PAPER *Special issue on spoken language systems and ...*

Large Vocabulary Continuous Speech Recognition: from Laboratory Systems towards Real-World Applications

Jean-Luc GAUVAIN† *and* **Lori LAMEL**†, *Nonmembers*

SUMMARY

This paper provides an overview of the state-of-the-art in laboratory speaker-independent, large vocabulary continuous speech recognition (LVCSR) systems with a view towards adapting such technology to the requirements of real-world applications. While in speech recognition the principal concern is to transcribe the speech signal as a sequence of words, the same core technology can be applied to domains other than dictation. The main topics addressed are acoustic-phonetic modeling, lexical representation, language modeling, decoding and model adaptation. After a brief summary of experimental results some directions towards usable systems are given. In moving from laboratory systems towards real-world applications, different constraints arise which influence the system design. The application imposes limitations on computational resources, constraints on signal capture, requirements for noise and channel compensation, and rejection capability. The difficulties and costs of adapting existing technology to new languages and application need to be assessed. Near term applications for LVCSR technology are likely to grow in somewhat limited domains such as spoken language systems for information retrieval, and limited domain dictation. Perspectives on some unresolved problems are given, indicating areas for future research.

key words: Speech recognition, spoken language systems, dictation, large vocabulary, speaker-independent continuous speech recognition, acoustic modeling, model adaptation, multilingual

1. Introduction

In the past few years large vocabulary, continuous speech recognition (LVCSR) has been one of the focal areas of research in speech recognition, serving as a test bed to evaluate models and algorithms. This technology push has been fostered by U.S. ARPA efforts in providing common corpora and organizing annual benchmark tests to assess progress. The interest in LVCSR is larger than simply building dictation systems for general English, it also serves to develop core technology that can be used in less demanding applications such as voice-interactive database access or limited-domain dictation. Progress in speech recognition can also boost other spoken language technologies such as speaker and language identification which rely on the same modeling techniques.

Speech recognition is principally concerned with the problem of transcribing the speech signal as a sequence of words. Today's best performing systems use statistical models of speech generation. From this point of view, message generation is represented by a language model which provides estimates of $\Pr(w)$ for all word strings w, and the acoustic channel encoding the message w in the signal x is represented by a probability density function $f(x|w)$. The speech decoding problem then consists of maximizing the *a posteriori* probability of w, or equivalently, maximizing the product $\Pr(w)f(x|w)$.

The principles on which these systems are based have been known for many years now, and include the application of information theory to speech recognition[5],[46], the use of a spectral representation of the speech signal [20],[21], the use of dynamic programming for decoding[92],[93], and the use of context-dependent acoustic models[15],[56],[86]. Despite the fact that some of these techniques were proposed well over a decade ago, considerable progress has been made in recent year making speaker-independent, continuous speech dictation feasible for relatively large vocabularies of up to 65,000 words. This progress has been substantially aided by the availability of large speech and text corpora and by significant advances made in microelectronics which have facilitated the implementation of more complex models and algorithms.

The same modeling techniques have been adapted to other related applications, such as speech understanding or spoken language systems or in the identification of what we refer to as "non-linguistic" speech features[54]. These feature-specific models may also be directly used to more accurately model the speech signal thus in consequence improving the performance of the speech recognizers.

In this paper we review the state-of-the-art in laboratory systems for LVCSR and give some example directions taken towards real-world applications. Most state-of-the-art LVCSR systems make use of hidden Markov models (HMM) for acoustics modeling [9],[19],[22],[32],[43],[59],[60],[64],[65],[76],[80],[83],[94]. Other approaches include segment based models [68],[100] and neural networks [42] to estimate acoustic observation likelihoods. However except for the acoustic likelihood estimation, all systems make use of the HMM framework to combine linguistic and acoustic information in a single network representing all possible sentences. Decoding is the search for the most likely word string, which is in most cases approximated

Manuscript received June 1, 1996.
Manuscript revised June 1, 1996.

†The authors are researchers at LIMSI, CNRS ...

by the Viterbi algorithm.

Moving towards real-world applications means building usable systems which involves reconsidering many design issues such as signal capture, and noise and channel compensation, while taking into account limitations in computational resources. For many applications rejection capabilities will be essential. The difficulties and costs of adapting existing technology to new languages or new applications must also be evaluated. Although not the direct topic of this article, it should not be forgotten that many applications of speech technology will require not the transcription of speech into text, but the understanding of speech in order to carry out appropriate actions. A related consideration for the application is whether it is provided as a central service or as a stand-alone system for an individual user, which has implications for the design.

While we attempt to take a general view on some of the outstanding problems in speech recognition and current approaches towards resolving them, for examples we refer mostly to our own work.

2. Acoustic-Phonetic Modeling

For HMM based systems, acoustic modeling consists of modeling the probability density function of a sequence of acoustic feature vectors. The acoustic features are chosen so as to reduce model complexity while trying to keep the relevant information (i.e. the linguistic information for the speech recognition problem). Most recognition systems use short-time cepstral features based either on a Fourier transform or a linear prediction model. For wide band analysis (usually 8kHz or 10kHz) the two most popular sets of features are cepstrum coefficients obtained with an MFCC [17] analysis or with a PLP [40] analysis. In both cases a Mel scale short term power spectrum is estimated on a fixed window (usually in the range of 20 to 30ms), with the most commonly used frame rate being 10ms. To get the MFCC cepstrum coefficients a cosine transform is applied to the log power spectrum, whereas a root-LPCC analysis is used to obtain the PLP cepstrum coefficients. Both set of features have been used with success for LVCSR, but PLP analysis has been found for some systems to be more robust in presence of background noise [49],[95]. Our experiance has been that the implementation details are not very important, but optimal tuning, which may be dependent on the language or the channel conditions, can result in slight performance improvements.

The front end configuration used in the LIMSI system for English has been optimized on the ARPA WSJ corpus. We use a 30ms frame window and 26 cosine filters on a Mel scale over the 8kHz bandwidth, from which 15 cepstrum coefficients and normalized energy are derived. As in most LVCSR systems, sentence-based cepstral mean removal [25] is performed to normalize the cepstrum features, rendering them more robust to channel variability. Cepstral mean removal is a very simple, yet efficient technique to deal with channel changes. It has also been found to slightly improve recognition performance with clean speech in matched channel conditions. First and second order time derivative features are used to partially overcome the HMM limitations in modeling the temporal dynamics. The three sets of features are then combined in a single stream and modeled by continuous density HMM (CDHMM).†

Most LVCSR systems use acoustic units corresponding to phonemic or phonetic units (or phones in context). However it is certainly possible to perform speech recognition without use of a phonemic lexicon, either by use of "word models" (as was the more commonly used approach 10 years ago) or a different mapping such as the fenonic lexicon [8]. Compared to word models, subword units reduce the number of parameters, enable cross word modeling and facilitate porting to new vocabularies. Fenones offer the additional advantage of automatic training, but lack the ability to include *a priori* linguistic models. Context-dependent (CD) phone models are today the most commonly used acoustic units for LVCSR. Compared to larger units such as *diphones, demisyllables* or *syllables*, a large spectrum of contextual dependencies can be implemented for CD phone models associated with back-off mechanisms to model infrequent contexts. Various types of contexts have been investigated from a single phone context (right- or left-context), left and right-context (triphone), generalized triphones [56], position-dependent triphones (cross-word and within word triphones), function word triphones, and quinphones [94]. The optimal set of modeled contexts is usually the result of a tradeoff between resolution and robustness, and is highly dependent on the available training data. This optimization is generally done by minimizing the recognizer error rate on development data. In fact, more than the number of CD phone models, what is really important is to match the total number of model parameters to the amount of available training data. A powerful technique to keep the models trainable without sacrificing model resolution is to take advantage of the state similarity among different models of a given phone by tying the HMM state distributions. This basic idea is used in most current LVCSR systems although there are slight differences in the implementation and in the naming of the resulting clustered states (*senones* [44], *genones* [18], *PELs* [10], *tied-states* [98]). Numerous ways of tying HMM parameters have been investigated [91],[96] in order to overcome the sparse

†In some systems based on discrete or tied-mixture distributions, the three streams are modeled separately by assuming independance of the feature sets which allows the use of smaller codebooks. For CDHMMs we found that a single stream outperforms the multiple stream approach.

training data problem and to reduce the need for distribution smoothing techniques. When HMM state tying is based on a phonetic decision tree it has the additional advantage of providing a means to build models for unseen contexts (i.e. those contexts which do not occur in the training data) [45], [97].

The LIMSI recognizer, which has state-of-art performance [73], makes use of continuous density HMM with Gaussian mixture for acoustic modeling. The main advantage continuous density modeling offers over discrete or semi-continuous (or tied-mixture) observation density modeling is that the number of parameters used to model an HMM observation distribution can easily be adapted to the amount of available training data associated to this state. As a consequence, high precision modeling can be achieved for highly frequented states without the explicit need of smoothing techniques for the densities of less frequented states. Discrete and semi-continuous modeling use a fixed number of parameters to represent a given observation density and therefore cannot achieve high precision without the use of smoothing techniques or tying techniques mentioned above.

The acoustic models are sets of context-dependent, position-independent phone models, which include both intra-word and cross-word contexts. The contexts are automatically selected based on their frequencies in the training data. The models include triphone models, right- and left-context phone models, and context-independent phone models. Each phone model is a three state left-to-right CDHMM with Gaussian mixture observation densities (typically 32 components). The covariance matrices of all the Gaussians are diagonal. Separate male and female models are used to more accurately model the speech data and state-tying is used to increase the triphone coverage. These models are obtained from speaker-independent seed models using Maximum *A Posteriori* estimators [30].

During system development for LVCSR, phone recognition experiments are useful to evaluate different acoustic model sets. It has been shown that improvements in phone accuracy are directly indicative of improvements in word accuracy when the same phone models are used for recognition [53]. Phone recognition provides the added benefit that the recognized phone string can be used to understand word recognition errors and problems in the lexical representation.

3. Lexical Representation

Lexical modeling provides the link between the lexical entries (usually words) used by the language model and the acoustic models, with each lexical entry being described as a sequence of elementary units. Experience with LVCSR has shown that systematic lexical design can improve system performance [50]. Lexical design entails two main parts - selection of the vocabulary items and representation of the pronunciation entry using the basic units of the recognition system. Vocabulary selection to maximize lexical coverage for a given size lexicon has been previously reported [13], [33]. On average, each out-of-vocabulary (OOV) word causes more than a single error, with rates of 1.6 to 2.0 additional errors reported. An obvious way to reduce the error rate due to OOVs is to increase the size of the lexicon. Increasing the lexicon size up to 65k words has been shown to improve performance, despite the potential of increased confusability of the lexical entries. In the LIMSI system, going from 20k words to 65k words, recovers on average 1.2 times as many errors as OOV words removed [33].

For LVCSR, the lexical unit of choice is usually phonemes or phoneme-like units, specific for the language (We use 46 for American English, 45 for British English, 35 for French, 49 for German, and 26 for Spanish.). In generating pronunciation baseforms, most lexicons include standard pronunciations and do not explicitly represent allophones. This representation is chosen as most allophonic variants can be predicted by rules, and their use is optional. More importantly, there often is a continuum between different allophones of a given phoneme and the decision as to which occured in any given utterance is subjective. By using a phonemic representation, no hard decision is imposed, and it is left to the acoustic models to represent the observed variants in the training data. Several efforts to automatically learn and generate word pronunciations have been investigated [14], [16], [79], [90].

However, there are a variety of words for which frequent alternative pronunciation variants are observed, and these variants are not due to allophonic differences. One common example is the suffix *-ization* which can be pronounced with a diphthong (/ɑʸ/) or a schwa (/ə/). Another example is the palatalization of the /k/ in a /u/ context resulting from the insertion of a /y/, such as in the word *coupon* (pronounced /kupɑn/ or /kyupɑn/). Alternate pronunciations are also needed for homographs (words spelled the same, but pronounced differently) which reflect different parts of speech (verb or noun) such as *excuse, record, produce.*

Fast speakers tend to poorly articulate unstressed syllables (and sometimes skip them completely), particularly in long words with sequences of unstressed syllables. Although such long words are typically well recognized, often a nearby function word is deleted. To reduce these kinds of errors, alternate pronunciations for long words such as *Minneapolis* (/mɪniæpəlɪs/ or /mɪniæplɪs/) and *positioning* (/pəzɪʃəniŋ/ or /pəzɪʃniŋ/), can be included in the lexicon allowing schwa-deletion or syllabic consonants in unstressed syllables. Alternative pronunciations can also be provided for common 3 syllable words such as *interest* (/ɪntrɪst/, /ɪntɚɪst/ or /ɪnɚɪst/, where the [n] in the latter example is often realized as a nasal flap ɾ̃

and *company* (/kʌmpəni/ or /kʌmpni/) which are often pronounced with only 2 syllables.

Phonological rules have been proposed to account for some of the phonological variations observed in fluent speech [67]. The principle behind the phonological rules is to modify the phone network to take into account such variations [16], [36], [52]. These rules are optionally applied during training and recognition. Using phonological rules during training results in better acoustic models, as they are less "polluted" by wrong transcriptions. Their use during recognition reduces the number of mismatches. The same mechanism has been used to handle liaisons, mute-e, and final consonant cluster reduction for French [31].

4. Language Modeling

Language models are used to model regularities in natural language, and can therefore be used in speech recognition to limit the decoding search space. The most popular methods, such as statistical n-gram models, attempt to capture the syntactic and semantic constraints by estimating the frequencies of sequences of n words. A backoff mechanism [48] is generally used to smooth the estimates of the probabilities of rare n-grams by relying on a lower order n-gram when there is insufficient training data, and to provide a means of modeling unobserved n-grams. Another advantage of the backoff mechanism is that LM size can be arbitrarily reduced by relying more on the backoff component, obtained by simply increasing the minimum number of required n-gram observations needed to include the n-gram. This property can also be used to reduce computational requirements. While bigram and trigram LMs are most widely used, small improvements have been reported with the use of longer span *4*-grams [9], [59], [94] and *5*-grams[41] or class *5*-grams[84]. Language models are typically compared by measuring the perplexity of a set of development texts.

Given a large text corpus it may seem relatively straightforward to construct n-gram language models. Most of the steps are pretty standard and make use of tools that count word and word sequence occurences. The main differences arise in the choice of the vocabulary and in the definition of words, such as the treatment of compound words or acronymns. There is, however, a significant amount of effort needed to preprocess the texts before they can be used. First, the texts must be put in a standardized format. For example, for the ARPA NAB text corpus [2], this formating work has been carried out by LDC using modified versions of text processing tools provided from Lincoln Labs [74]. The main conditioning steps are text markup and conversion for LVCSR. Text markup consists of tagging the texts (article, paragraph and sentence markers) and garbage bracketing.[†] Then numerical expressions are expanded, and isolated letters marked, and finally the text is transformed to upper case. At LIMSI similiar processing has been carried out on over 90M words of newspaper texts from *Le Monde*. Further semi-automatic processing is necessary to correct frequent errors inherent in the texts or arising from processing with the distributed text processing tools. The error correction consists primarily of correcting obvious mispellings (such as MILLLION, OFFICALS, LITTLEKNOWN), systematic bugs introduced by text processing tools, and expanding abbreviations and acronyms in a consistent manner. Better language models can be obtained using texts transformed to be closer to the observed reading style, where the transformation rules and corresponding probabilities are automatically derived by aligning prompt texts with the transcriptions of the acoustic data. For example, the word HUNDRED followed by a number is replaced by HUNDRED AND 50% of the time. Similarly, half the occurences of ONE EIGHTH are replaced by AN EIGHTH, and 15% of MILLION DOLLARS are replaced with simply MILLION. After treating the texts, a reduced perplexity of 5 points on development data was reported [33], along with a better coverage of the 65k lexicon.

	Lexicon			
Test set	*Baseline 20k*	*20k*	*40k*	*65k*
Dev94	2.7	2.2	0.8	0.4
Eval94	2.5	2.0	0.8	0.4

Table 1 OOV rate (%) on development and test sentences for 20k, 40k, and 65k lexicons. The baseline 20k vocabulary contains the most common 20k words in the training texts (processed texts distributed by LDC).

A common way of selecting a recognition vocabulary is to measure the OOV rate on development data. For the 1994 ARPA NAB task, it was found that the best lexical coverage was obtained by selecting the vocabulary on a subset of the training data (the most recent 2 years), as opposed to using all the available data [13], [33]. This is to be expected as the development test data were selected from a time period following the training text material, and the vocabulary coverage reflects recency effects.

The lexical coverages of several LIMSI lexicons in Table 1 reflect the combined effect of text cleaning and vocabulary selection. The OOV rate with the 20k wordlist is significantly smaller than that of the baseline 20k wordlist. The OOV rate with the 65k word list on the 1994 development data (Dev94) is 0.39% which is a pretty accurate indicator of the 0.42% observed on the evaluation data (Eval94).

[†]Garbage includes not only corrupted text materials, but all text material unsuitable for sentence-based language modeling, such as tables and lists.

5. Decoding

One of the most important problems in implementing the decoder of a large vocabulary speech recognizer is the design of an efficient search algorithm to deal with the huge search space, especially when using language models with a longer span than two successive words, such as *3*-grams and *4*-grams. Even for research purposes where real-time recognition is not needed there is a limit on computing resources (memory and CPU time) above which the development process becomes too costly.

The most commonly used approach for small and medium vocabulary sizes is the one-pass frame-synchronous Viterbi beam search [62] which uses a dynamic programming procedure. This basic strategy has been extended to deal with large vocabularies by adding features such as fast match [7],[37], word-dependent phonetic trees [63], forward-backward search [4], N-best rescoring [85], progressive search[29],[61] and one-pass dynamic network decoding [66]. An alternative to the frame-synchronous Viterbi beam search is an asynchronous search based on the A* algorithm such as *stack decoding* [6],[75] or the *envelope search* [39].

Single pass decoders such the stack decoder [75] or the one-pass dynamic network decoder [66] which use all the knowledge sources (e.g. cross word triphones and trigram language models) in one step are certainly very attractive to minimize search errors. However, many LVCSR systems under development use multiple pass decoders to reduce the computational resources needed for evaluation runs. In this case, information is transmitted between passes by means of word lattices, word graphs or N-best lists. (Lattices are graphs where nodes correspond to particular frames and where arcs representing word hypothesis have associated acoustic and language model scores.)

The two-step approach used in the LIMSI research system is based on the idea of progressive search where the information between levels is transmitted via word graphs [29]. Due to memory constraints, each step may consist of one or more passes, each using successively more refined models. All decoding passes use cross-word CD triphone models.

The first step of the decoder uses a bigram-backoff LM with a tree organization of the lexicon for the backoff component. This one-pass frame-synchronous beam search, which includes intra- and inter-word CD phone models, and gender-dependent models, generates a list of word hypotheses resulting in a word lattice. The tree representation of the backoff component (first introduced in our Nov92 CSR system) provides an efficient way of arbitrarily reducing the search space and of limiting the computational requirements of the first pass which represent on the order of 75% of the computation need for the entire decoding process. Additionally, this strategy allows us to use a static graph instead of building it dynamically, therefore avoiding the computational bookkeeping costs associated with dynamic network decoding. The key elements of the procedure used to generate the word graph from the word lattice are the following.† First, a word graph is generated from the lattice by merging three consecutive frames (i.e. the minimum duration for a word in our system). Then, "similar" graph nodes are merged with the goal of reducing the overall graph size and generalizing the word lattice. This step is reiterated until no further reductions are possible. Finally, based on the trigram backoff language model, a trigram word graph is generated by duplicating the nodes having multiple language model contexts. Bigram backoff nodes are created when possible to limit the graph expansion. The trigram step may be carried out in more that one pass, using successively larger language models.

Evidently, the first pass used to generate the initial word lattice must be accurate enough to not introduce lattice errors which are unrecoverable with further processing. In our 65k system the graph error is usually small ($\sim$ 2%), but poor speakers tend to have higher graph errors, and higher graph errors are obtained on telephone and noisy data.

6. Model Adaptation

Model adaptation can be used to reduce the mistmatch between test and training conditions or to improve model accuracy based on the observed test data. Adaptation can be of the acoustic models or the langugae models, or even to the pronunciation lexicon. One of the main challenges in LVCSR is building robust systems that keep high recognition accuracy when testing and training environmental conditions are different. Two classes of techniques to increase system robustness can be identified: signal processing techniques which attempt to compensate for the mismatch between testing and training by correcting the speech signal to be decoded; and model adaptation techniques which attempt to modify the model parameters to better represent the observed signal. Signal processing based approaches include normalization techniques that remove variability, thereby increasing the system accuracy under mismatched conditions but often resulting in reduced word accuracy under matched conditions, and compensation techniques which rely on a mismatch model and/or speech models. Model adaptation is a much more powerful approach, especially when the signal processing relies on a speech model. Therefore when computational resources are not an issue, model adaptation is the prefered approach to compensate for mismatches.

Acoustic model adaptation can be used to compen-

†In our implementation, a word lattice differs from a word graph only because it includes word endpoint information.

sate mismatches of various natures due to new acoustic environments, to new transducers and channels, or to particular speaker characteristics, such as the voice of a non-native speaker. The most commonly used techniques for acoustic model adaptation are parallel model combination (PMC), maximum *a posteriori* (MAP) estimation, and transformation methods such as maximum likelihood linear regression (MLLR). PMC is only used to account for environmental mismatch due to additive noise whereas MAP estimation and MLLR are general tools that can be used for speaker adaptation and environmental mismatch.

PMC approximates a noise corrupted model by combining a clean speech model with a noise model [26]. For practical reasons, it is generally assumed that the noise density is Gaussian and that the noise corrupted speech model has the same structure and number of parameters as the clean speech model – typically a continuous density HMM with Gaussian mixture. Various techniques have been proposed to estimate the noisy speech models, including the log-normal approximation approach, the numerical integration approach, and the data driven approach[27]. The log-normal approximation is crude especially for the derivative parameters, and all three approaches require making some approximations to estimate derivative parameters other than first order differences.

MAP estimation can be used to incorporate prior knowledge into the CDHMM training process, where the prior information consists of prior densities of the HMM parameters [35]. In the case of speaker adaptation, MAP estimation may be viewed as a process for adjusting speaker-independent models to form speaker-specific ones based on the available prior information and a small amount of speaker-specific adaptation data. The joint prior density for the parameters in a state is usually assumed to be a product of Normal-Gamma densities for the mean and variance parameters of the Gaussian mixture components and a Dirichlet density for the mixture gain parameters. MAP estimation has the same asymptotic properties as ML estimation but when independent priors are used for different phone models the adaptation rate may be very slow, particularly for large models. It is therefore avantageous to represent correlations between model parameters in the form of joint prior distributions [88], [99].

MLLR is used to estimate a set of transformation matrices for the HMM Gaussian parameters in order to maximize the likelihood of the adaptation data [58]. This adaptation method was originally used for speaker adaptation, but it can equally be applied to environmental mismatch [95]. Since the number of transformation parameters is small, large models can be adapted with small amounts of data. To obtain ML asymptotic properties it is necessary to adjust the number of linear tranformations to the amount of available adaptation data. This can be done efficiently by arranging the mixture components into a tree and dynamically defining the regression classes [57]. It should be noted that both MAP estimation and MLLR adaptation can be used for supervised or unsupervised model adaptation.

Model adaptation can evidently also be applied to the language model. In most LVCSR systems one or more language models are used, but these LMs are usually static. Various approaches have been taken to adapt the language model based on the observed text so far, including the use of a *cache model* [47], [82], a *trigger model* [81], or *topic coherence modeling* [87]. The cache model is based on the idea that words appearing in a dictated document will have an increased probability of appearing again in the same document. For short documents the number of words appearing is small, and as a consequence the benefit is small. The trigger model attempts to overcome this by using observed words to increase the probabilities of other words that often co-occur with the trigger word. In topic coherence modeling, selected keywords in the processed text are used to retrieve articles on similar topics with which sublanguage models are constructed and used to rescore N-best hypotheses. Despite the growing interest in adaptive language models, thus far only minimal improvements have been obtained compared to the use of very large, static n-gram models.

7. Assessment Driven Technology Development

The most widely known evaluation experiments in speech recognition have been coordinated by NIST (National Institute for Science and Technology) and sponsored by the U.S. ARPA program. Through the objective evaluation of different recognition systems, the community has been able to contrast different methods, sharing reliable information among participants. The initial evaluations were carried out on the 1000-word Resource Management (RM) task [69] and on the 5000-word and 20,000 word Wall Street Journal (WSJ) task [70], [71], and most recently on the unlimited vocabulary North American Business News (NAB) task [72] with high quality read speech and in more challenging acoustic conditions with unknown microphones and background environmental noise (MUM) [73]. The baseline tests constrain the acoustic and language model training data, as well as fix the vocabulary and language model so as to permit cross-site comparisons in acoustic modeling for speech recognition. Non-baseline conditions relax these constraints, allowing the use of additional acoustic and language model training materials. The results of the last 5 baseline evaluations for speaker-independent LVCSR, held in September 1992 (RM), November 1992 (WSJ) and November 1993 (WSJ), November 1994 (NAB), and November 1995 (MUM) are given Table 2. The commonly used metric of "word error" rate is defined as: *%word error*

Test	*Test Conditions*	*Vocabulary*	*Word Error (%)*
Sep92 RM	1k wordpair, closed vocabulary	1k	4.4 - 11.7
Nov92 WSJ	5k bg, closed vocabulary	5k	6.9 - 15.0
	20k bg	20k	15.2 - 25.2
Nov93 WSJ	20k open tg	20k	11.7 - 19.0
	5k bg	5k	8.7 - 17.7
	5k tg	5k	4.9 - 9.2
	5k tg, local telephone	5k	12.8 - 25.5
Nov94 NAB	20k tg, unlimited	20k	10.5 - 22.8
	unlimited	20 - 65k	7.2 - 17.4
	unlimited, telephone	40 - 65k	22.5 - 24.6
Nov95 MUM	unlimited, noise, unknown mic.	65k	13.5 - 55.5
	unlimited, noise, Sennheiser	65k	6.6 - 20.2

Table 2 Results on ARPA sponsored evaluation tests from 1992 to 1995. The tests were carried out on increasingly more difficult tasks and conditions. The lowest and highest word errors are given for each test.

= %substitutions + %insertions + %deletions.

Despite the increasing task difficulties, the word error rates are seen to decrease over time. In 1992, the 5k baseline test was carried out in a closed-vocabulary condition, meaning that the commonly used vocabulary included all the words in the test data. In contrast, for the open-vocabulary condition the test data are selected without ensuring that all lexical items appear in the known recognition vocabulary. Since the 1994 evaluation, the test data have been selected without limitations on the vocabulary and the use of a common LM is no longer imposed. The first table entry for 1994 gives results with an imposed tg LM and in the second entry no constraints were imposed. An important observation of this benchmark test is that by increasing the size of the recognition vocabulary (up to 65k words), the errors introduced by OOV words are reduced despite the potential for increased acoustic confusability of the larger lexicon. In 1994 and 1995 some sites used longer span language models (*4*-gram and *5*-gram). The comparative tests with telephone speech in 1993 and 1994 had performance levels significantly worse (over twice the word error rate) than on the clean speech data. In 1993 the telephone data was collected with subjects at SRI, over local Palo Alto lines, while in 1994 data were collected remotely over long distance channels.

Several points should be made about the above results. First, it can be seen that typically for a closed-vocabulary test, word errors are quite low - as low as 4% with a 1000 word vocabulary and 5% with 5000 words. Second, increasing the vocabulary size does not harm recognition performance, given a sufficient language model. Third, while the table shows average word error rates, for the same system there can be a factor of 10 difference in the word error rates for the best and worst speakers. Finally, while the benchmark tests have been extended to conditions closer to those of possible applications (telephone, noise conditions, multi-microphone, spontaneous dictation), they still remain in the domain of laboratory systems, with significant advances needed for real-world usage.

8. Towards Multilinguality

Speech recognition in multiple languages is essential in Europe, where the national language(s) are closely linked to the national cultures and identities. Even in the United States, a "monolingual" country, there is such a large immigrant population that there is increasing interest in multilingual speech recognition. It is thus of interest to assess the applicability of commonly used speech recognition techniques for different languages, and the issues involved in porting a speech recognizer to a new language.

To build a recognizer in a new language, the first step is obtaining the necessary acoustic and language model training data, and a pronunciation lexicon. System parameters or components which are dependent on the language (such as the phone set, the need for pronunciation alternatives or phonological rules) evidently must be changed. Other language dependent factors are related to the acoustic confusability of the words in the language (such as homophone, monophone, and compound word rates) and the word coverage of a given size recognition vocabulary. These factors will influence the size of the recognition vocabulary and the choice of acoustic units (context-independent or context-dependent), as well as the choice of language model (bigram, trigram, class-*n*-grams).

Taking into account language specificities can improve recognition performance. For example, in German, glottalized segments are good indicators of morpheme boundaries, a characteristic which, when accounted for in the lexicon and the acoustic models, has led to better recognition [3]. While word-initial glottalization also occurs in other languages such as English, its occurrence is less systematic and therefore more difficult to model.

At the lexical level, a given size lexicon will have different coverage across languages. Highly inflected languages require a larger lexicon to adequately represent the language. For example, comparing the number of distinct words in newspaper text corpora for En-

glish, French, German and Italian, the German corpus contains over twice as many distinct words as French, which has more than Italian and English [51].† The larger number of distinct words stems mainly from the number and gender agreement in nouns, adjectives and past participles, and the high number of different verb forms. While in English there is only one form for the definite article *the*, in French there are 3 forms *le, la, les* (masculine singular, feminine singular, plural), and in German are found singular forms *der, die, das* (male, female, neuter) and the plural form *die*. Declension case distinction adds 3 additional forms *des, dem, den* to the nominative form *der*. As a consequence, to obtain a lexical coverage of 95%, an English lexicon need only contain 5000 words, compared to 20,000 for French and Italian, and 65,000 for German.

Homophone rates differ across languages. A comparative study of French and English showed that, given a perfect phonemic transcription, 23% of words in the *WSJ* training texts are ambiguous, whereas 75% of the words in the *Le Monde* training texts have an ambiguous phonemic transcription [31]. Another difficulty specific to French is that most of the phonemes are also words (we refer to these as "monophone" words), and often have several graphemic forms (the phoneme */ɛ/* can stand for *ai, aie, aies, ait, aient, hais, hait, haie, haies, es, est* and */s/* can stand for *s', c'*). These words that are short and frequent can easily be inserted and deleted by the recognizer, having the result that any out-of-vocabulary word (OOV) can be replaced by a sequence of highly probable phonemes. An extreme example is the OOV *"s'épanousissait"* which was recognized as the word sequence *"c'est pas nous oui c'est"* [23].

The LRE SQALE (Speech recognizer Quality Assessment for Linguistic Engineering) project aimed to assess language-dependent issues in multilingual recognizer evaluation [89]. In the project the ARPA evaluation paradigm was used to assess the performance of the same system on comparable tasks in different languages (American English, British English, French and German) to determine cross-lingual differences, as well as different systems on the same data so as to compare different methods. Table 3 summarizes the experimental conditions and results of the evaluation, in which the test data were selected so as to control the OOV rate. This exercise demonstrated that the same recognition technology and evaluation methodolgy could be successfully adapted to these 4 languages.

†The newspaper text corpora compared are the *Wall Street Journal* (English, 37M words) [2], *Le Monde* (French, 38M words) [31], *Frankfurter Rundschau* (German, 36M) [1], and *Il Sole 24 Ore* (Italian, 26M words) [24], where the total number of words of text material are given in parentheses.

9. Towards Usable Systems

In adapting a state-of-the-art speech recognizer developed in a laboratory for real-world use, all aspects of the speech recognizer must be reconsidered from signal capture to adaptive acoustic and language models. Given application constraints, standard laboratory development procedures may need to be revised. At LIMSI we have recently faced this challenge in the context of the ESPRIT MASK (Multimodal-Multimedia Automated Service Kiosk) project, aimed at providing access to rail travel information [28]. The speech recognition requirements for the MASK information kiosk are: speaker-independence; real-time spontaneous, continuous speech recognition; a recognition vocabulary including 600 station/city names; and robustness as the expected background noise level for the MASK kiosk located in a Parisian train station is on the order of 63dBA SPL. In this section we address issues related to signal capture and real-time decoding in non-ideal environmental conditions. Issues related to recognition of spontaneous speech are discussed in the next section.

In order to better simulate the acoustic conditions of the final kiosk, a data collection kiosk has been built according to the physical specifications supplied by ergonomics experts. This data collection kiosk, shown in Figure 1, is being used to carry out laboratory experiments and to record data under more realistic conditions, by placing the users in conditions closer to that of real use. The touch screen (1) is located so as to accomodate a wide variety of user sizes. In order to account for the different customer heights and positions when using the kiosk, 3 PCC (Phase Coherent Cardioid) microphones have been positioned around the screen cavity on the top (2), left (3) and right (4) of the screen. Based on the SNR of each channel, the output of one of the three microphones is selected. Beam forming was considered but found to not be efficient for the kiosk configuration, since the distance between the speaker and the closest microphone is less than the distance between microphones. A fourth channel is used to capture the signal played over the loudspeaker, coming from the message synthesizer or from video soundtracks, in order to compensate for the acoustic feedback on the microphones.

In order to simulate the environmental conditions of the kiosk, measurements were carried out in a Parisian train station to estimate the expected mid working day background noise. Laboratory subjects are recorded in both quiet and noisy conditions, so as to model potentially different user behaviors. A touch-to-talk mechanism is used to get a rough estimate of the query endpoints, as well as to avoid processing queries not directed to the system. The system response signal is cancelled only until the user's speech is detected, and the response signal is stopped as soon as possible after the touch is detected.

Language	*Training Corpus*	*# Participants*	*Vocabulary Size*	*OOV rate*	*Word Error (%)*
AmEng	WSJ0	4	20k	1.43	12.9 – 14.7
BritEng	WSJCAM0	3	20k	1.66	13.8 – 15.4
French	BREF-80	3	20k	1.70	15.1 – 16.1
German	Phondat/FR	3	64k	1.85	16.1 – 19.7

Table 3 Results in % word error of the SQALE evaluation for speech recognition in four languages (American English, British English, French and German) with 20k/64k trigram LMs.

Fig. 1 The LIMSI MASK data collection kiosk, (1) touch screen, (2), (3) and (4) are microphones, and (5) loudspeaker.

An important aspect of real-time speech recognition is the design of a fast search algorithm that maintains high recognition accuracy. Even though the MASK task is less ambitious than our laboratory 65k system, decoding is still not trivial, particularly in the presence of noise which slows down the decoder. Since an immediate response is required, not too much time can be spent in multipass decoding. Recognizer optimization is trickier given the constraint of real-time decoding, as performance may be more dependent upon other factors (such as the pruning level) than on the accuracy of the acoustic models. For laboratory systems our experience has been that improving model accuracy both improves recognition performance, and leads to better decoding due to more efficient pruning. However, if the decoding strategy remains the same, the trade off between accuracy and speed is dependent upon the total number of model parameters. Several techniques have been combined to achieve the MASK goals: a lexicon tree, multipass decoding, distributed LM weights, Gaussian shortlists and gender dependent (GD) acoustic models.

The network used in the bigram pass is built in such a way that the word tails (the last phone or last few phones of the word) are shared between the lexicon tree and the linear representation of the words, so as to minimize the number of interword connections. Evidently single phoneme words are represented only once. Bigram decoding with CI phone models is realized in real-time (RT), where real time is defined as taking 1s to process a 1s utterance. The language model weights are distributed over the phone graph so as to allow the use of a reduced pruning threshold, enabling both faster and more accurate search. When a trigram LM is used, a second decoding pass is carried out using a word graph generated with the bigram. The result of the first decoding pass is used to guide the search of the second pass, enabling the use of a dynamic pruning threshold. This second pass uses more accurate acoustic and language models and can be carried out in about 20% of CPU time of the first pass.

For small and medium vocabulary tasks, the state likelihood computation can represent a significant portion of the overall computation. One way to speed up this computation is to reduce the number of Gaussians needing to be considered to compute the likelihood for a state by preparing a Gaussian short list for each HMM state and each region of the quantified feature space [12]. Doing so, only a fraction of the Gaussians of each mixture is considered during decoding. This approach allows us to reduce the average number of examined Gaussians per mixture from 12 to 4 without any loss in accuracy.

One easy way to improve the accuracy of the recognizer is to use GD acoustic models. By building two separate networks and carrying out frame-synchronous decoding on the two networks in parallel, recognition accuracy can be improved without increasing the decoding time since after only a few frames the network corresponding to the speaker's gender is under consideration [52]. The small overhead of searching the 2 networks at the start of the sentence is largely compensated by more efficient pruning due to the use of more accurate models.

To deal with noisy conditions, the data-driven model adaptation scheme used in the LIMSI Nov95

NAB system [34] is applied. Related to model combination schemes [26], [27], adaptation is based on the following model of the observed signal y given the input signal x: $y = (x + n) * h$, where n is the additive noise and h the convolutional noise. Compensation is performed iteratively, where refined estimates of n and h are obtained before each decoding process. To adapt to different conditions at different times of day, noise estimation and compenstation can be performed at regular intervals or inbetween customer sessions. In order to perform the speech analysis in real-time, sentence-based cepstral mean removal is approximated by removing the mean of the previously observed frames, where the cepstrum mean is updated at each frame with a first order filter $(1 - 0.995z^{-1})$.

10. Towards Natural Speech and Speech Understanding

The capabilities of speech recognition systems in multiple languages reported here have all been obtained using read-speech, recorded in laboratory conditions. To approach more closely future speech recognition applications, it is necessary to be able to recognize naturally spoken utterances. It is well-known that spontaneous speech does not respect written grammar, and has common phenomena such as hesitations, filler words, false starts, repetitions and repairs. The speaking style is often more relaxed than read speech, and more phonological modifications are observed in which word realizations can differ from their canonical lexical representation. For a wide range of applications it is also likely that the system will need to understand the linguistic content of the utterance, not only to simply transcribe it into words. For many tasks a dialog component will be necessary, which can also be used to reduce the task perplexity by using different language models in different dialog states. A dialog component in turn requires a response generation component, optionally with vocal output. Although in this paper we address only the speech recognition aspects of spoken language understanding systems, we acknowledge the important roles of the dialog management and response generation components in system design and development.

Recognition of spontaneous speech implies several consequences for the recognizer, including identifying available information sources which can be used to bring up an initial system. In contrast to a dictation application where it is relatively straight-forward to select a recognition vocabulary from large written corpora, for specific tasks, *a priori* even the vocabulary size is not known, and there usually are no application-specific training data (acoustic or textual) available. A commonly adopted approach for data collection is to start with an initial system (that may involve a Wizard of Oz configuration to replace non-existant system components) and to collect a set of data which can be used to start an iterative development cycle. The recognition vocabulary and language model are initially based on the designers' expectations and task domain knowledge, and augmented according to the collected corpus. The capacity to easily add new words is thus essential.

The most effective manner of obtaining representative speech data is with preliminary versions of a complete system. It has been observed that as the system improves, subjects speak more easily and use longer and more varied sentences [55]. They are also more likely to perceive that errors are their own fault, rather than the system's. As a result they continue to speak relatively naturally to the system, enabling the collection of more representative spontaneous speech.

Different approaches have been taken for interfacing between the speech recognizer and the natural language (NL) understanding component. In most systems a bottom up approach is taken, where the output of the recognizer is passed to the NL component. The recognizer output can be the best word string, an N-best list of word strings, or a word lattice. In the latter cases, the NL component can be used to filter the recognizer output.

The most widely known work in this area are the ARPA ATIS task[78] and the ESPRIT SUNDIAL project[77]. More recent projects are the ESPRIT MASK project [28] and the Language Engineering Multilingual Action Plan (LE-MLAP) projects RAILTEL and MAIS. The range of results obtained for different sites in the ARPA ATIS benchmark tests [71], [72] are shown in Table 4. The performance of the best system is seen to have significantly improved from 1993 to 1994.† The word error rates of the best system are quite low, and the spoken language system (SLS) understanding error based on the spoken input is not much larger than the NL understanding error obtained using the orthographic transcription of the query. The performance of the L'ATIS system[11], a French ATIS system developed at LIMSI, is within the same range of the ATIS systems. For the MASK system, reducing the word error from 15% in to 10%, led to a 29% reduction in SLS error to 15%. The current MASK NL understanding error is 7%. We expect that further improvements in recognition performance will reduce the difference in NL and SLS understanding error rates, as was observed for the ARPA ATIS task.

For spoken language applications, global evaluation measures and subjective user ratings are likely to be more important than word error and query understanding rates. An important need for such applications is the capability to reject out of domain queries. Our strategy is to estimate the *a posteriori* sentence probability for the recognizer hypothesis,

†Although benchmark SLS tests were carried out prior to 1993, the scoring used a weighted error which makes it difficult to compare with these results.

Test	*SPREC Word Error (%)*	*NL Error (%)*	*SLS Error (%)*
ATIS'93	3.3 - 9.0	9.3 - 43.1	13.2 - 46.8
ATIS'94	1.9 - 14.1	5.9 - 41.7	8.6 - 55.3
L'Atis Jan'95	6.0	11.0	12.0

Table 4 Range of spoken language system results for the ATIS task. Speaker-independent speech recognition results (SPREC) are given in terms of word error. Natural language (NL) and Spoken Language System (SLS) results are in unweighted error, which is the sum of *#(no answer) + #(wrong answer)*. Results are given for queries of type A+D (A answerable without context, D answerable with context).

i.e. $\Pr(w|x)$, by modeling the talker as a source of phones with phonotactic constraints provided by phone bigrams. We approximate $\Pr(w|x)$ by $\Pr(\varphi_w|x) \simeq f(x|\varphi_w)\Pr(\varphi_w)/\max_\varphi f(x|\varphi)\Pr(\varphi)$, where φ_w is the recognized phone transcription corresponding to the recognizer hypothesis w. $\Pr(\varphi_w|x)$ is then compared to a fixed threshold to decide whether to accept or reject the query. This procedure requires only a small amount of additional computation if you use simple models and a tight prunning threshold.

11. Summary and Perspectives

In this paper we have provided an overview of the state-of-the-art in laboratory speaker-independent, large vocabulary continuous speech recognition systems, and discussed some of the issues involved in adapting such technology to the requirements of real-world applications. Much of the recent progress made over the last 5-10 years in LVCSR has been made possible by the availability of large corpora for training and testing speech recognition and understanding technology. However, despite our experience as a community, constructing corpora that are representative, complete, and yet at the same time not too big, is an open research area. It is extremely hard to even demonstrate the effects of different corpus design strategies. Yet at the same time, the performance of all recognition systems is acknowledged to be quite dependent on the training data.

For dictation tasks, it is relatively easy to obtain text data for training language models. After processing of the texts to clean them and to transform them to be closer to observed reading styles, a task vocabulary can be selected and language models trained. A subset of texts can be selected to ensure good phonetic coverage and used as prompts to obtain spoken data. Obtaining representative data for spontaneous speech is much more difficult and expensive. It is difficult, if not impossible, to control the content of the speech data, be it at the semantic, lexical or phonetic level, or the speaking style. The Switchboard corpus [38] contains a rich set of telephone conversations on a variety of topics. Even with the detailed orthographic transcriptions, language modeling for this task remains a challenge.

For LVCSR, we attempt to obtain speaker-independence by recording speech from many different speakers, hoping to cover the speaker population. Opinions differ as to the number of speakers needed: some favor more data from a fewer number of speakers, while others favor less data per speaker from more speakers. In order to have models that are relatively task independent, it is important to cover many different phonetic contexts in the training corpus. More generally speaking, we do not know how to design and train accurate task-independent models that can be used for various applications without the need for additional data collection.

While rapid progress has been made in LVCSR, there are many factors that are observed to influence the speech recognition performance, and many outstanding problems. Some of these unsolved problems are inter-speaker variability, speaking rate, and lexical and language modeling. Regarding inter-speaker variability, even todays best systems have a huge difference in performance (sometimes as much as a factor of 30) between the word error of the best speaker (1-2%) and the word error of the worst speaker (25-30%). These performance differences are often related to differences in speaking rate - speakers that are much faster or slower than the norm tend to have much higher word error rates. Differences in speaking rate affect not only the acoustic level, but also the phonological level and maybe even the word level. At the lexical level, it should be possible to choose among pronunciation variants according to observed pronunciations for the given speaker. A person who pronounces a word in a given manner is likely to say derived forms, and other similar words with a similar form. Similarly, at the cross-word level, different speakers make use of different phonological rules. For most speakers, the choice of rules is systematic, yet no system that we know of is able to make use of this consistency. More generally, today's systems do not easily adapt to new accents, be they different dialects or speech of non-native speakers. As humans we usually are able to do this rather quickly.

Concerning language modeling, the n-gram language models which are reasonably successful for dictation in English, are less efficient for more highly inflected languages (such as French and German). Higher order n-grams or class-based n-grams may be more appropriate for such languages. Efforts in adaptinve language modeling are enticing, but still have not resulted in significant performance improvements. There certainly remains room for a lot of research in this area, particularly in language modeling for spontaneous speech, where models trained on written texts are sure to be less effective. Perhaps the ultimate question is how far can we go in recognizing speech without understanding it? We do not know this limit.

References

[1] ACL-ECI CDROM, distributed by Elsnet and LDC.

[2] "CSR corpus. Language model training data," *NIST Speech Disc 22-1 and 22-2*, Produced by LDC, Aug. 1994.

[3] M. Adda-Decker, G. Adda, L.F. Lamel, J.L. Gauvain, "Developments in Large Vocabulary, Continuous Speech Recognition of German," *Proc. IEEE ICASSP-96*, Atlanta, GA, **I**, pp. 153-156, May 1996.

[4] S. Austin, R. Schwartz and P. Placeway, "The Forward-Backward Search Strategy for Real-Time Speech Recognition," *Proc. IEEE ICASSP-91*, Toronto, Canada, pp. 697-700, May 1991.

[5] L.R. Bahl, J.K. Baker, P.S. Cohen, N.R. Dixon, F. Jelinek, R.L. Mercer, and H.F. Silverman,"Preliminary results on the performance of a system for the automatic recognition of continuous speech," *ICASSP-76*, 1976.

[6] L.R. Bahl, F. Jelinek and R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, **PAMI-5**(2), pp. 179-190, March 1983.

[7] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo, and M. Picheny, "A Fast Match for Continuous Speech Recognition Using Allophonic Models," *ICASSP-92*, San Francisco, CA, **1**, pp. 17-21, March 1992.

[8] L.R. Bahl, P. Brown, P. de Souza, R.L. Mercer, and M. Picheny, "Acoustic Markov Models used in the Tangora Speech Recognition System," *Proc. IEEE ICASSP-88*, New York, NY, **1**, pp. 497-500, April 1988.

[9] L.R. Bahl, S. Balakrishnan-Aiyer, M. Franz, P.S. Gopalakrishnan, R. Gopinath, M. Novak, M. Padmanabhan, and S. Roukos, "The IBM Large Vocabulary Continuous Speech Recognizer for the ARPA NAB News Task," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 121-126, Jan. 1995.

[10] J. Baker, J. Baker, P. Bamberg, K. Bishop, L. Gillick, V. Helman, Z. Huang, Y. Ito, S. Lowe, B. Peskin, R. Roth and F. Scattone, "Large Vocabulary Recognition of Wall Street Journal Sentences at Dragon Systems," *Proc. DARPA Speech and Natural Language Workshop*, pp. 387-392, Feb. 1992.

[11] S.K. Bennacef, H. Bonneau-Maynard, J.L. Gauvain, L.F. Lamel and W. Minker, "A Spoken Language System For Information Retrieval," *Proc. ICSLP'94*, Yokohama, Japan, **3**, pp. 1271-1274, Sep. 1994.

[12] E. Bocchieri, "Vector quantization for efficient computation of continuous density likelihoods," *Proc. IEEE ICASSP-93*, Minneapolis, MN, **2**, pp. 692-695, May 1993.

[13] L. Chase, R. Rosenberg, A. Hauptmann, M. Ravishankar, E. Thayer, P. Placeway, R. Weide and C. Lu, "Improvements in Language, Lexical and Phonetic Modeling in Sphinx-II," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 60-65, Jan. 1995.

[14] F. Chen, "Identification of contextual factors for pronunciations networks," *Proc. IEEE ICASSP-90*, Albuquerque, NM, pp. 753-756, April 1990.

[15] Y.L. Chow, R. Schwartz, S. Roukos, O. Kimball, P. Price, F. Kubala, M.O. Dunham, M. Krasner and J. Makhoul, "The Role of Word-Dependent Coarticulatory Effects in a Phoneme-Based Speech Recognition System", *Proc. IEEE ICASSP-86*, Tokyo, Japan, **3**, pp. 1593-1596, April 1986.

[16] M. Cohen, *Phonological Structures for Speech Recognition*, PhD Thesis, U. Ca. Berkeley, 1989.

[17] S. Davis and P. Mermelstein, "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, **28**(4), pp. 357-366, 1980.

[18] V. Digalakis and H. Murveit, "Genones: Optimization the Degree of Tying in a Large Vocabulary HMM-based Speech Recognizer," *Proc. IEEE ICASSP-94*, Adelaide, Australia, **1**, pp. 537-540, April 1994.

[19] V. Digalakis, M. Weintraub, A. Sankar, H. Franco, L. Neumeyer, and H. Murveit, "Continuous Speech Dictation on ARPA's North American Business News Domain," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 88-93, Jan. 1995.

[20] J. Dreyfus-Graf, "Sonograph and Sound Mechanics," *J. Acoust. Soc. America*, **22**, pp. 731, 1949.

[21] H. Dudley and S. Balashek, "Automatic Recognition of Phonetic Patterns in Speech," *J. Acoust. Soc. America*, **30**, pp. 721, 1958.

[22] C. Dugast, R. Kneser, X. Aubert, S. Ortmanns, K. Beulen, and H. Ney, "Continuous Speech Recognition Tests and Results for the NAB'94 Corpus," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 156-161, Jan. 1995.

[23] C. Dugast, X. Aubert, and R. Kneser, "The Philips Large-Vocabulary System for American English, French and German," *Proc. ESCA Eurospeech'95*, Madrid, Spain, **1**, pp. 197-200, Sep. 1995.

[24] Marcello Federico, personal communication.

[25] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Trans. on Acoustics, Speech and Signal Processing*, **ASSP-29**, pp. 342-350, 1981.

[26] M.J.F. Gales and S.J. Young, "An improved approach to hidden Markov model decomposition of speech and noise," *Proc. IEEE ICASSP-92*, pp. 233-236, March 1992.

[27] M.J.F. Gales and S.J. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," *Computer Speech & Language*, **9**(4), Oct. 1995.

[28] J.L. Gauvain, S.K. Bennacef, L. Devillers, L.F. Lamel, and S. Rosset, "The Spoken Language Component of the Mask Kiosk," *Proc. Human Comfort and Security Workshop*, Brussels, Belgium, Oct. 1995.

[29] J.L. Gauvain, L.F. Lamel, G. Adda and M. Adda-Decker, "The LIMSI Nov93 WSJ System," *Proc. ARPA Spoken Language Technology Workshop*, Princeton, NJ, pp. 125-128, March 1994.

[30] J.L. Gauvain and C.H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. Speech and Audio Processing*, **2**(2), pp. 291-298, April 1994.

[31] J.L. Gauvain, L.F. Lamel, G. Adda and M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**(1-2), pp. 21-37, Oct. 1994.

[32] J.L. Gauvain, L.F. Lamel and M. Adda-Decker, "Developments in Large Vocabulary Dictation: The LIMSI Nov94 NAB System," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 131-138, Jan. 1995.

[33] J.L. Gauvain, L.F. Lamel and M. Adda-Decker, "Developments in Continuous Speech Dictation using the ARPA WSJ Task," *Proc. IEEE ICASSP-95*, Detroit, MI, pp. 65-68, May 1995.

[34] J.L. Gauvain, L.F. Lamel, G. Adda and D. Matrouf, "Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task," *Proc. IEEE ICASSP-96*, Atlanta, GA, **I**, pp. 73-76, May 1996.

[35] J.L. Gauvain and C.H. Lee, "Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models," *Proc. DARPA Speech and Natural Language Workshop*, Pacific Grove, CA, pp. 272-277, Feb. 1991.

[36] E. Giachin, A.E. Rosenberg and C.H. Lee, "Word Juncture Modeling using Phonological Rules for HMM-based Continuous Speech Recognition," *Computer Speech & Language*, **5**, pp. 155-168, 1991.

[37] L. Gillick and R. Roth, "A Rapid Match Algorithm for Continuous Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, pp. 170-172, June, 1990.

[38] J. Godfrey, E. Holliman and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," *Proc. IEEE ICASSP-92*, San Francisco, CA, pp. 517-520, March 1992.

[39] P.S. Gopalakrishnan, L.R. Bahl and R.L. Mercer, "A tree search strategy for large-vocabulary continuous speech recognition," *Proc. IEEE ICASSP-95*, Detroit, MI, **1**, pp. 572-575, May 1995.

[40] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. America*, **87**(4), pp. 1738-1752, 1990.

[41] D.M. Hindle, A. Ljolje, and M.D. Riley, "Recent Improvements to the AT&T Speech-to-Text (STT) System," *Proc. ARPA Speech Recognition Workshop*, Feb. 1996.

[42] M.M. Hochberg, S.J. Renals, A.J. Robinson, and D. Kershaw, "Large vocabulary continuous speech recognition using a hybrid connectionist-HMM system," in *Proc. ICSLP'94*, Yokohama, Japan, pp. 1499-1502, Sep. 1994.

[43] X. Huang, F. Alleva, M.Y. Hwang, and R. Rosenfeld, "An Overview of the SPHINX-II Speech Recognition System," *Proc. ARPA Human Language Technology Workshop*, Princeton, NJ, pp. 81-86, March 1993.

[44] M. Hwang and X. Huang, "Subphonetic Modeling with Markov States - Senone," *Proc. IEEE ICASSP-92*, San Francisco, CA, **1**, pp. 33-36, March 1992.

[45] M.Y. Hwang, X. Huang and F. Alleva, "Predicting Unseen Triphones with Senones," *Proc. IEEE ICASSP-93*, Minneapolis, MN, **II**, pp. 311-314, April 1993.

[46] F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. of the IEEE*, **64**(4), pp. 532-556, April 1976.

[47] F. Jelinek, B. Merialdo, S. Roukos and M. Strauss, "A Dynamic Language Model for Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop*, Pacific Grove, CA, pp. 293-295, Feb. 1991.

[48] S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. Acoustics, Speech, and Signal Processing*, **ASSP-35**(3), pp. 400-401, March 1987.

[49] D. Kershaw, A.J. Robinson and S.J. Renals, "The 1995 Abbot hybrid connectionist-HMM large-vocabulary recognition system," *Proc. ARPA Speech Recognition Workshop*, Feb. 1996.

[50] L.F. Lamel and G. Adda, "On Designing Pronunciation Lexicons for Large Vocabulary, Continuous Speech Recognition," *Proc. ICSLP'96*, Oct. 1996.

[51] L.F. Lamel and R. DeMori, "Speech Recognition of European Languages," *Proc. IEEE Automatic Speech Recognition Workshop*, Snowbird, Utah, pp. 51-54, Dec. 1995.

[52] L.F. Lamel and J.L. Gauvain, "Continuous Speech Recognition at LIMSI," *Proc. ARPA Workshop on Continuous Speech Recognition*, Stanford, CA, pp. 59-64, Sep. 1992.

[53] L.F. Lamel and J.L. Gauvain, "High Performance Speaker-Independent Phone Recognition Using CDHMM," *Proc. ESCA Eurospeech'93*, Berlin, Germany, pp. 121-124, Sep. 1993.

[54] L.F. Lamel and J.L. Gauvain, "A Phone-based Approach to Non-Linguistic Speech Feature Identification," *Computer, Speech and Language*, **9**(1), pp. 87-103, Jan. 1995.

[55] L.F. Lamel, S. Rosset, S.K. Bennacef, H. Bonneau-Maynard, L. Devillers, and J.L. Gauvain, "Development of Spoken Language Corpora for Travel Information", *Eurospeech'95*, Madrid, Spain, **3**, pp. 1961-1964, Sept. 1995.

[56] K.-F. Lee, "Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system," *Ph.D. Thesis*, Carnegie-Mellon University, 1988.

[57] C.J. Leggetter and P.C. Woodland, "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 110-115, Jan. 1995.

[58] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, **9** , pp. 171-185, 1995.

[59] A. Ljolje, M.D. Riley, D.M. Hindle and F. Pereira, "The AT&T 60,000 Word Speech-To-Text System," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 162-165, Jan. 1995.

[60] T. Matsuoka, K. Ohtsuki, T. Mori, S. Furui, and K. Shirai, "Large-Vocabulary Continuous Speech Recognition using the Japanese Business Newspaper (Nikkei) Task," *Proc. ARPA Speech Recognition Workshop*, Harriman, NY, Feb. 1996.

[61] H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub, "Large-Vocabulary Dictation using SRI's Decipher Speech Recognition System: Progressive Search Techniques," *Proc. IEEE ICASSP-93*, Minneapolis, MN, pp. II-319-322, April 1993.

[62] H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, **ASSP-32**(2), pp. 263-271, April 1984.

[63] H. Ney, R. Haeb-Umbach, B.H. Tran and M. Oerder, "Improvements in Beam Search for 10000-Word Continuous Speech Recognition," *Proc. IEEE ICASSP-92*, San Francisco, CA, **I**, pp. 9-12, March 1992.

[64] L. Nguyen, T. Anastasakos, F. Kubala, C. LaPre, J. Makhoul, R. Schwartz, N. Yuan, G. Zavaliagkos, and Y. Zhao, "The 1994 BBN/BYBLOS Speech Recognition System," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 77-81, Jan. 1995.

[65] Y. Normandin, D. Bowness, R. Cardin, C. Drouin, R. Lacouture, and A. Lazaridès, "CRIM's November 94 Continuous Speech Recognition System," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 153-155, Jan. 1995.

[66] J.J. Odell, V. Valtchev, P.C. Woodland, and S.J. Young, "A One Pass Decoder Design for Large Vocabulary Recognition," *Proc. ARPA Human Language Technology Workshop*, Princeton, NJ, pp. 405-410, March 1994.

[67] B.T. Oshika, V.W. Zue, R.V. Weeks, H. Neu, and J. Aurbach, "The Role of Phonological Rules in Speech Understanding Research," *IEEE Trans. Acoustics, Speech, Signal Processing*, **ASSP-23**, pp. 104-112, 1975.

[68] M. Ostendorf, A. Kannan, O. Kimball and J. R. Rohlicek, "Continuous Word Recognition Based on the Stochastic Segment Model," *Proc. ARPA Workshop on Continuous Speech Recognition*, Stanford, CA, pp. 53-58, Sep. 1992.

[69] D.S. Pallett, J.G. Fiscus and J.S. Garofolo, "Resource Management Corpus: September 1992 Test Set Benchmark Results," *Proc. ARPA Workshop on Continuous Speech Recognition*, Stanford, CA, pp. 1-18, Sep. 1992.

[70] D.S. Pallett, J.G. Fiscus, W.M. Fisher, and J.S. Garofolo, "Benchmark Tests for the DARPA Spoken Language Program," *Proc. ARPA Human Language Technology Workshop*, Princeton, NJ, pp. 7-18, March 1993.

[71] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund, and M.A. Pryzbocki, "1993 Benchmark Tests for the ARPA Spoken Language Program," *Proc. ARPA Human Language Technology Workshop*, Princeton, NJ, pp. 49-74, March 1994.

[72] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund, A.F. Martin and M.A. Przybocki, "1994 Benchmark Tests for the ARPA Spoken Language Program," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 5-36., Jan. 1995.

[73] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, A.F. Martin and M.A. Przybocki, "1995 Hub-3 Multiple Microphone Corpus Benchmark Tests," *Proc. ARPA Speech Recognition Workshop*, Harriman, NY, Feb. 1996.

[74] D.B. Paul and J.M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *Proc. ICSLP'92*, Banff, CA, **2**, pp. 899-902, Oct. 1992.

[75] D.B. Paul, "An efficient A* stack decoder algorithm for continuous speech recognition with a stochastic language model," *Proc. DARPA Workshop on Speech and Natural Language*, Harriman, NY, pp. 405-409, Feb. 1992.

[76] D.B. Paul, "New Developments in the Lincoln Stack-Decoder Based Large Vocabulary CSR System," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 143-147, Jan. 1995.

[77] J. Peckham, "A New Generation of Spoken Dialog Systems: Results and Lessons from the SUNDIAL Project", *Proc. ESCA Eurospeech'93*, Berlin, Germany, pp. 33-40, Sep. 1993.

[78] P. Price, "Evaluation of Spoken Language Systems: The ATIS Domain," *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, pp. 91-95, June, 1990.

[79] M.D. Riley and A. Ljojle, "Automatic Generation of Detailed Pronunciation Lexicons", in *Automatic Speech and Speaker Recognition*, Kluwer Academic Pubs, Ch. 12, pp. 285-301, 1996.

[80] I. Rogina and A. Waibel, "The JANUS Speech Recognizer," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 166-169, Jan. 1995.

[81] R. Rosenfeld and X. Huang, "Improvements in Stochastic Language Modeling," *Proc. DARPA Workshop on Speech and Natural Language*, Harriman, NY, pp. 107-111, Feb. 1992.

[82] R. Rosenfeld, *Adaptive Statistical Language Modeling,* PhD Thesis, Carnegie Mellon University, 1994. (also *Tech. rep. CMU-CS-94-138*)

[83] R. Roth, L. Gillick, J. Orloff, F. Scattone, G. Gao, S. Wegmann and J. Baker, "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer, *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 116-120, Jan. 1995.

[84] A. Sankar, A. Stolke, T. Chung, L. Neumeyer, M. Weintraub, H. Franco, and F. Beaufays, "Noise-Resistant Feature Extraction and Model Training for Robust Speech Recognition," *Proc. ARPA Speech Recognition Workshop*, Feb. 1996.

[85] R. Schwartz, S. Austin, F. Kubala, and J. Makhoul,"New uses for N-Best Sentence Hypothesis, within the BYBLOS Speech Recognition System," *Proc. IEEE ICASSP-92*, San Francisco, CA, **I**, pp. 1-4, March 1992.

[86] R. Schwartz, Y. Chow, S. Roucos, M. Krasner, and J. Makhoul, "Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition," *Proc. IEEE ICASSP-84*, San Diego, CA, **3**, pp. 35.6.1-35.6.4, March 1984.

[87] S. Sekine and R. Grishman, "NYU Language Modeling Experiments for the 1995 CSR Evaluation," *Proc. ARPA Speech Recognition Workshop*, Harriman, NY, Feb. 1996.

[88] B. Shahshahani, "A Markov Random Field Approach to Bayesian Speaker Adaptation," *Proc. IEEE ICASSP-95*, pp. 697-700, May 1995.

[89] H.J.M. Steeneken and D.A. Van Leeuwen, "Multi-Lingual Assessment of Speaker Independent Large Vocabulary Speech-Recognition Systems: the SQALE Project, *Proc. ESCA Eurospeech'95*, Madrid, Spain, **2**, pp. 1271-1274, Sep. 1995.

[90] G. Tajchman, E. Fosler and D. Jurafsky, "Building Multiple Pronunciation Models for Novel Words Using Exploratory Computational Phonology," *Proc. ESCA Eurospeech'95*, Madrid, Spain, **3**, pp. 2247-2250, Sep. 1995.

[91] S. Takahashi and S. Sagayama, "Four-level Tied Structure for Efficient Representation of Acoustic Modeling," *Proc. IEEE ICASSP-95*, Detroit, MI, pp. 520-523, May 1995.

[92] T.K. Vintsyuk, "Speech discrimination by dynamic programming," *Kibnernetika*, **4**, p. 81, 1968.

[93] T.K. Vintsyuk, "Elements-wise recognition of continuous speech composed of words from a specified dictionary," *Cybernetics*, **7**, pp. 133-143, March-April 1971.

[94] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev, and S.J. Young, "The development of the 1994 HTK large vocabulary speech recognition system," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 104-109, Jan. 1995.

[95] P.C. Woodland, M.J.F. Gales, D. Pye and V. Valtchev, "The HTK large vocabulary recognition system for the 1995 ARPA H3 task," *Proc. ARPA Speech Recognition Workshop*, Feb. 1996.

[96] S.J. Young, "The General Use of Tying in Phoneme-Based HMM Speech Recognisers," *Proc. IEEE ICASSP-92*, San Francisco, CA, pp. 569-572, March 1992.

[97] S.J. Young, J.J. Odell and P.C. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modeling," *Proc. ARPA Human Language Technology Workshop*, Princeton, NJ, pp. 307-312, March 1994.

[98] S.J. Young, P.C. Woodland, "The Use of State Tying in Continuous Speech Recognition," *Proc. ESCA Eurospeech'93*, **3**, pp. 2203-2206, Berlin, Germany, Sep. 1993.

[99] G. Zavaliagkos, R. Schwartz, J. McDonough, "Maximum *a Posteriori* Adaptation for Large Scale HMM Recognizers," *Proc. IEEE ICASSP-95*, pp. 725-728, May 1995.

[100] V. Zue. J. Glass, M. Phillips, and S. Seneff, "The MIT SUMMIT Speech Recognition System: A Progress Report", *Proc. DARPA Speech and Natural Language Workshop*, pp. 179-189, Philadelphia, Feb. 1989.