceedings of ICASSP-89*, Glasgow, Scotland, 1989, pp. 683–686.
29. M. Weintraub et al., "Linguistic Constraints in Hidden-Markov-Model-Based Speech Recognition," *Proceedings of ICASSP-89*, Glasgow, Scotland, 1989, pp. 699–702.
30. L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of IEEE 77*, 1989, pp. 257–286.
31. X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, Edinburgh, Scotland, 1990.
32. J. R. Bellegarda and D. Nahamoo, "Tied-Mixture, Continuous-Parameter Modeling for Speech Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-38, No. 12, 1990, pp. 2033–2045.
33. R. Pieraccini and C.-H. Lee, "Factorization of Language Constraints in Speech Recognition," *Proceedings of American Computational Linguistics 91*, Berkeley, California, 1991.
34. B. Lowerre and D. R. Reddy, "The HARPY Speech-Understanding System" in W. Lea (ed.), *Trends in Speech Recognition*, Prentice-Hall Inc., 1980, pp. 340–346.
35. H.-W. Hon, "Vocabulary-Independent Speech Recognition: The VOCIND System," Doctoral Thesis, School of Computer Science, Carnegie Mellon University, 1992.
36. J.-L. Gauvain and C.-H. Lee, "MAP Estimation of Continuous-Density HMM: Theory and Applications," *Proceedings of the DARPA Speech and Natural-Language Workshop*, Harriman, New York, 1992, pp. 185–190.
37. J.-L. Gauvain and C.-H. Lee, "Bayesian Learning for Hidden Markov Models With Gaussian Mixture State Observation Densities," *Speech Communication*, Vol. 11, Nos. 2-3, 1992, pp. 205–214.
38. J.-L. Gauvain and C.-H. Lee, "Improved Acoustic Modeling with Bayesian Learning," *Proceedings of ICASSP-92*, San Francisco, California, 1992, pp. 481–484.
39. R. Pieraccini et al., "A Speech-Understanding System Based on Statistical Representation of Semantics," *Proceedings of ICASSP-92*, San Francisco, California, 1992, pp. 193–196.
40. C.-H. Lee, B.-H. Juang, and F. K. Soong, "A Segment-Model-Based Approach to Speech Recognition," *Proceedings of ICASSP-88*, New York City, 1988, pp. 501-504.
41. A. E. Rosenberg et al., "Experiments in Automatic Talker Verification Using Subword-Unit Hidden Markov Models," *Proceedings of ICSLP-90*, Kobe, Japan, 1990.

(Manuscript approved October 1993)

Chin-Hui Lee *is a member of the technical staff in the speech research department at AT&T Bell Laboratories in Murray Hill, New Jersey. He is involved in research on speech and speaker recognition and speech and signal modeling. Mr. Lee has a B.S. from National Taiwan University in Taipei, an M.S. from Yale University, New Haven, Connecticut, and a Ph.D. from the University of Washington in Seattle, all in electrical engineering. He joined AT&T in 1986.*

Jean-Luc Gauvain, *a visiting researcher at AT&T Bell Laboratories in Murray Hill, New Jersey, from June 1990 through November 1991, was a member of the speech research department. He was responsible for identifying and developing large-vocabulary speech-recognition techniques. Mr. Gauvain has an M.S. in telecommunications and a Ph.D. in electronics from the University of Paris in France.*

Roberto Pieraccini *is a member of the technical staff in the speech research department at AT&T Bell Laboratories, Murray Hill, New Jersey. He focuses his efforts on speech recognition and understanding, as well as language modeling. Mr. Pieraccini received a Ph.D. in electrical engineering from the University of Pisa in Italy. He joined AT&T in 1990.*

Lawrence R. Rabiner *is director of the information principles research laboratory at AT&T Bell Laboratories in Murray Hill, New Jersey. He leads research efforts in several key areas of information sciences: speech and image processing, interactive systems (including handwriting recognition), digital signal processing, and communications. Mr. Rabiner received B.S., M.S., and Ph.D. degrees in electrical engineering, all from the Massachusetts Institute of Technology in Cambridge. He joined AT&T in 1962.*